JYVÄSKYLÄ STUDIES IN COMPUTING
225

Sergey Chernov

Detecting Cellular Network Anomalies Using the Knowledge Discovery Process

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 225

Sergey Chernov

Detecting Cellular Network Anomalies Using the Knowledge Discovery Process

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen Beeta-salissa
joulukuun 11. päivänä 2015 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, Beeta hall, on December 11, 2015 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2015

Detecting Cellular Network Anomalies Using the Knowledge Discovery Process

JYVÄSKYLÄ STUDIES IN COMPUTING 225

Sergey Chernov

Detecting Cellular Network Anomalies Using the Knowledge Discovery Process

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2015

Editors
Timo Männikkö
Department of Mathematical Information Technology, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-6392-7
ISBN 978-951-39-6392-7 (PDF)

ISBN 978-951-39-6391-0 (nid.)
ISSN 1456-5390

Copyright © 2015, by University of Jyväskylä

Jyväskylä University Printing House, Jyväskylä 2015

ABSTRACT

Chernov, Sergey
Detecting Cellular Network Anomalies Using the Knowledge Discovery Process
Jyväskylä: University of Jyväskylä, 2015, 62 p.(+included articles)
(Jyväskylä Studies in Computing
ISSN 1456-5390; 225)
ISBN 978-951-39-6391-0 (nid.)
ISBN 978-951-39-6392-7 (PDF)
Finnish summary
Diss.

Analytical companies unanimously forecast the exponential growth of mobile traffic consumption over the next five years. The densification of a network structure with small cells is regarded as a key solution to meet growing capacity demands. The manual management of a multi-layer network is a very expensive, error prone, and sluggish process. Hence, the automation of the whole life cycle of network operation is highly anticipated. To this aim 3GPP introduces a self-management concept referred to as SON. It is envisioned that SON updates information concerning the latest network conditions through the MDT mechanism. MDT enables a network operator to collect radio and service quality measurements from regular mobile phones. Self-healing is SON's functionality that implements fault management in radio networks. The automated and timely detection of a malfunctioning cell is one of the crucial challenges for network operators.

The thesis investigates the topic of self-organizing radio networks and proposes a cell outage detection framework based on MDT measurements and advanced data mining techniques. The sequential analysis of LTE network events underlies the proposed idea. The conducted research demonstrates the feasibility of the original idea and designs the KDD process for the automated analysis of cell failures. The second part of the study improves the computational complexity and performance of the proposed solution. Besides, the research discovers the impact of location accuracy and scarcity of MDT measurements on the quality of cell outage detection. The validation of the framework has been conducted on the state-of-the-art LTE/LTE-A system level simulator. Results demonstrate reliable and timely detection of a malfunctioning cell. Therefore, the developed cell outage detection solution can be considered for the practical validation and implementation.

Keywords: radio networks, LTE, MDT, self-organizing networks, self-healing, cell outage, KDD, data mining, anomaly detection

Author

Sergey Chernov
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland

Supervisors

Professor Dr. Tapani Ristaniemi
Department of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland

Dr. Dmitry Petrov
Magister Solutions Ltd.
Jyväskylä, Finland

Reviewers

Dr. Seppo Hämäläinen
Ooredoo Group
Doha, Qatar

Professor Dr. Yevgeni Koucheryavy
Department of Electronics and Communication Engineering
Tampere University of Technology
Tampere, Finland

Opponent

Dr. Tech. Jarno Niemelä
Elisa Ltd.
Helsinki, Finland

ACKNOWLEDGEMENTS

While I am writing this text, my thesis is ready and I am happy to be at the final stage of my Ph. D. studies. Being relaxed and looking back into the past, I recognize that I have made a big effort during the last four years. However, my efforts would never have been enough to accomplish the research without the encouragement and comprehensive support from my colleagues and friends.

My deepest gratitude goes to my first supervisor, Professor Dr. Tapani Ristaniemi, for helping me in Ph. D. research with patience and invaluable support, especially during the first few years. I more than appreciate timely advice, positive discussions, and beneficial cooperation with my second supervisor, Dr. Dmitry Petrov. I am very thankful to Dr. Fedor Chernogorov for the broad support, not only in the research, without which I would not come to Finland.

Furthermore, I am very grateful to my co-authors Professor Dr. Tapani Ristaniemi, Kimmo Brigatti, Dr. Fedor Chernogorov, Dr. Dmitry Petrov, Michael Cochez, and Assistant Professor Dr. Mykola Pechenizkiy, for their significant help and inspiring collaboration. Special thanks go to Assistant Professor Dr. Mykola Pechenizkiy, who invited me for a research visit to the Eindhoven University of Technology. During those brief two months I obtained an unmeasurable life and research experience, being a part of his tight-knit team. In addition, I wish to acknowledge my colleagues Dr. Alexander Semenov and Alexandr Maslov, without whom the research would have been much harder and less enjoyable.

I am grateful to my supervisors and reviewers of my dissertation, Dr. Seppo Hämäläinen and Professor Dr. Yevgeni Koucheryavy, for their constructive comments and Dr. Tech. Jarno Niemelä for being my opponent.

I wish to express my appreciation to the University of Jyväskylä; the Graduate School in Electronics, Telecommunications and Automation (GETA); the Foundation of Nokia Corporation; Doctoral training network in Electronics, Telecommunications and Automation (DELTA), who made the research financially possible and enabled me to focus on the study.

During the long journey, that is called the Ph.D. study, the support from friends was very important and valuable to me. I really enjoyed spending my time with the international friends I made in Jyväskylä. It is impossible to overestimate the support from my close friends, Dr. Alexander Khodunin, Irina Kashnikova, Michael Likholet, Nadezhda Kuznetsova, Alexander Belov, Nikita Potapov, Dmitriy Spiridonov, Andrew Schneider, Alexander Dvoichenkov, Pavel&Daria (with Youri) Gonchukovi, and Kirill Markov, who found the time and opportunity to visit me and create a homelike atmosphere. Moreover, I want to thank my school friends, Anton Gunygin and Anton Kliuchevskii, for their advice and continuous involvement in my life.

My warmest thanks go to my mother Galina, father Anatoliy, and elder brother Evgeniy, who spent many evenings with me on Skype and enabled me to feel at home away from home. Their strong support and faith in me inspired and gave me confidence throughout the whole study.

LIST OF ACRONYMS

k-NN	k Nearest Neighbors
2G	2nd Generation
3G	3rd Generation
3GPP	3rd Generation Partnership Project
4G	4th Generation
5G	5th Generation
A-GPS	Assisted Global Positioning System
bps	bits per second
BSC	Base Station Controller
BTS	Base Transceiver Station
Capex	Capital expenses
CM	Configuration Management
CQI	Channel Quality Indicator
CRISP-DM	Cross-Industry Standard Process for Data Mining
CS	Circuit-Switched
E-UTRAN	Evolved UTRAN
eNB	evolved Node B
FindCBLOF	Find Cluster-Based Local Outlier Factor
FM	Fault Management
FN	False Negative
FP	False Positive
FPR	False Positive Rate
GERAN	GSM EDGE Radio Access Network
GPS	Global Positioning System
GSM	Global System for Mobile Communication
HO	Handover
HOF	Handover Failure
HSPA	High Speed Packet Access
Hz	Hertz
ID	Identification
IMEI	International Mobile Station Equipment Identity
Impex	Implementation expenses
IMS	Internet protocol based Multimedia Subsystem
IP	Internet Protocol

KDD	Knowledge Discovery in Databases
KDnuggets	Knowledge Discovery nuggets
KPI	Key Performance Indicator
LSH	Local-Sensitive Hashing
LTE	Long Term Evolution
LTE-A	LTE Advanced
MCA	Minor Component Analysis
MDT	Minimization of Drive Tests
MHz	Megahertz
MIMO	Multiple Input Multiple Output
MT	Mobile Terminal
NB	Node B
NGMN	Next Generation Mobile Networks
NSN	Nokia Siemens Networks
OAM	Operation, Administration, and Maintenance
Opex	Operating expenses
PAD	Probabilistic Anomaly Detection
PLP	Physical Layer Problem
PM	Performance Monitoring
PR	Precision Recall
PS	Packet-Switched
RACH	Random Access Channel
RAE	Ration of Adjacent Eigenvalues
RAN	Radio Access Network
RAT	Radio Access Technology
RLF	Radio Link Failure
RMSE	Root Mean Squared Error
RNC	Radio Network Controller
ROC	Receiver Operating Characteristic
RRC	Radio Resource Control
RSRP	Reference Signal Received Power
RSRQ	Reference Signal Received Quality
SEMMA	Sample, Explore, Modify, Model, and Assess
SOM	Self-Organizing Map
SON	Self-Organizing Network
SRI	Stanford Research Institute

TCE	Trace Collection Entity
TCO	Total Cost of Ownership
TN	True Negative
TNR	True Negative Rate
TP	True Positive
TPR	True Positive Rate
UE	User Equipment
UTRAN	Universal Terrestrial Radio Access Network
WCDMA	Wideband Code Division Multiple Access
Wi-Fi	Wireless Fidelity

LIST OF FIGURES

CONTENTS

LIST OF INCLUDED ARTICLES

PI Fedor Chernogorov, Tapani Ristaniemi, Kimmo Brigatti, Sergey Chernov. *N*-gram Analysis for Sleeping Cell Detection in LTE Networks. *38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada*, 2013.

PII Sergey Chernov, Fedor Chernogorov, Dmitry Petrov, Tapani Ristaniemi. Data Mining Framework for Random Access Failure Detection in LTE Networks. *25th IEEE Annual International Symposium on Personal, Indoor, and Mobile Radio Communication (PIMRC), Washington DC, USA*, 2014.

PIII Sergey Chernov, Michael Cochez, Tapani Ristaniemi. Anomaly Detection Algorithms for the Sleeping Cell Detection in LTE Networks. *81st IEEE Vehicular Technology Conference (VTC) Spring, Glasgow, Scotland*, 2015.

PIV Sergey Chernov, Dmitry Petrov, Tapani Ristaniemi. Location Accuracy Impact on Cell Outage Detection in LTE-A Networks. *11th IEEE International Wireless Communications & Mobile Computing Conference (IWCMC), Dubrovnik, Croatia*, 2015.

PV Fedor Chernogorov, Sergey Chernov, Kimmo Brigatti, Tapani Ristaniemi. Sequence-based Detection of Sleeping Cell Failures in Mobile Networks. *Wireless Networks: The Journal of Mobile Communication, Computation and Information*, 2015.

PVI Sergey Chernov, Mykola Pechenizkiy, Tapani Ristaniemi. The Influence of Dataset Size on the Performance of Cell Outage Detection Approach in LTE-A Networks. *10th IEEE International Conference on Information, Communications and Signal Processing (ICICS), Singapore*, 2015.

1 INTRODUCTION

A cell phone is the most wide spread piece of high technology. With the original purpose of making calls, nowadays it supports a variety of multimedia services. As a result, a smartphone has become an integral part of our everyday life. We leverage smartphones for writing emails, watching YouTube videos, or even making video calls. With the growing number of mobile phone subscribers, the quantity of media services increases. At the same time end users are becoming more and more demanding in terms of quality provided services. These factors will lead to the exponential grows of mobile traffic consumption over the coming years, (Ericsson, 2015).

In order to meet the increasing demands to mobile traffic consumption, network operators have to improve spectral efficiency, widen signal bandwidths, and leverage frequency reuse principle. Spectral efficiency represents the number of bits per second (bps) transmitted via 1 Hertz (Hz) of frequency band. For example, Wideband Code Division Multiple Access (WCDMA) (Release 6) demonstrates the spectral efficiency of 3 bps/Hz. With the recent introduction of Multiple Input Multiple Output (MIMO) technology – one device transmits and receives radio signals via multiple antennas – the spectral efficiency reaches 16 bps/Hz and 30 bps/Hz for Long Term Evolution (LTE) (Release 8) and LTE Advanced (LTE-A) (Release 10) technologies respectively, see (Sesia et al., 2011). The term signal bandwidth is a spectrum band utilized for data transmission. For the aforementioned technologies the allowed signal bandwidth equals 5 Megahertz (MHz) (Release 6), 20 MHz (Release 8), and 100 MHz (Release 10). Frequency reuse principle, in its turn, implies the reusing of the same frequency band. It becomes possible due to signal attenuation in the propagation. In future networks the dense deployment of small cells will heavily utilize frequency reuse principle. Small cells will enable to increase the quality of service and offload spots with high mobile traffic consumption, (Hwang et al., 2013).

The advances in cellular technologies pose considerable challenges and require intelligent solutions from network operators. On one side, the radio network architecture becomes very complex, which increases maintenance costs. On the other side, revenue is not growing that fast. This effect is known as the scissors effect, (Blennerud, 2010). In order to decrease expenses and prevent the scissors effect, the alliance of mobile operators Next Generation Mobile Networks (NGMN) reports network use cases that need automation, (NGMN, 2008). In the report NGMN focuses on the automation of cheap, but frequent operations. The automation of expensive and rare operations obtained low priority, because it requires significant efforts and looks unprofitable. As a response to the initiatives of mobile operators, 3rd Generation Partnership Project (3GPP) introduces a Self-Organizing Network (SON) concept, which aims to automate network management, improve network performance, and reduce network deployment and maintenance expenses, (3GPP, 2014b; Hämäläinen et al., 2012).

SON concept consists of three solutions, namely self-configuration, self-optimization, and self-healing. Self-configuration refers to the automated configuration of the newly established networks. Self-optimization mechanism automatically adjusts network parameters in order to guarantee high quality of service and reduce maintenance costs. The purpose of the self-healing mechanism is to automatically detect and address network operation problems, avoiding significant impact on subscribers' experience and reducing operational expenses.

Later in (NGMN, 2010) NGMN formulates top 10 operational efficiency recommendations for SON. The list includes standardized Minimization of Drive Tests (MDT) method and a cell outage detection function, which is part of the self-healing mechanism. MDT is a method that enables mobile operators to collect measurements from a regular cell phone. A mobile operator can employ these measurements to improve capacity, coverage, and performance characteristics of a radio network. One of the use cases is the identification of malfunctioning cells. The automation of this process allows fast and effective cell outage detection, while reducing the risk of human error to a minimum. Despite its benefits, the implementation of MDT functionality is a challenging task for network operators. It requires an intelligent management of data flow, data storage, and data processing mechanisms.

1.1 Objectives and Scope

MDT is a data collection process and an auxiliary mechanism for SON functions, such as self-healing. MDT concerns radio and service quality measurements from a regular user device. Since the measurements are collected from all devices, MDT database may include an enormous amount of records. The management of this huge database, especially processing, is a very challenging task for network operators. Knowledge Discovery in Databases (KDD) is known to be an efficient tool for the processing and extraction of information from large volumes of data. KDD is an umbrella term for machine learning and data mining techniques, which address anomaly detection, clustering, and other types of information retrieval problems. Therefore, network operators can employ KDD methods to analyze huge MDT databases and improve a network's performance.

SON includes a self-healing mechanism, which can leverage anomaly detection methods for the automated detection of malfunctioning cells. Cell failures is a highly important problem, because it directly affect on customer satisfaction of wireless services. Unsatisfied clients may change a telecommunication provider, leading to the loss of the provider's revenue. Therefore, effective and timely detection of cell failures is a vital problem for mobile operators. Sleeping cell is a special case of cell outage, which makes mobile service unavailable for subscribers even though the cell still appears to be operable from the networks point of view. Usually, sleeping cell outage lasts for days until it becomes unveiled by multiple user complaints or after a detailed analysis of expensive drive tests measurements. For these reasons the timely detection of sleeping cells is of high importance for network operators.

The primary goal of the thesis is to leverage MDT measurements and to develop an advanced data mining framework for the detection of sleeping cells in LTE/LTE-A networks. To fulfill the goal, the thesis formulates the following tasks:

- Define and simulate a suitable model for sleeping cell problem in LTE/ LTE-A networks.
- Consider the feasibility of using advanced data mining and machine learning methods for cell outage detection.
- Design and develop a data mining framework to pinpoint anomalous network behavior via MDT measurement.

The actual sleeping cell failure is modeled in LTE/LTE-A system level simulator via a malfunctioning of a random access procedure. The simulator is a validated and verified tool used by the Nokia Research Center and Nokia Solutions and Networks research groups. Random access procedure is a constitutional part of Random Access Channel (RACH) and is deployed for initial access, radio resource connection protocol re-establishment, uplink synchronization, and handover. RACH failure makes it impossible for a user to establish a connection or to make a handover to the malfunctioning cell. However, this cell still carries traffic for those users who were connected before the failure has occurred. Typically, RACH problem would be detected only after a long observation time or after a clients complaint.

We employ data mining techniques to examine MDT radio measurements and to discover abnormal network behavior. As it is specified in Release 10, MDT reports provide the location information of the conducted measurements (Baumann, 2014). In our approach, the location information plays an essential role in the identification of the problem area. The current study aims not only to design and verify a cell outage detection approach, but also to examine and to improve its constitutional modules.

The main research questions of the current thesis are as follows:

- *RQI*: Feasibility of the detection of a malfunctioning cell by means of user-level statistics and data mining techniques.

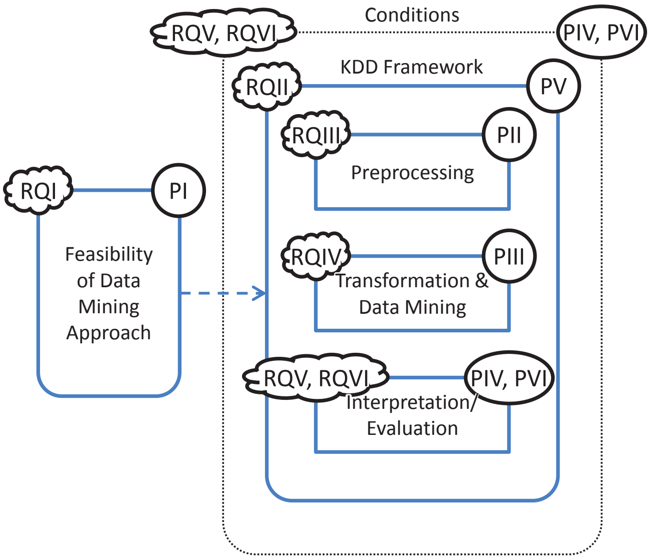


FIGURE 1 Relationship between the research questions and the included articles

- *RQII*: Design of KDD based cell outage detection framework via user-level statistics.
- *RQIII*: Optimize parameter N in N-gram analysis with respect to the complexity and performance of cell outage detection framework.
- *RQIV*: Examine different data mining algorithms in order to enhance the detection quality of impaired cells, while preserving low computational complexity of KDD framework.
- *RQV*: What is the positioning accuracy impact of user-level measurements on the performance of cell outage detection framework?
- *RQVI*: How does the density of user devices influence on the performance of cell failure identification approach in LTE/LTE-A networks?

The defined research questions are addressed in six scientific articles included in the thesis. Figure 1 illustrates the relationship between the research topic, research questions, and the included articles.

1.2 Main Contribution

The study of cell outage detection problem has been conducted in collaboration with researchers from the University of Jyväskylä, Magister Solutions Ltd., and Eindhoven University of Technology. KDD based cell outage detection framework is the result of this international and fruitful collaboration. Supervised learning is the underlying principle and is implemented in two phases: training and testing. During the training phase, the framework processes MDT measurements and learns a regular network operation. The training is needed to adjust and choose thresholds of the utilized algorithms of the KDD process. The next phase is testing. During this phase, the framework discovers the set of incoming MDT reports and automatically detects abnormal behavior in the network. In the conducted experiments, the abnormal behavior is caused by a sleeping cell. Location information of MDT measurements plays a crucial role in the detection of a sleeping cell. The main outcome of the study is the developed framework that enables a mobile operator to improve its network performance and save maintenance expenses.

The results of this research are published in six scientific articles, see Figure 1. Paper PI investigates the question whether data mining algorithms can be brought to the detection of malfunctioning cells in modern radio networks. Based on the findings of Paper PI, Paper PV elaborates a complete KDD framework. Further articles develop and examine constitutional steps of the proposed KDD process. Paper PII is concerned about the choice of parameters for the preprocessing step, as long as it impacts strongly on the computational efficiency. Paper PIII considers data mining step, which is crucial for the performance of the cell outage detection solution. Paper PIV investigates the impact of MDT location error on the accuracy of the proposed framework. Paper PVI evaluates the quality of the outage detection solution under the low density of user devices.

The author's contribution to the included articles lies in the design and development of the KDD cell outage detection framework. The results of all included papers are based on data that comes from LTE/LTE-A system level simulator. The author has not participated in the implementation of network simulations, but performed data validation. The elaborate description of author's contribution to the included articles is as follows. Paper PI designs a data mining approach for cell outage detection via the analysis of MDT measurements. The approach encompasses traditional and original methods. Traditional methods include N-gram feature selection algorithm, principal component analysis for dimensionality reduction, and Find Cluster-Based Local Outlier Factor (FindCBLOF) for anomaly detection. The discovered method is referred to as symmetry analysis. The author contributed ideas to the general approach, participated in the implementation and design of the data mining framework, and proposed to utilize the FindCBLOF algorithm. Paper PII concerns the preprocessing step of the developed cell outage detection solution. Parameter N in N-gram analysis is optimized with respect to the complexity and performance of the solution. The author introduced the original idea of the article, performed the analysis, and interpreted the results. Also the author proposed a heuristic measure for the performance analysis of the whole solution. Paper PIII examines the data mining step of the developed cell outage detection framework. The author proposed the original idea of the article, implemented anomaly detection algorithms, performed the analysis, and interpreted the results. Paper PIV investigates the location accuracy impact on cell outage detection in LTE-A networks. The author proposed the original idea of the article, performed analysis, and interpreted the results. Journal Paper PV provides a thorough description of the developed cell outage detection framework. The author played a key role in the development of the whole KDD process, actively participated in the writing process, and worked on comments from reviewers. Paper PVI demonstrates the influence of data set size on the performance of the cell outage detection approach in LTE-A networks. The author proposed the original idea of the article, performed analysis, and interpreted the results.

1.3 Structure of the Dissertation

The original thesis is composed of five chapters. The first one is an introduction chapter. Chapter 1 draws a general overview of the current state of modern cellular networks, describes a research gap in this field, and narrows down to the topic of the study. Also, the introduction chapter outlines the main results of the conducted research. The purpose of Chapter 2 is to introduce modern radio networks and their latest advances. The chapter provides necessary background knowledge about general network architecture and a manual network management, economic challenges for network providers and the outlook of mobile traffic consumption. The narration continues with self-organizing networks concept, which is a crucial step towards the automation of radio network operation. Minimization of drive tests is also presented in this chapter. The rest of the chapter is devoted to the review of cell outage detection approaches, which are developed by academic and industry researchers. Chapter 3 introduces origins and principles of data mining, which is a common approach for the analysis of large volumes of data. The general description of a KDD data analysis process is also provided. The main focus of the chapter is made on the developed cell outage detection framework. The framework and its elements are covered in detail. Chapter 4 focuses on thesis's research contribution. The main contribution is a cell outage detection framework, based on advanced data mining techniques. The chapter is split into two parts. The first part outlines problems and results of the research about the approach feasibility and framework's design. The second part summarizes the problems and results of the framework's analysis. Chapter 5 concludes the thesis. Here the author summarizes the results, discusses the limitation of the proposed approach, and outlines future challenges of cellular networks from the perspective of self-management.

2 MODERN RADIO NETWORKS

The main requirements of gradually and inevitably growing mobile cellular networks are high throughput, low capital expenditure as well as low operational costs. These aspects are dictated by the growing demands of high-speed access to communication services for less money. To this aim, radio access technologies and cellular networks are constantly evolving and achieve more efficient usage of radio resources. For example, the second generation (2G) of cellular technologies originates in 1991 in Finland, where Global System for Mobile Communication (GSM) standard was officially launched. The business use of the third generation (3G) networks began in 2001 in Japan. LTE is a common term for the fourth generation (4G). Nowadays, LTE is the most advanced commercially used wireless technology, which allows voice calling and fast access to data exchange services. The most anticipated features of 4G are autonomous maintenance and management. This concept is referred to as self-organization, which encompasses three functional parts: self-configuration, self-optimization, and self-healing. Such autonomous mechanisms are being developed by many research institutes and reveal a great field for research activity. As long as the current research focuses on the development of self-healing mechanisms, this dissertation misses to properly introduce LTE technology. The thorough description of LTE can be found in external sources such as (Sesia et al., 2011; Holma and Toskala, 2011).

This chapter introduces a general architecture of modern cellular networks. The narration continues with the outlook and economical aspects of wireless technologies. The following sections elaborate on the self-organizing networks concept and minimization of drive tests functionality. The section about cell outage detection problems and research activities in this field concludes the chapter.

2.1 Radio Networks Architecture

The architecture of modern radio networks is a very complex system, which is the result of a long-lasting evolution of wireless technologies. The main requirements for the network architecture are interoperability between various wireless standards and seamless handovers between different networks and services. The other challenge in the network design is the compatibility of network equipment manufactured by different vendors. The integration of heterogeneous network with the further deployment of 4th Generation (4G) is going to signify the aforementioned challenges even more dramatically. To handle the growing complexity, 3GPP has designed a cellular network that supports interoperability of such multi-network, multi-service, multi-vendor, and heterogeneous environments.

(Olsson et al., 2013) book highlights the latest aspects in the mobile network evolution. Also, the book provides a high-level illustration of 3GPP network architecture, see Figure 2. 3GPP defines the following components of the network: User Equipment (UE), Radio Access Network (RAN), core network, and external network. UE is a handset or modem, which is placed at the beginning of the system. A mobile user makes voice calls, consumes media resources, or generates data traffic by means of this device. UE supports one or more Radio Access Technologies (RATs) in order to connect to a core network via RAN. RAN is responsible for radio related functionality such as radio access, mobility management, and radio resource management. Handovers between different base stations, frequency carriers, or RATs are performed by RAN elements. The heart of a cellular network is a core network. It handles and redirects a subscriber's call within the current network or to another operator's network. Also the core network performs charging mechanisms, controls data connection to external networks, and implements other management procedures. An external network consists of Circuit-Switched (CS) and Packet-Switched (PS) networks. These technologies transmit information in conceptually different ways. A CS network establishes a dedicated channel, which is unavailable for other transmissions. The dedicated channel is used for voice calls and telephony services, like integrated services digital network, see (Wikipedia, 2015b). A PS approach splits a user message into a number of smaller pieces, called packets. These packets achieve the destination point via different routes. Afterward, the original message is assembled on the receiver's side. The Internet is an example of PS network.

RAN is comprised of network elements, which are related to different generations of RATs. Depending on the access technology the network elements perform distinct functionalities. The common part is a base station, which is usually deployed on a building or a post. The base station transmits and receives a radio signal via sector antennas. The coverage area of a signal from one antenna is referred to as a cell. Multiple distributed base stations split the target coverage area into cells. Cellular structure is the underlying principle of modern radio networks. This principle allows localizing signal interference and an efficient utilization of radio resources. For instance, if a base station allocates a mobile subscriber a band on a frequency of 900 MHz, then the transmitted signal power attenuates in direct proportion to the square of the distance. Thus, a distant enough base station can reuse the same frequency band on 900 MHz because the interference from the original base station is negligible. GSM EDGE Radio Access Network (GERAN) and Universal Terrestrial Radio Access Network (UTRAN)

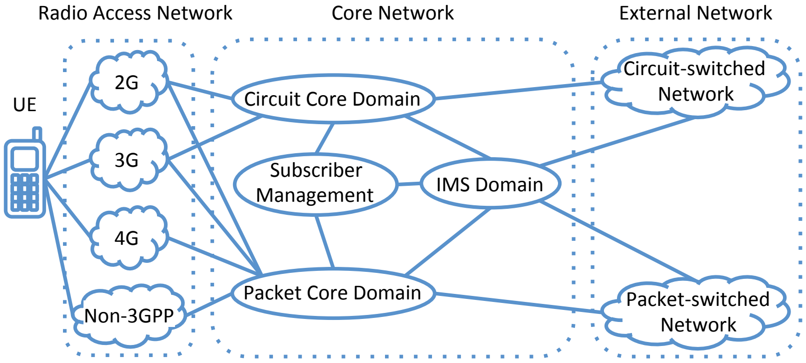


FIGURE 2 3GPP radio network architecture

are the common names for RANs of 2nd Generation (2G) and 3rd Generation (3G) respectively. The management of multiple base stations – Base Transceiver Stations (BTSs) – is carried out by a Base Station Controller (BSC) in GERAN. Node B (NB) is a term for a base station in UTRAN. One Radio Network Controller (RNC) coordinates multiple NBs in 3G radio network. 4G radio access network is called Evolved UTRAN (E-UTRAN) and features flatter structure, than the previous generations. The complete functionality of E-UTRAN is placed in one network element referred to as evolved Node B (eNB). Such restructuration achieves lower signal latency. Non-3GPP is a packet access network, which is unspecified by 3GPP. Wireless Fidelity (Wi-Fi), see (Wikipedia, 2015c), or Bluetooth, see (Wikipedia, 2015a), are well known examples of non-3GPP wireless standards.

A core network includes a subscriber management component and circuit core, packet core and Internet protocol based Multimedia Subsystem (IMS) domains, as shown in Figure 2. These network domains are connected via interfaces, which are completely defined by 3GPP. Subscriber management element consolidates and coordinates all subscribers' information. 2G and 3G allow access to the external network via both the circuit core and packet core domains. LTE and non-3GPP technologies do not support circuit-switch services. The complete functionality of these technologies is provided over PS connections. IMS is an Internet Protocol (IP) based framework for providing multimedia services. IP allows achieving more flexible network architecture; hence, the overall operational expenses decrease. Besides, IMS allows network operators to increase revenues by providing a wider range of services. For instance, this domain is responsible for establishing sessions between a user handset and the Internet. To find a detailed description of IMS refer to (3GPP, 2015c) 3GPP technical specification.

3GPP explains general requirements to the elements of network architecture without detailed specifications. Interfaces between the elements, on contrary, are unambiguously defined. This approach serves a double purpose. On one side,

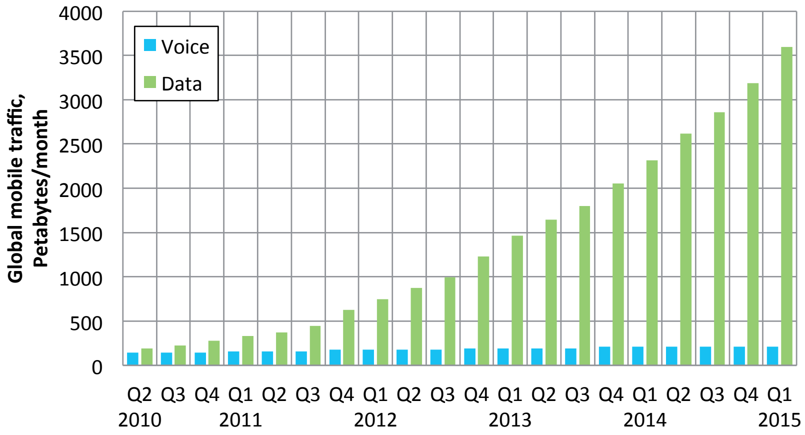


FIGURE 3 Global monthly mobile data and voice traffic

network manufacturers are enabled to develop the internal functionality of their equipment. On the other side, 3GPP standards guarantee interoperability between the elements of a complex radio network.

2.2 Outlook and Economics of Radio Networks

Ericsson, Cisco, and other companies reveal the constant growth of global mobile data and voice traffic in their annual analytic reports (Ericsson, 2015; Cisco, 2015). This process would be impossible without the appearance of high-end user devices such as smartphones and the development of high-speed wireless technologies. In Figure 3 Ericsson analysis illustrates the exponential growth of global data usage. As you can see, voice calls and messaging traffic slightly increases, while its relative share is rapidly declining. For instance, in the fourth quarter of 2010 mobile data exceeded twice that of voice traffic. At the beginning of 2015 this ratio increased up to seventeen times.

As for the upcoming five years, Ericsson report forecasts the substantial growth of generated mobile traffic, see Figure 4. This trend is caused by the gradually increasing number of mobile subscriptions all over the Globe. However, the key factors of growing data usage differ from region to region. In mature markets, the number of consumed bytes per individual has a significant role. This is driven by the high development of wireless technologies in these regions. The other reason is one person generates traffic from multiple devices, such as tablets, mobile computers, and mobile routers. In developing countries, the growth of mobile data usage is largely caused by new subscribers, as long as smartphones are becoming more affordable and LTE achieves a bigger share in regions. As it is shown on the Figure 4, Ericsson anticipates that during the upcoming five years

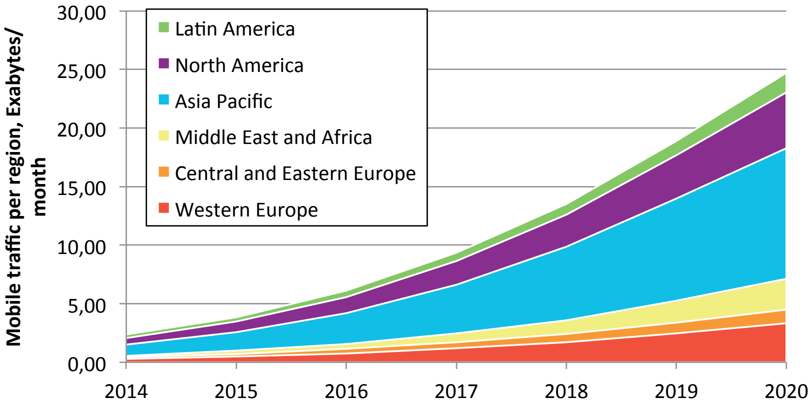


FIGURE 4 Global mobile traffic per region

mobile traffic consumption will exhibit a moderate increase in the economically developed markets such as Central and Eastern Europe. In opposite, mobile data usage in the developing Asia Pacific region will experience significant growth and exceed half of the traffic generated in the world.

In modern cellular networks, smartphones constitute the main share of the connected devices. The appearance of 5th Generation (5G) technology in 2020 will fulfill increasing demands for short latency, high-speed connection, and robustness. Advanced capabilities are going to significantly expand the range of business models. The major applications are expected to be machine-to-machine and device-to-device type communications. For instance, 5G technology will enable communication between cars and the control of smart houses. Therefore, we anticipate an even more dramatic increase in the number of connected devices and data consumption after 2020.

To fulfill ever growing data usage demands cellular technologies are moving along the way of widening frequency bands, improving the spectral efficiency, and utilization of MIMO technology. The other approach is to make use of the fundamental principle of cellular networks – frequency reuse. The idea implies the splitting of the coverage area into smaller areas or cells. It localizes interference, allows less transmission power from base stations, and guarantees higher network robustness. The cellular concept formed the basis of the principle referred to as heterogeneous networks. The fact is that the majority of mobile traffic is generated indoors, i.e. within buildings and malls. Such locations require a macro-cell to an extra power in order to transmit the signal through the concrete walls. Deployment of small cells – also known as pico-cells, micro-cells, or femto-cells – within the buildings will enhance the performance of the network in target areas while freeing the resources of a macro-cell. The disadvantage of such a multi-layer approach is the massive infrastructure of cellular networks.

The constant cellular evolution leads to the growing complexity of network infrastructure and increases operational expenses. Mobile operators already support multiple standards, such as GSM, High Speed Packet Access (HSPA), and LTE. As (Pingping et al., 2013) points out, modern cellular networks are faced with the following three major challenges. The first is the efficient utilization of radio resources. Then multi-RAT networks need to guarantee a decent quality of wireless service regardless of the used RAT. The third task is to optimize and simplify the coordination of complex cellular network architecture and provide a seamless connection between different RATs. Thus, the management of multi-RAT networks is a nontrivial task, which involves highly qualified human labor. As a result, maintenance and operational expenditures remain high and increase with further network evolution.

The economic component of network deployment and maintenance is a complex term. The purchase price of network equipment depends upon the vendor's name and embedded proprietary functionality. However, this price is not the main financial concern for mobile operators. Radio communication providers deal with Total Cost of Ownership (TCO), which is the estimation of the total expenses for a certain period of time. Usually, network equipment becomes obsolete and needs to be updated in five years. Therefore, TCO is analyzed over a duration of five years. TCO combines three components, which are Capital expenses (Capex), Implementation expenses (Impex), and Operating expenses (Opex). Capex includes the market price of network equipment plus the necessary materials and tools. In the case of a base station deployment, Capex also covers network planning and other relevant activities. The actual installation of network equipment is related to Impex. In the base station example, it includes transportation and installation expenses. Besides, all the activities concerning the initial configuration and launch are covered by Impex. Opex mainly consists of operation and maintenance costs. These cover rental and electricity bills. Also, Opex includes a bunch of activities like network optimization, replacement of broken devices, necessary measurement campaigns, and other relevant aspects. (Cassidian, 2013) points out that Opex is a substantial value, which may reach 80% of Capex.

2.3 Radio Networks Operation

Radio network operation is a continuous process aimed at monitoring and improving a network performance. The process features centralized management architecture, which is referred to as Operation, Administration, and Maintenance (OAM). As (Hämäläinen et al., 2012) shows, the operation tasks are split between network planning, network monitoring, and network optimization centers, see Figure 5. Although, network operators intend to automate these procedures, human supervision is still required at each level. Network monitoring center regulates Performance Monitoring (PM), Fault Management (FM), and Configuration Management (CM) processes. PM collects measurements from different levels of a radio network. The measurements are referred to as Key Performance Indica-

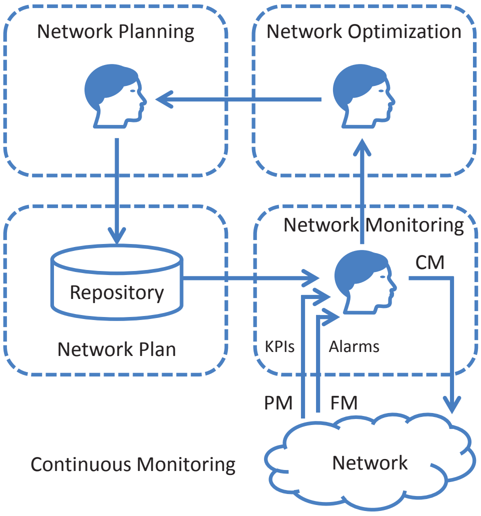


FIGURE 5 Radio networks operation

tors (KPIs). The FM process detects failures and triggers alarms. Failure detection involves the analysis of KPIs and depends on the experience of a human operator. CM adjusts the network parameters for network optimization or troubleshooting. As for planning and optimization, these departments focus on the improvement of coverage and capacity characteristics while reducing Opex. This is achieved by analyzing data received from the network monitoring center.

The planning department is engaged in the tasks related to rolling out a network or the deployment of a base station. Important aspects in this process are the policies of a communication provider regarding signal coverage, network capacity, and service quality. Network validation follows the planning step. The performance of newly deployed network is usually analyzed during drive testing measurements. A more detailed overview of the planning and optimization processes can be found in (Laiho et al., 2005).

Continuous network monitoring is needed to maintain and guarantee a target level of provided service. For these purposes, the PM process collects KPI measurements by multiple ways (Laiho et al., 2007). KPIs statistics can be roughly divided into field, user equipment, and network equipment based measurements. Drive tests are field measurements that are intended for performance and signal coverage monitoring. The collected data is very important for network optimization and constitutes a significant part of Opex. Field measurements are carried out for troubleshooting or when radio environment changes due to major construction. Note that drive testing campaigns are conducted by means of a vehicle, which is equipped with measuring devices. Therefore, mobile operators are missing the picture of indoor network performance. The indoor radio measurements can be carried out by phone applications and sent to network management centers. These user-level statistics mirror the provided service quality and radio network performance. In its turn, a network equipment log contains detailed information, which is automatically collected during network operations. These measurements allow the monitoring of conditions of network elements and detecting changes. In the case of a hardware or software failure, detailed information from network equipment logs serves for problem solving and root cause analysis.

Failure management or troubleshooting is placed among network monitoring tasks. Failure management functions encompass timely detection of network faults or performance degradation and further reactions. Malfunctioning is detected by direct network alarms, via the analysis of KPIs, or after multiple user complaints. The next phase initiates necessary failure compensation, root cause analysis, and repair actions. In the case of cell outage and the consequent arising coverage issues, the most efficient compensation action is the adjustment of antenna tilts of the neighboring base stations, see (Amirijoo et al., 2011). Root cause analysis involves the detailed study of KPIs. Sometimes, a network operator may even carry out drive tests if extra measurements are needed. The goal of the repair work is to completely fix the problem. This action may require adjusting parameters, repair, or even replacement of the network equipment.

The network optimization center aims to improve the systems performance. The process is carried out via the continuous optimization of the network equipment's configuration. Firstly, KPIs are analyzed and the need for network replanning is assessed. After the reconfiguration, performance indicators evaluate the efficiency of the optimization procedure. The necessity for the continuous network optimization is explained by the following reasons. These are imperfect planning of newly deployed network and radio environmental changes due to construction.

2.4 Self-Organizing Networks Concept

High needs towards the automation of radio networks management are dictated by the increasing complexity of the network architecture and demands for cost reductions. In 2007 and 2008 the alliance of mobile operators NGMN introduced a set of network use cases where the automation is desirable. The use cases are related to typical network problems and are divided into planning, deployment, optimization, and maintenance groups. NGMN activities and network automation research projects had an impact on the 3GPP standardization process. As a result, 3GPP introduced the SON concept, which aims to optimize network management. In technical report (3GPP, 2011), 3GPP published a set of SON use cases and solutions. They include coverage and capacity optimization, interference reduction, energy savings, and others. The SON concept and its requirements have been under development since Release 8 onwards, see (3GPP, 2014b).

SON concept consists of three solutions, namely self-configuration, self-optimization, and self-healing (Hämäläinen et al., 2012). Self-configuration refers to the automated configuration of the newly established networks. Self-optimization mechanism automatically adjusts network parameters in order to guarantee high quality of service and reduce maintenance costs. The purpose of the self-healing mechanism is to automatically detect and address network operation problems, avoiding significant impact on subscribers' experience and reducing operational expenses.

The overall goal of SON is to automate network operation and to reduce or completely exclude manual work. On one hand, the exclusion of manual operations minimizes the risk of human error in everyday tedious tasks. On the other hand, the automation of low-level management functions saves Opex and enables network operators to focus on more significant tasks such as the design of network policies. SON vision of the human role is providing high-level instructions and monitoring the network condition via KPIs.

The implementation of SON significantly impacts radio network architecture. Changes of network conditions may require either immediate actions or delicate analysis. For example, eNB SON functionality can immediately tackle the problem of a cell outage and a consequent coverage hole by increasing signal powers of neighboring cells. Besides, SON functions can be implemented in a core network. 3GPP is responsible for the standardization of radio measurements, interfaces between network modules, and the fundamentals of self-management system. The actual logic of SON algorithms is the responsibility of network equipment vendors. As a result, cellular networks will leverage a mixture of proprietary SON mechanisms within the unified hybrid SON architecture.

The scientific community demonstrates a high interest in the SON concept. The list of related research projects includes Celtic GANDALF (2005-2006) (Altman, 2006), EU FP7 E3 (2008-2009) (König, 2009), EU FP7 SOCRATES (2008-2011) (Kürner and et al., 2010), and FP7 SEMAFOUR (2012-2015) (Willcock, 2015). GANDALF project tackles automation and optimization problems in the environment of 2G and 3G networks. Based on the self-organization principles, E3 focuses on the design and development of wireless network architecture for beyond 3G systems. SOCRATES focuses on self-optimization, self-configuration, and self-healing methods in LTE radio networks. SEMAFOUR is the most recent SON related research that finished in August 2015.

SEMAFOUR is a European project, which is the result of the collaboration of research institutions and major telecommunication companies towards the automation of network planning and maintenance, (Willcock, 2015). The main trigger of the project was the numerous number of SON solutions, which had been under 3GPP standardization process or already in specifications. These solutions had an unsystematic nature and mostly tackled problems of RAT layers. Under the pressure of potential difficulties to handle numerous SON solutions the SEMAFOUR project was established. The project aims at the development of an integrated self-management scheme that facilitates network management and reduces Opex. To achieve this aim SEMAFOUR formulates two major goals. The first goal is to come up with common SON functions that support network configuration and maintenance across different RATs and cell layers. The second one is to build a unified SON framework that enables an operator to fulfill performance requirements for a heterogeneous network condition via developed common SON functions.

As a result, the SEMAFOUR project developed a unified self-management system for the automated planning and maintenance of multi-RAT and multi-layer networks, (Hahn and et al., 2015). Elaborating on major achievements, the project proposes a transparent and effective SON mechanism, impacts on the 3GPP standardization process, and improves network performance and spectral usage. The designed solutions were examined via extensive experiments on the Hannover simulation scenario, (Rose et al., 2013). Simulations demonstrate multiple SON approaches developed within SEMAFOUR. For example, multi-layer traffic steering approach enables to offload LTE traffic via a Wi-Fi network. User steering approach predicts user mobility behavior and avoids unnecessary Handovers (HOs). Sectorisation employs antenna beam steering to localize interference and save energy. Dynamic spectrum allocation optimizes spectrum usage across multiple RATs. Also, SEMFOUR addresses the problem of conflict avoidance between SON functions.

Based on the conducted research and feedback from mobile operators, SEMAFOUR provides future vision and key challenges for a unified self-management system, see (Eisenblatter et al., 2013). Key challenges are covered by the following three topics. First, conflict diagnosis and conflict resolution between the different SON functions executed in parallel. Second topic focuses on the development of a mapping process from high-level objectives of a network operation into particular SON functions. And third one is devoted to the interpretation of global network policies through high-level objectives.

Communication technology companies already present business solutions for self-management of radio networks. Nokia is one of the biggest players on the market of SON solutions. Nokia Eden-NET solution presents automation for network configuration, optimization, and healing operations. For more details about implemented multi-vendor and multi-RAT functions visit (Nokia, 2015).

2.5 Minimization of Drive Tests

Drive tests are field measurements performed by mobile operators and aimed at gathering network statistics about coverage, capacity, and quality of service. Drive tests allow the finding of weak parts of the network and ways for performance improvement. Hence, field measurements are essential for network planning, optimization, and troubleshooting processes. Technically, drive tests are carried out by means of special vehicles. The vehicles transport the measuring equipment along the streets and roads. Global Positioning System (GPS) receivers associate conducted radio measurements with actual geographic locations. As a result, a mobile operator obtains coverage map and network condition as a function of coordinates. Despite its benefits, a drive testing campaign features noticeable drawbacks. Drive tests miss radio signal measurements everywhere: pedestrian zones, buildings, and other car inaccessible spots are out of scope. Also, drives tests constitute a valuable part of Opex because the tests require manual work of qualified specialists and entail expenses related to the usage of complex measuring devices. Optimization of Opex and limited reachability are the ruling factors for the innovation in collecting field measurements.

3GPP introduced MDT functionality in Release 10 (3GPP, 2014a). Release 10 responded to high demands of network operators for the automation of driven tests (NGMN, 2007, 2008). The underlying idea is to employ user handsets for the collection of network performance information. In 2010 NGMN alliance placed MDT among the top 10 solutions to ensure network operation efficiency (NGMN, 2010). In opposite to traditional field measurements, MDT allows for the collection of network data throughout the whole map, including indoor and outdoor locations.

The original emphasis of the MDT study was made on coverage optimization, see (3GPP, 2010). Coverage represents signal strength at a particular point. The space distribution of signal strength – coverage map – is a crucial aspect for the assessment of network performance because it directly influences on service quality. From a customer's point of view, coverage maps are easy to interpret and often play a key role in choosing a communication provider. Thus, the development of solutions for coverage optimization received a high priority. The common solution implies the detailed visualization of signal strength distribution over the network area. Signal strength distribution facilitates the detection of areas with excessive, weak, or no coverage.

The further development of MDT assisted functionality involves mobility optimization, capacity optimization, and quality of service verification (Hämäläinen et al., 2012). An example of mobility solutions is the optimization of handover parameters. At first, the analysis of MDT location information provides users' mobility patterns. Then, mobility patterns enable better configuration of handover settings. Also, MDT facilitates capacity planning. The first step is to identify the relationship between traffic flow and geographical location. The relationship spots areas with high or increasing traffic usage. Then the deployment of new base stations in these areas offloads busy network cells and improves scheduling of radio resources. Besides, MDT verifies the quality of service. Verification gives the comparison between expected and actual quality indicators. Significant discrepancies of the indicators may force a mobile operator to reconsider scheduling policies.

2.5.1 MDT Architecture

MDT architecture is built on top of a subscriber and equipment trace concept, see (Hämäläinen et al., 2012). Tracing enables to keep track of all user activities in a cellular network, see (3GPP, 2015d). 3GPP distinguishes two ways of tracing activation: signaling based and management based. Signaling based trace sessions monitor all activities of a chosen handset. Monitoring continues if the handset moves throughout the whole network area. Management based tracing captures all users within a chosen area of several cells. This type of tracing is activated when a new user comes into the specified area and is interrupted when a user leaves the area. Both tracing approaches formed the basis for subscription based MDT and area based MDT. Subscription based mechanism collects radio network measurements of a particular user equipment. The user equipment is identified by International Mobile Station Equipment Identity (IMEI) in a core network. Area based MDT retrieves information of network conditions from all active calls within a specified area. Both MDT mechanisms can be activated on a mobile device only after obtaining a permission from a subscriber. The basis of the current research is the analysis of measurements from multiple devices. Therefore, the developed algorithm implies the utilization of area based MDT.

3GPP implements measurement and reporting procedures by immediate MDT and logged MDT approaches, see (3GPP, 2015e). Immediate MDT or so called online approach stands for the information collection from a mobile terminal, when it is synchronized with RAN. This operating mode is referred to as Radio Resource Control (RRC) connected mode. If a user device is in RRC idle mode – the device is disconnected from the network – then field measurements can be implemented over logged MDT or so called offline approach.

Immediate MDT and logged MDT approaches can be utilized for the management of area based MDT. Figure 6 illustrates the architecture of both approaches at high-level. OAM performs the original configuration of MDT settings and distributes it by means of the trace functionality to RAN elements (3GPP, 2015e). RAN elements manage MDT settings on end-user devices via a control plane. RNC nodes and eNBs configure and retrieve radio measurements from user handsets in 3G and 4G networks correspondingly. If a handset is in RRC connected mode, then user-level data is reported as soon as a triggering condition is met. Upon retrieving the data, RAN node adds a time stamp. Logged MDT approach specifies the data collection process for a handset in idle mode. MDT configuration is carried out when the handset is in connected state. While being offline the handset measures network conditions and stores relevant information in the memory. As soon as the user equipment returns to a connected state, the stored information is reported to the RAN node. RAN nodes retrieve, process, and send MDT measurements from mobile terminals to Trace Collection Entity (TCE).

2.5.2 MDT Measurements

MDT configuration parameters extend the subscriber and equipment trace concept, see (Hämäläinen et al., 2012). The configuration parameters define conditions for data collection and reporting processes as well as the list of radio measurements. The existing immediate MDT and logged MDT data management schemes feature similarities, but also possess significant differences.

Data collection and reporting conditions are standardized in (3GPP, 2015e). User-level measurements are triggered by either a network event or a periodic

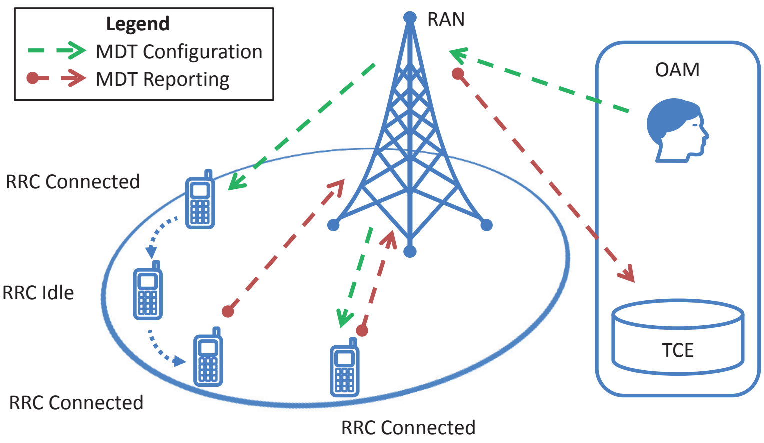


FIGURE 6 MDT architecture

counter. Only online MDT approach performs event-based measurements. As soon as an event-based measurement is taken, it is immediately sent to a base station. The base station stores the received data item and assigns a time stamp to it. Periodic-based measurements are carried out by both immediate MDT and logged MDT. During a logged MDT session a user device stores the data with a time stamp. The periodicity of measurements and validity of MDT configuration parameters are defined by logging interval and logging duration. In the case of immediate MDT, similar parameters are referred to as report interval and report amount. For example, in LTE networks the time between two sequential reports varies from 120 milliseconds up to 1 hour.

The current thesis makes use of immediate MDT scheme with the data collection process triggered by network events, see (3GPP, 2015b). The reported measurements represent timely ordered sequences, because a base station assigns a time stamp upon receiving a data item. The sequential analysis of network events underlies the proposed cell outage detection approach, see Section 3.5. In simulations, we considered the following set of network events: ”A2 Enter”, ”A2 Leave”, ”A3”, ”HO Command Received”, ”HO Command Complete”, ”Physical Layer Problem (PLP)”, and ”Radio Link Failure (RLF)”, see (3GPP, 2015b). Triggering conditions for LTE events ”A2 Enter” and ”A2 Leave” are the Reference Signal Received Power (RSRP) of the serving cell becomes worse or better than a specified threshold correspondingly. An ”A3” event occurs when the signal power of a neighboring cell becomes offset better than the power of a serving cell. ”HO Command Received” and ”HO Command Complete” events are related to a handover procedure. A sequence of ”out-of-sync” indications triggers a ”PLP” event and starts a T310 or T313 timer. ”RLF” is reported upon expiry of these timers. An ”out-of-sync” trigger criteria is either the quality of the dedicated radio channel becoming worse than the threshold or the last 20 transport blocks are received incorrectly. As soon as one of the events occurs, a user device performs a measurement procedure. An MDT parameter ”List of Measurements” indicates what kind of measurements should be taken, see (3GPP, 2015e). This list includes such network characteristics as data volume, throughput, and signal power. In the case of an LTE network the following measurements can be scheduled on a user device:

- RSRP and Reference Signal Received Quality (RSRQ) measurement
- Power headroom
- Received interference power
- Data volume separately for downlink and uplink connections
- Scheduled IP throughput separately for downlink and uplink connections
- Any combination of aforementioned items

(3GPP, 2014c) enhances MDT report with information about RLFs and Handover Failures (HOFs) experienced by a user equipment.

With the introduction of Release 11, 3GPP enriches MDT with location measurement of a user device, see (Johansson et al., 2012). (Ericsson, 2011) lists LTE positioning methods. The location can be presented as cell Identification (ID), radio frequency fingerprinting, or geographical coordinates. Assisted Global Positioning System (A-GPS) provides the highest accuracy for outdoor positioning. However, A-GPS is not supported by all handsets and features relatively high power consumption.

MDT enables a mobile operator to employ field measurements in order to improve a networks performance and reduce Opex. However, MDT technology features noticeable drawbacks. Huge amount of field measurements pose additional challenges related to data management. Operators are needed to build data centers, optimize information flows, and perform intelligent data analysis. Besides, data collection and reporting processes impact on a user device battery life, which is significantly valuable for modern smartphones. Furthermore, MDT is not capable of completely replacing traditional drive tests, see (Baumann, 2014). For example, a rolled out cellular network needs validation before being put into operation.

2.6 Cell Outage Detection

Automatic detection of malfunctioning cells is part of the self-healing mechanism, see (Hämäläinen et al., 2012). In general, cell outage takes place due multiple reasons: hardware or software failures, external failures of power supply, erroneous configuration or even environmental changes. For instance, cell failure can be caused by improper network planning, misconfiguration of multi-vendor equipment, erroneous antenna tilt, or a software bug. Usually, the detection of a malfunctioning cell is performed via the analysis of alarms, KPIs, or multiple customer complaints.

Depending on the ability to provide a service, cell outage is roughly classified into three types: degraded, crippled, and catatonic, see (Cheung et al., 2006).

The most difficult to detect is a degraded cell. A degraded cell is able to carry network traffic, but not as much as a properly functioning cell. A crippled cell is characterized by severely degraded performance, but still provides a service to a few users. A cell which experiences complete inoperability is referred to as a catatonic cell.

A sleeping cell is a cell outage type, which is invisible for network operators via traditional alarms (Hämäläinen et al., 2012). This peculiarity makes a sleeping cell problem a very challenging task. Usually, this special type of a cell outage becomes visible after drive tests, via the detailed analysis of KPIs, or even due to user complaints. (Xiaojin, 2012) distinguishes the following failure scenarios resulting in the appearance of a sleeping cell. These are physical channel failures, Cell/LTE base station failures, and transport channel failures.

With the introduction of MDT functionality, it became possible to collect user-level statistics from regular user devices. Such an approach provides much larger volumes of data to be analyzed and poses problems related to data management schemes. In order to process enormous user-level measurements data mining techniques look as a very promising tool. One of the major use cases is the human free detection of cell outage. Corresponding anomaly detection algorithms are proposed and analyzed in the number of papers.

Advanced data mining and machine learning techniques are carried out for the detection of cell outage in publications (Chernogorov et al., 2011; Turkka et al., 2012). The approach in (Chernogorov et al., 2011) analyzes the data set of signal strength and quality measurements reported by mobile terminals. These measurements contain both serving and neighboring cell measurements in an LTE network according to MDT. Advanced data mining techniques exhibit the ability to reveal latent abnormal behavior in the high dimensional data set of measurements. Thus, the developed approach pinpoints a problematic cell in the network. In (Turkka et al., 2012) the same approach is extended by targeting to find similarities between periodical measurements and reports related to failures happening at the radio link. The investigated solution enables to substantially increase the number of samples, which indicate the existence of a problem in specific cells. Thus, more reliable and fast detection is achieved.

In a series of scientific articles Nokia Siemens Networks (NSN) researchers solely and later in a cooperation with a group from Stanford Research Institute (SRI) address the problem of automatic cell outage detection and root cause analysis in LTE networks. The key outcome is the design and development of anomaly detection and diagnosis framework for mobile network operators. The underlying idea is the analysis of channel quality degradation through cell-wise profiles of network KPIs. If the deviation of a KPI profile from normal behavior is detected, then an alarm is raised. The deviation may be caused by the degradation of cell performance or a cell outage; hence, even small deteriorations can be identified. When the framework detects a malfunctioning cell, then root cause analysis is performed. Root cause analysis involves an automatic investigation of problem KPIs and diagnosis regarding failure reasons. To be fully automated the diagnosis process learns numerous decisions of human experts.

In (Nováczki and Szilágyi, 2011) Nováczki and Szilágyi utilize Channel Quality Indicators (CQIs) reported by user devices as KPIs. The designed framework is evaluated in simulated LTE environment. Paper (Szilágyi and Nováczki, 2012) considers three KPIs and tackles arising problems regarding fault management. Nováczki and Szilágyi utilize 12 KPIs and demonstrate the capabilities of their intelligent system on 70 cells in live 3G network over 180 days, see (Nováczki and Szilágyi, 2012). All further investigations regarding the automatic detection and diagnosis framework are conducted on real 3G network data. (Nováczki, 2013) makes one more step in the improvement of the profile and anomaly detection modules of the framework. (Ciocarlie et al., 2013) is the first collaborative result of NSN and SRI. The paper proposes an ensemble method that assigns weights and combines different anomaly detection algorithms, which include an empirical cumulative distribution function, support vector machine, and autoregressive integrated moving average. With the optimized weights the ensemble method outperforms any other single algorithm. (Ciocarlie et al., 2014b) demonstrates the feasibility of the framework implementation in a real network composed of 10 cells and estimates the execution time of key processes. As for training and testing of the framework, the corresponding times are insufficient. The detection delays are estimated in hours. The practical deployment involves configuration of parameters like observation window size, acceptable false rates, and others. (Ciocarlie et al., 2014a) is the next step towards the practical deployment of the intelligent framework. The paper introduces a platform that enables a network operator to visualize and adjust modules of the developed anomaly detection and diagnosis framework. Based on the designed framework, (Ciocarlie et al., 2014a) proposes SON verification system. The system aims to coordinate and avoid possible conflicts of automatic network management processes.

An important factor for data-driven approaches is the amount of available data. During nighttime or in rural areas the number of users is significantly smaller than during the daytime in a city. In (Mueller et al., 2008) Muller et al. consider the scarcity factor for the detection of antenna gain failure. The proposed algorithm employs statistical classification techniques, which are utilized to construct graphs from UE reports of neighboring cell patterns. Changes in the constructed graphs served as indicators of a cell outage. The algorithm is evaluated in a system level simulator for a macro-cell scenario. Although, the antenna failure detection approach shows good performance, a detection error rate increases with the decreasing number of active users.

Scarce user statistics is also a distinct feature of small or femto-cells. Small cells will be deployed by regular users without any planning, what potentially leads to the high probability of misconfiguration and malfunctioning of the network equipment. As long as the wide spread of small cells is expected in the near future, the development of appropriate self-healing algorithms is of high importance. (Wang et al., 2011) proposes a cell-aware transfer scheme for femto-cell outage detection. In order to overcome the problem of scarce data, neighboring femto-cells share information between each other. Each femto-cell weights the coming flow of information from surrounding units. The optimal choice of weights allows achieving higher accuracy than traditional methods. (Wang et al., 2014) introduces a cooperative femto-cell outage detection framework. The self-healing function employs trigger and detection mechanisms, which process spatially and temporally correlated user statistics. The trigger phase utilizes collaborative filtering to analyze correlated measurements and to initiate the detection mechanism. During the detection phase an original cooperative detection rule pinpoints the problem small cell. The developed framework demonstrates low communication overhead and high detection accuracy.

The original thesis tackles the problem of RACH failure, which belongs to the group of physical channel failures, see (Xiaojin, 2012). RACH is a constitutional part of LTE uplink transport and physical levels, see (3GPP, 2015a). According to this specification random access procedure is deployed for initial access, radio resource connection protocol re-establishment, uplink synchronization, and handover. When RACH failure occurs, it prevents a user device from establishing a connection or making a handover to the malfunctioning cell. However, the problem cell still carries traffic for those users who were connected before the failure has occurred. Since, only a few users notice the problem the cell belongs to impaired outage type. Due to mobility, users leave the malfunctioning area and the cell becomes catatonic. Typically, RACH problem would be detected only after a long observation time or after multiple user complaints. Thus, a timely detection of RACH failures is an important issue in LTE networks.

The detection of a sleeping cell relies on the detailed analysis of network KPIs (MDT measurements), because traditional alarms are not raised. Manual analysis of KPIs is a tedious, error-prone, and expensive process. This thesis employs data mining, which is an intelligent approach for data discovery. Usually, a data mining algorithm requires setting thresholds for decision making. Setting a threshold is a key aspect in the algorithm's adjustment process. A non-optimal threshold may lead to a high false alarm rate or high misdetection rate. A high false alarm rate requires additional network examinations, which increase Opex. A high misdetection rate decreases quality of service and is potentially dangerous for an operator's revenue.

3 DATA MINING

We live in a world, where a vast amount of digital data – also called big data – is collected from different sources every day. As reported by McKinsey Global Institute (Manyika et al., 2011) the analysis of such big data brings forth business competition to the next level of innovation and productivity. Therefore, the extraction and interpretation of hidden patterns in data sets is of great importance. Data mining is a modern tool that aims to discover meaningful knowledge from large data sets and prediction trends. Data mining offers not only a retrospective view on a business process, but also enables humans to develop a successful market strategy.

This chapter presents the origins of ”data mining” term. The next section considers problems that are addressed by data mining techniques. The following section illustrates knowledge discovery and databases processes. The chapter is concluded by a detailed description of developed cell outage detection framework.

3.1 Origins

”Data mining” originates in the 80s, when it was introduced and utilized within a research community (Coenen, 2011). The term itself is misleading, because it refers to the extraction or ”mining” of knowledge from data. Nevertheless, the ”data mining” term entered the scientists' jargon, unlike the more appropriate and verbose name ”knowledge mining from data”.

Piatetsky-Shapiro coined the term KDD, which was the name for KDD workshop in 1989. Nowadays, KDD refers to the whole analysis process introduced by Fayyad et al (Fayyad et al., 1996b,a). Data mining, in its turn, is defined as the component of KDD process and deals with the exploration of inner patterns in databases. Besides that KDD is concerned about the evaluation and interpretation of discovered patterns. Although, exact meanings of KDD and data mining terms differ from each other, often they are used interchangeably in the literature. In this dissertation I utilize KDD and data mining as synonyms, if it is not specified.

Hand et al in their book (Hand et al., 2001) give the following definition of data mining:

> *Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.*

Data mining computational methods find themselves in the intersection of classical statistics, artificial intelligence, and machine learning (Kotu and Deshpande, 2015). Data mining as a whole knowledge discovery process also involves many disciplines, such as databases, data cleaning, visualization, exploratory analysis, and performance evaluation. KDD leverages the solutions of these disciplines to benefit in different domains, such as business, medicine, and others.

3.2 Methods

Data mining techniques are categorized into supervised, semi-supervised, and unsupervised methods. Supervised method learns the relationship between the input items and their target values. Then the model predicts values for previously unknown data. Learning is completed with the help of a training data set, where the mapping between input records and output values is provided. Unlike the supervised approach, the unsupervised technique is fed by unknown training data and looks for the hidden rules and patterns between the input records. Also data scientists identify semi-supervised learning, which is similar to a supervised one. In this case, only a portion of training samples is labeled; hence, these samples have corresponding target values. Semi-supervised learning takes place when the data preparation – labeling of all training samples – is an expensive, time consuming, or even unrealizable task.

Data mining aims to discover existing patterns and predict possible trends. Thus, data mining methods can be roughly split into predictive and descriptive groups. As in (Kotu and Deshpande, 2015) I categorize the predictive group into classification, time series forecasting, regression, and anomaly detection methods. As for the descriptive group, it is broken into clustering and association analysis methods, see Figure 7.

Predictive:

- *Classification* aims to categorize unseen input data records into known classes. The assignment model or classifier learns from the training data set, where the relationship between records and classes is provided.
- *Time series forecasting* predicts the future value of a target function based on the previously observed measurements.

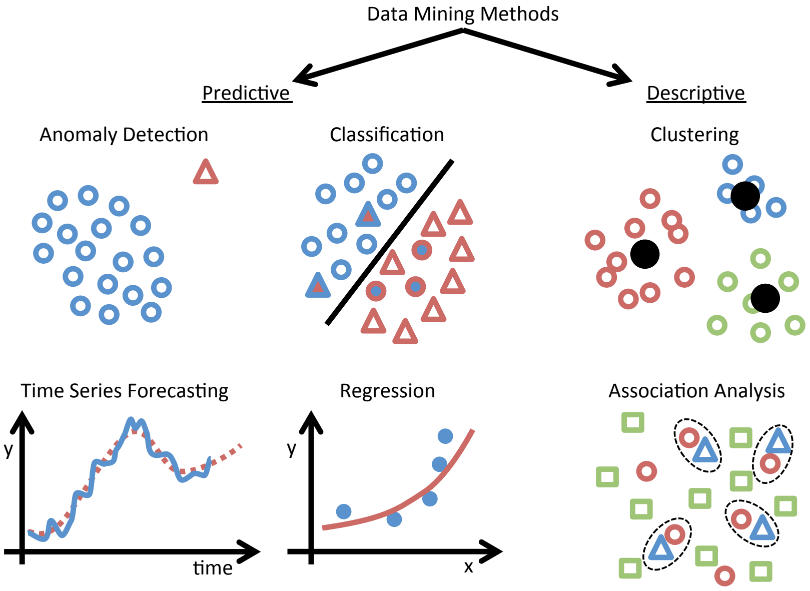


FIGURE 7 Data mining methods

- *Regression* aims to predict numerical values for input data records. The mapping function learns from the training data set, where the relationship between records and their values is known.
- *Anomaly detection* extracts points or outliers that are considerably different from the rest manifold of data points.

Descriptive:

- *Clustering* identifies manifolds of points – called clusters – with similar properties or behaviors.
- *Association analysis* discovers relationships between records within the same data set.

The research work conducted for the actual thesis addresses anomaly detection task. To solve the task we utilized supervised learning, because training data set consists of the entries of the same class – normal or anomaly free items. The developed data-driven solution is briefly described in the following sections.

3.3 Knowledge Discovery in Databases Process

Data mining is not only about computational methods that retrieve hidden dependencies, but it also embraces data preparation, interpretation of results, and other steps. This nontrivial process is described in different standards. In accordance with Knowledge Discovery nuggets (KDnuggets) survey (KDnuggets,

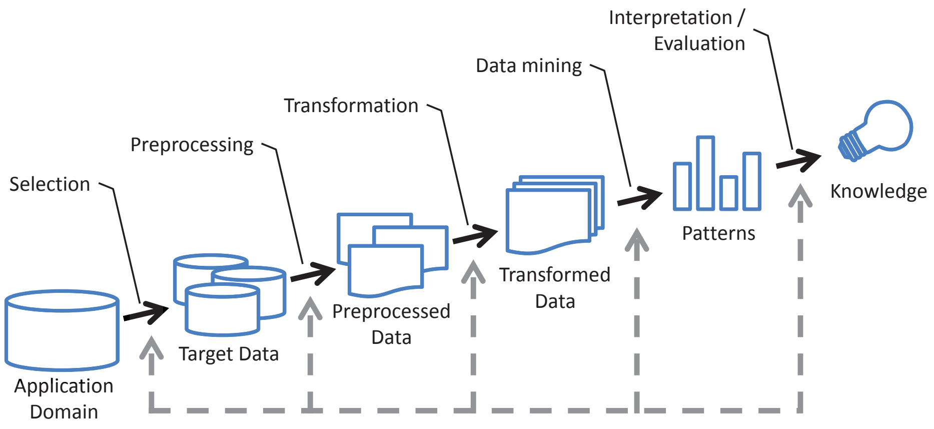


FIGURE 8 An overview of KDD process

2014) in 2014 the most popular methodologies utilized for data analysis are Cross-Industry Standard Process for Data Mining (CRISP-DM); Sample, Explore, Modify, Model, and Assess (SEMMA); and KDD. SEMMA and CRISP-DM are industrial standards. KDD is more traditional and academic in approach. Although, these methodologies are used by specialists from different spheres, they essentially follow the same principle of sequential and iterative data analysis, as it is shown in (Azevedo and Santos, 2008). In my research I leverage KDD process, because it is common knowledge discovery approach in academia.

KDD is a nontrivial process, which aims to reveal hidden patterns and extract useful information from data. Fayyad in his model (Fayyad et al., 1996a,b) defines data mining as a particular step in the whole exploratory analysis. Data mining refers to the set of computational methods that extract valuable patterns from original data. Additionally, KDD process is concerned about manipulation with massive data, scaling algorithms for better performance, proper interpretation of retrieved information, and human interaction with the overall process.

KDD process is a sequential analysis that includes the following steps: selection, preprocessing, transformation, data mining, and information interpretation steps, see Figure 8. However, this sequential knowledge extraction approach may involve iterations, because at any point the data analyst can change settings and repeat previous steps again. Thus, the basic KDD sequence may include loops. The knowledge exploration process starts with the development of necessary theoretical and practical background in the application domain. The understanding of relevant knowledge is important to achieve customer's goals. Here you have the short overview of the following KDD steps:

Selection It implies the selection of the target data set. The target data set refers to a set of items or data sources, which relationships and inner patterns are going to be discovered. Often the set of available data sources or records is redundant or uninformative. It is the work of data scientists to select the most relevant sources for the solution of the customer's problem.

Preprocessing The quality of the selected data is often inappropriate for further analysis, because of multiple reasons. Outliers (records featuring infeasible or abnormal values), missing variables, or high level of noise during the measurements require special data strategies – or preprocessing – for handling these issues. This list also comprises the merging of data from multiple sources and converting the values into correct formats.

Transformation The preprocessed data is described by a number of features. This imposes many difficulties to find the perfect combination of them, in order to solve the original problem. Transformation projects an original data into a low dimensional space – embedded space – and includes linear and nonlinear method. The reduced set of embedded features allows visual inspection and facilitates the further mining of knowledge.

Data mining The core element of the KDD process is the data mining phase, which includes several steps. Depending on the customer's goal, a specific data mining task is chosen – classification, anomaly detection, or other, see Section 3.2. Then a human specialist selects the most suitable calculation method for retrieval of information. This is done by matching the peculiarities of existing methods and ultimate goals. Finally, the chosen data mining algorithm is executed to search for underlying patterns and valuable knowledge.

Interpretation/Evaluation The final step of the KDD process is interpretation and evaluation of the retrieved information. This step involves techniques for visual analysis and a number of performance metrics. The correct interpretation of results is important, because it allows checking assumptions and tuning parameters of preceding KDD components.

Finally, the discovered hidden patterns and designed KDD algorithm may be integrated into an existing business model. The possible usage scenarios encompass reporting and prediction, optimization and automation of the business processes.

3.4 Modeling

In data mining, modeling is the creation of a model that represents relationship between data and underlying knowledge. Figure 9 illustrates the modeling of predictive and descriptive data mining; however, descriptive techniques lack testing data set. Evaluation step presents in predictive and descriptive techniques. The evaluation is needed to build a stable model and to avoid the overfitting and underfitting phenomena (Kotu and Deshpande, 2015). Overfitting is

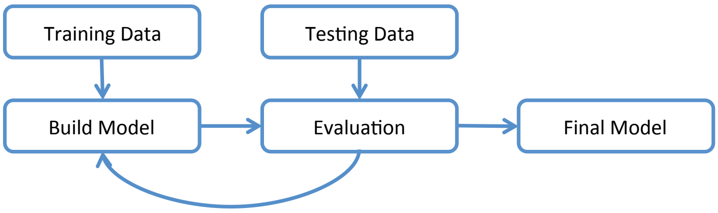


FIGURE 9 Data mining methods

referred to as memorization of training data, what causes significant under performance on unseen data. Underfitting occurs when the level of generalization of the build model is too high and the model fails to capture detail.

3.5 Cell Outage Detection Framework

The developed cell outage detection framework is illustrated in Figure 10. This approach leverages the KDD process (see Section 3.3) and successfully pinpoints a sleeping cell (see Section 2.6), as it is demonstrated in the included articles. The detection of the sleeping cell is carried out in two phases. The first phase is referred to as training. During the training process, the overall scheme is fed by a normal data set. Normal data is the data that reflects the regular – failure free – network behavior. The detection framework extracts the profile of normal data and learns the necessary decision boundaries and parameters. The second phase is testing. This time the framework processes a new data set and assigns outage scores to the network cells. These scores explain a current network state. The higher the outage score, the more abnormal the behavior exhibited by a particular cell. Besides, the proposed framework is able to make strict decisions about cells' operability, i.e. label cell as ″normal″ or ″broken″. The necessary decision threshold learns during the training phase.

In a nutshell, to identify a broken cell the proposed framework explores MDT logs, reported by Mobile Terminals (MTs). Data mining analysis extracts anomalous MTs – MTs, whose logs deviate from normal behavior. Then anomalous MTs are mapped on the network topology. The area with a high concentration of anomalous MTs is considered problematic.

The proposed cell outage detection framework is designed in accordance with KDD standard, which involves selection, preprocessing, transformation, data mining, and interpretation or evaluation steps. The selection step is straightforward and simply extracts target data – data to be processed – from MDT logs. A target data instance includes a time stamp and MT coordinate, information about triggered LTE network event and target cell. As long as data instances are ordered in time, our framework employs algorithms for the analysis of sequential data. The detailed description of the subsequent steps is given in the following paragraphs.

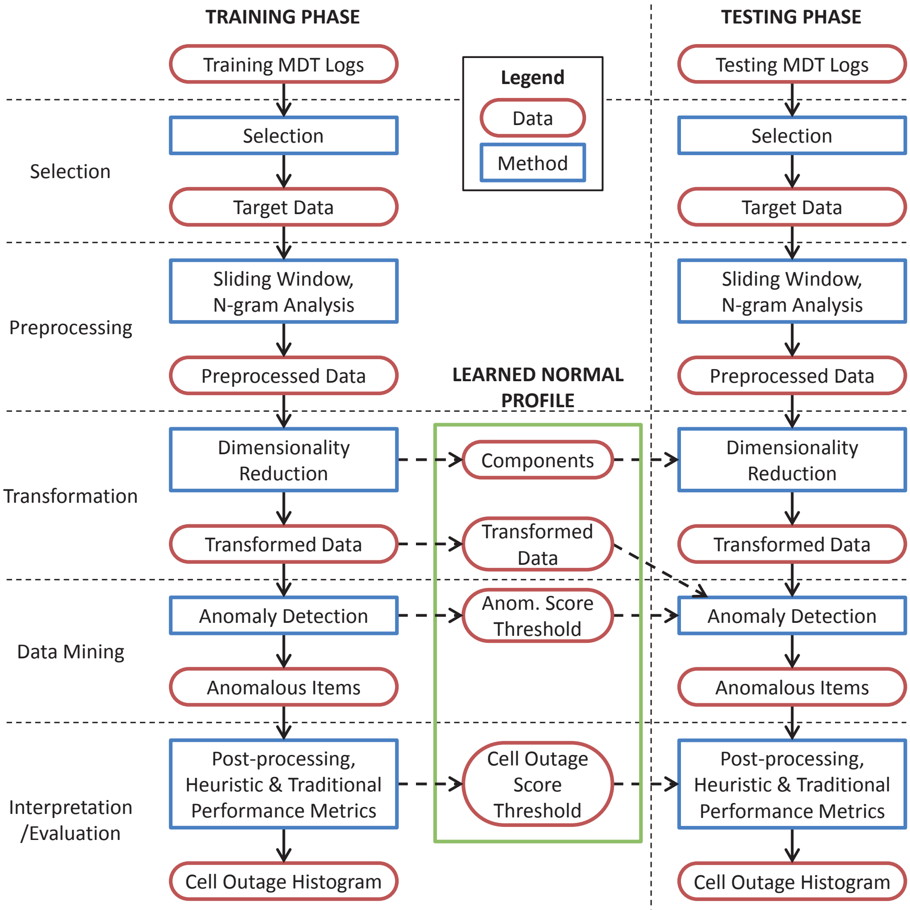


FIGURE 10 Cell outage detection framework

Preprocessing The preprocessing step is carried out during the training and testing phases. It employs sliding window and N-gram analysis methods for the processing of sequential data. Sliding window is a technique for slicing data instances into smaller pieces, which can be considered as separate units. When MT makes a call or data is transmitted, then a sequence of LTE network events is triggered, see Section 2.5.2. Due to the random nature of these connections the sequences have different lengths. Sliding window slices each sequence into subsequences by a segment – window – with a predefined size and step. Large window hides abnormal behavior of subsequences; whereas, the small one provides too much extensive granularity. Also, slicing performs length-wise normalization and localizes subsequences on the map. N-gram is a feature selection algorithm utilized for the analysis of sequential data. Sequential data is data with instances ordered in time or space. N-gram finds itself in many applications. For example, this algorithm is used for the analysis of whole-genome protein sequences (Ganapathiraju et al., 2002), for computer virus detection (Choi et al., 2011), and in many other fields.

N-gram analysis decomposes subsequences provided by sliding window method into small pieces. This new representation is stored in a feature vector. The feature vector conveys frequencies of how often each particular N-gram happens in the original subsequence. The number of elements of the feature vector is defined as all possible permutations of size N of unique instances in the considered data set. The higher N is chosen, the more elaborate the description of the original data is obtained. The detailed description of the original data is potentially beneficial for the performance of data mining algorithm. However, the increase of N raises both the dimensionality of the feature space and computational complexity. Thus, the choice of N is a tradeoff between the overall algorithm performance and computational efficiency. In this study $N = \{1, 2\}$. For the sake of clarity, let's apply N-gram to the string 'atadata'. Then 1-gram analysis of 'atadata' is

1-gram	a	d	t
Frequency	4	1	2

and 2-gram analysis of 'atadata' is

2-gram	aa	ad	at	da	dd	dt	ta	td	tt
Frequency	0	1	2	1	0	0	2	0	0

Transformation High dimensional data requires extra computational resources. To save the resources transformation carries out dimensionality reduction. Dimensionality reduction extracts the most meaningful features from the matrix provided by the preprocessing step. The designed data mining framework maps the manifold of points into a low dimensional space – embedded space – with the help of Minor Component Analysis (MCA) and model order selection algorithm Ration of Adjacent Eigenvalues (RAE). RAE chooses the actual components for embedded space. It is worth noting that If the KDD process uses 1-gram analysis, then the dimensionality reduction component is excluded.

Embedded space is composed of the subset of eigenvectors of covariance matrix of the training data set. MCA forms embedded space out of the eigenvectors with the smallest eigenvalues or minor components. The formal and elaborate derivation of eigenvectors and eigenvalues is proposed in (Josse et al., 2011). The splitting point between the smallest and the highest eigenvalues is chosen by RAE, see (Cong et al., 2012). RAE defines the splitting point as the index of the maximum value of the ratio of adjacent eigenvalues. Minor components learn once while training and utilized during training and testing phases.

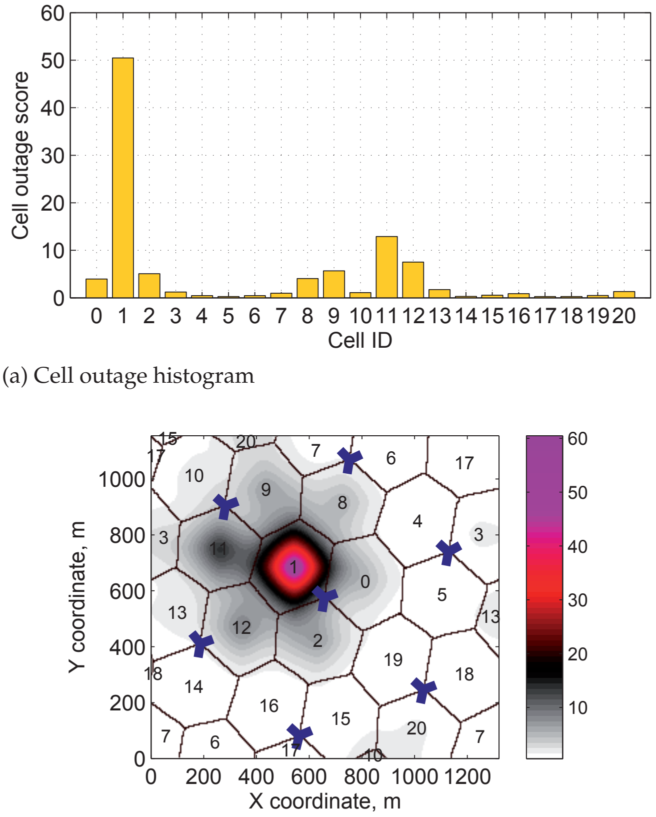
Data Mining The developed data-driven framework extracts anomalous subsequences or points from the data set; hence, anomaly detection task is solved. Firstly, a data mining algorithm assigns anomaly scores to points. Secondly, points having the largest anomaly scores are classified by a decision threshold as ”anomalous”, others are treated as ”normal”. Assuming that the normal – anomaly free – data set is provided, the supervised learning technique is carried out. The necessary decision threshold is learned during the training phase. The threshold is set such that n% of training points are labeled as ”normal”, while others are ”anomalous”. Then the threshold is applied to training data set.

In the conducted research I applied and compared different anomaly detection methods, because it is a key element of the KDD scheme, see Figure 10. These methods encompass distance (k Nearest Neighbors (k-NN), (Angiulli and Pizzuti, 2002)) and cluster (FindCBLOF, (He et al., 2003)), centroid distance (glssom, (Kohonen et al., 2001)) and probabilistic based techniques (Local-Sensitive Hashing (LSH) (Indyk and Motwani, 1998), Probabilistic Anomaly Detection (PAD) (Mutz et al., 2006)). Comparison involves the analysis of theoretical computational complexities and traditional performance metrics, such as Receiver Operating Characteristic (ROC) and Precision Recall (PR) curves (Davis and Goadrich, 2006).

Interpretation/Evaluation Interpretation/Evaluation step infers the current condition of a mobile cellular network based on subsequences of LTE network events, see Section 2.5.2. Each LTE event is associated with network cell ID in accordance with either the MT location or MT target cell. Then a post-processing method assigns outage scores to each cell. The higher the outage score, the more abnormal behavior is shown by a particular cell. Cell outage histogram and network outage map presents a visual interpretation of cell outage scores, see Figure 11. To make a strict decision about down cells a cell-wise threshold learns during the training phase and is applied to the testing data set, see Figure 12. In this study the threshold is equal to the cell-wise mean plus three standard deviations of outage scores. The optimal choice of the decision boundary is the matter of future research. To evaluate developed post-processing methods we propose traditional and heuristic performance metrics.

Introduced post-processing methods assign outage scores to network cells. Post-processing methods are grouped into deviation and non-deviation ones. Deviation methods separately calculate cell outage scores for sets of all training and ”anomalous” testing subsequences. Then these vectors are subtracted from each other, and a normalized result provides the final estimate. Non-deviation methods utilize only ”anomalous” testing subsequences to calculate cell outage scores. The low-level description of post-processing methods is presented below.

Subsequences of LTE events are used to calculate cell outage scores. Firstly, LTE events are associated with cell IDs by means of either coordinates or target cell features reported in MDT logs. Afterwards, post-processing methods consider adjacent pairs of LTE events within one subsequence. A pair of events can be related to the same cell or to two different cells. Hence, post-processing methods are classified into residing (a pair of events related to the same cell) and symmetry (a pair of events related to different cells) methods. The residing method leverages the property of uneven distribution of event pairs over cells. It implies that the number of event pairs residing in a cell depends upon whether the cell



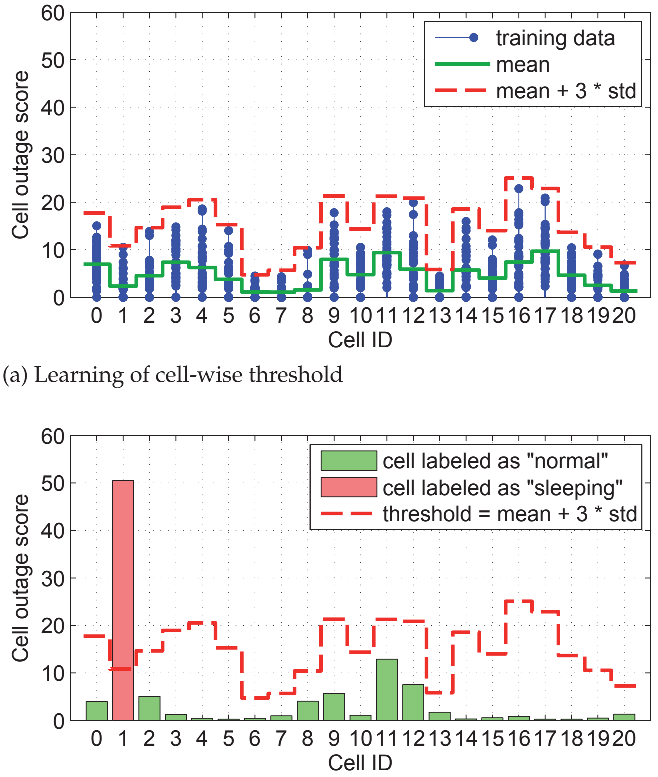
(a) Cell outage histogram

(b) Cell outage map

FIGURE 11 Visualization of cell outage detection

is operating or broken. The symmetry property, in its turn, implies that the number of incoming and outgoing event pairs of the same kind is roughly equal for a normal cell. However, this property does not hold in the presence of a sleeping cell. The difference between the number of incoming and outgoing event pairs serves as an estimate of cell operability. Then regardless a residing or symmetry method is chosen, the estimated numbers – one number for each cell – are normalized by the number of considered subsequences. Thus, at this point we have a vector of scores corresponding to the network cells. The next step is called amplification. As long as ”anomalous” testing subsequences are localized within the actual sleeping cell and its vicinity, the neighboring cells get unfairly high scores. An amplification procedure lowers this effect and enhances the score of the sleeping cell. Practically, the score of each cell is divided by the sum of scores of other cells, but not its neighbors. The vector of amplified scores is normalized by the sum of these scores and multiplied by 100. The normalized vector of amplified scores is referred to as cell outage scores, which range from 0 to 100.

KDD framework evaluates the quality of cell outage detection by making use of traditional performance metrics: accuracy, precision, F-measure, True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), ROC curve,



(a) Learning of cell-wise threshold

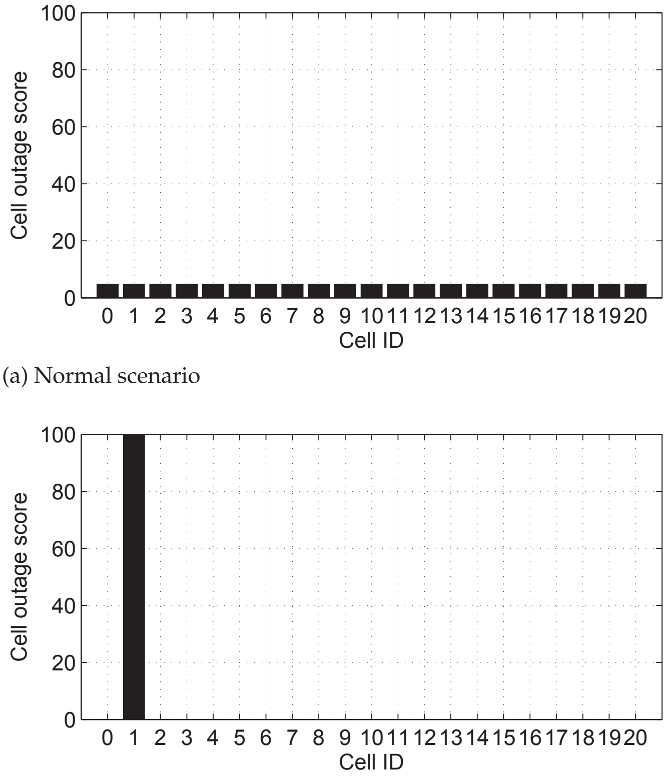
(b) Strict decision making

FIGURE 12 Cell-wise decision threshold

and PR curve, as it is specified in (Davis and Goadrich, 2006). The listed metrics are derived from elements of coincidence matrix: True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN). TP denotes the case when an impaired cell has been successfully detected. TN occurs when a cell has been correctly identified as normal. FP stands for the case when a normal cell has been recognized as broken. FN is the case when a malfunctioning cell has been labeled as normal. The elements of coincidence matrix directly depend on the chosen cell-wise decision threshold. The higher the threshold is set, the less FPs occur and the other way around.

Before describing other utilized performance evaluation criteria, there is a need to introduce an ideal cell outage detection algorithm. If a cellular network is functioning normally, then the ideal algorithm assigns equal outage scores to all cells. Due to normalization such a vector of scores is composed of $100/N_{cells}$ values, where N_{cells} is the number of cells in the network. In the case of a malfunctioning scenario, the ideal algorithm assigns an outage score of 100 to the sleeping cell and zero values to the rest of the cells. Figure 13 illustrates cell outage histograms for normal and problem (cell 1 fails) scenarios.

Root Mean Squared Error (RMSE) and an original heuristic metric are additional ways for the performance evaluation. Both criteria estimate the devia-



(a) Normal scenario

(b) Problem scenario. Cell 1 fails

FIGURE 13 Cell outage histograms of the ideal algorithm

tion of cell outage scores of a post-processing algorithm from the ideal vector of scores. RMSE measures the root of Euclidean distance between vectors of ideal outage scores and scores estimated by a post-processing algorithm. The developed heuristic metric maps the vector of outage scores to a point in a heuristic space. The heuristic performance is equal to the Euclidean distance between two points related to the ideal solution and a proposed post-processing algorithm. The heuristic space is defined a coordinate plane (”cumulative standard deviation”; ”sleeping cell magnitude”). A ”sleeping cell magnitude” coordinate is measured as the highest value among cell outage scores and can reach 100 due to normalization. A ”cumulative standard deviation” coordinate is equal to the standard deviation of cell outage scores, except the highest one. There are two points of interest corresponding to the ideal solutions on the heuristic plane. The first point $[0; 100/N_{cells}]$ is the mapping of cell outage histogram of normal network scenario, see Figure 13a. The second point $[0; 100]$ refers to the problem scenario, see Figure 13b. Note that RMSE and heuristic metrics represent the distance from a proposed solution to the ideal one. Therefore, the smaller the distance, the better the performance of a cell failure detection algorithm.

4 RESEARCH CONTRIBUTION

The main contribution of this research is the design and analysis of a framework, which automatically detects impaired cells. Such automatic functionality enables mobile operators to increase network robustness and decrease maintenance costs. The proposed framework utilizes novel data collection and data processing approaches, namely, data mining algorithms analyze MDT measurements reported by MTs. The developed knowledge discovery approach is verified and evaluated via multiple experiments in LTE/LTE-A system level simulator, which runs macro-cell and wrap-around network scenarios.

The study deals with malfunctioning cells, which are invisible for network operators via traditional alarms and are referred to as sleeping cells. Usually, operators detect sleeping cells after multiple client complaints or after the detailed analysis of drive tests measurements. The considered cell outage is caused by a failure in random access procedure. Due to this outage MTs are not able to establish a connection or to make a handover to the problem cell. However, the cellular network still services MTs, which had been synchronized with the broken cell before the problem occurred. Thus, with full signal coverage there is a lack of connectivity. This characteristic of RACH failure makes it a challenging problem.

The developed cell outage detection framework successfully detects a sleeping cell, without the need for expensive drive tests. The main findings of the research have been presented in six scientific articles. The articles can be split into two categories. PI and PV consider feasibility of the original idea and design KDD process for cell outage detection framework. The rest of the articles carry out elaborate analysis of the key components of the proposed data mining approach. The following subchapters exhibit the posed research problems and summarize the results.

4.1 Feasibility and Design of Cell Outage Detection Framework

4.1.1 Research Problem

PI and PV consider the problem of automatic detection of malfunctioning cells in LTE/LTE-A cellular networks. In practice hardware or software failures cause partial or complete degradation of cell performance. Most of the studies are focused on cell outages due to antenna gain failure or non-optimal network planning. We deal with the degradation of network service caused by RACH failure. Unlike the traditional failure cases, RACH failure does not cause signal coverage holes and therefore poses additional challenges for network operators. The other major objective of the research is to employ data mining techniques and user-level statistics to identify broken cells. The research questions to be answered are:

– *RQI*: Feasibility of the detection of a malfunctioning cell by means of user-level statistics and data mining techniques.

– *RQII*: Design of KDD based cell outage detection framework via user-level statistics.

4.1.2 Results

The paper PI introduces an original data mining framework for cell outage detection in LTE/LTE-A networks. The essence of the idea is to analyze sequences of measurements reported by MTs. At first N-gram algorithm transforms these sequences or items into feature vectors. Then dimensionality reduction and anomaly detection techniques end up with outlier items – the items, whose behavior significantly deviates from the majority. The extracted set of abnormal sequences reveals the location of the malfunctioning cell.

Journal article PV refines the original idea and describes in detail the proposed KDD approach and experimental settings. The KDD preprocessing step is complemented with the sliding window approach, which serves for the normalization of the examined sequences. As compared to PI, major improvements affect interpretation/evaluation step of the KDD process. Here we offer different post-processing techniques and compare their performances. The comparison is based on classical data mining metrics, such as precision, accuracy, and ROC curve.

The developed data mining framework processes MDT logs. The results demonstrate reliable, timely, and automatic detection of cells experiencing RACH failure in LTE networks.

4.2 Analysis of Cell Outage Detection Framework

4.2.1 Research Problems

Data mining framework introduced in our earlier publication PI exploits the KDD process: selection, preprocessing, transformation, data mining, and interpretation/evaluation steps. With an aim to improve computational complexity and framework's performance, we independently examine key KDD components and the whole model under different conditions.

Paper PII considers preprocessing and transformation steps of the developed KDD framework. The preprocessing step exploits *N*-gram analysis and maps input data on a feature space. The optimal choice of *N* is a nontrivial task, that requires finding a compromise. High *N* preserves more information about the original data and enables one to distinguish better between input items. At the same time, high value of *N* increases the dimensionality of the feature space and poses the curse of dimensionality problem. The problem is usually tackled by dimensionality reduction techniques. However, in our model the dimensionality reduction is unnecessary if *N* is equal to 1. Thus, the goal is to optimize the parameter *N* in *N*-gram analysis with respect to the complexity and performance of cell outage detection framework.

Paper PIII examines data mining step of the developed KDD framework. Data mining component is the heart of the proposed cell outage detection approach. In this paper we are aiming to compare multiple supervised learning algorithms, which address the anomaly detection task. The essence of the data mining process is, firstly, the discovering of the profile and properties of a normal data set. Afterwards, the trained algorithm assigns anomaly scores and labels input data records during the testing phase. The main challenge of this study is to enhance the detection quality of impaired cells, while preserving low time complexity of both training and testing phases.

Paper PIV considers the interpretation/evaluation step of the developed KDD framework and takes into account location measurement noise. The framework utilizes MTs' locations to detect a malfunctioning cell. Although, it is beneficial to know the exact coordinates of MTs, that much high accuracy is unachievable in practice. 3GPP Release 9 standardized different positioning methods, which provide estimates within meters and hundred meters of error range. In this regard we tackle the question how the positioning accuracy of measurements impacts on the performance of the developed approach.

Paper PVI discovers the influence of the data set size on the performance of the cell outage detection approach. As long as the developed framework collects user-level statistics, the density of MTs in a cellular network is of major importance – the more data available, the more reliable the estimation of the network behavior is provided. The density of active users in modern cellular networks varies as a function of time and location. For instance, traffic at night is significantly lower than during the day; also, network usage in urban areas is much higher than in rural ones. Besides, the future appearance of heterogeneous networks poses one more significant use case – malfunctioning of small cells. The malfunctioning of small cells is likely to happen more frequently than the failure of macro-cells, because small cells are going to be deployed by end users. Therefore, the main focus of paper PVI is to evaluate the proposed data mining approach under different densities of active users.

The following research questions summarize the aforementioned problems:

- *RQIII*: Optimize parameter N in N-gram analysis with respect to the complexity and performance of cell outage detection framework.
- *RQIV*: Examine different data mining algorithms in order to enhance the detection quality of impaired cells, while preserving the low computational complexity of KDD framework.
- *RQV*: What is the positioning accuracy impact of user-level measurements on the performance of cell outage detection framework?
- *RQVI*: How does the density of user devices influence on the performance of cell failure identification approach in LTE/LTE-A networks?

4.2.2 Results

Paper PII optimizes parameter N in N-gram analysis with respect to the complexity and performance of cell outage detection framework. Parameter N has a significant impact on time and space complexity of the approach. Results demonstrate that $N = 2$, as opposed to $N = 1$, more efficiently distinguishes between normal and outlier sequences of LTE events, but leads to higher dimensionality of feature space. The latest raises the need for dimensionality reduction. However, the advantage of high dimensional feature space is mitigated by the following components of our KDD model. Paper PII concludes that $N = 1$ as compared to $N = 2$ exhibits significantly lower space and computational complexity, while the overall framework performance is comparable.

The key element of the introduced KDD framework is a data mining procedure. Paper PIII makes use of different data mining techniques to enhance the detection quality of impaired cells and to preserve low computational complexity of the KDD framework. For that purpose we leverage and compare the following anomaly detection methods: distance based (k-NN), centroid distance based (Self-Organizing Map (SOM)), and statistic based (PAD, LSH). Practical comparison involves analysis of ROC and PR curves. Also, we present theoretical estimations of time complexities for the training and testing phases. The paper PIII concludes that PAD overtakes the other data mining methods when comparing ROC and PR curves. However, the cell outage detection quality of the framework insignificantly depends on the performance of the anomaly detection algorithm. The main reason is post-processing methods mitigate the differences between data mining algorithms.

Paper PIV discovers the impact of the location accuracy of MT measurements on the performance of the cell outage detection framework. The accuracy of MTs' coordinates depends on a used positioning technology. Post-processing methods of the KDD process utilize MTs' locations to estimate the cellular network state. In the paper, we propose and compare three post-processing algorithm in the presence of a location error: 'Target Cell O-I', 'Dominance Cell C', and 'Dominance Cell O-I'. In our experiments, the location error follows the model of Rayleigh noise. RMSE and percent gain are utilized for the performance evaluation of post-processing algorithms. As long as there is no a superior post-processing algorithm at any level of location noise, the combined method is introduced. This method combines the output of two distinct post-processing algorithms and exhibits higher reliability.

Paper PVI investigates the influence of data set size on the performance of cell outage detection approach in LTE/LTE-A networks. The data set size is a function of the density of active users and the size of the observation interval. The paper demonstrates that the proposed approach is able to detect cells experiencing random access failures, but the performance is vulnerable to the amount of available data. The performance is evaluated in terms of an ROC curve, TPR, RMSE, and fall out risk as a function of active users per cell per observation interval. The metrics demonstrate that the cell outage detection quality improves gradually, while the density of active users increases. The investigated dependences hold for different classification algorithms based on distance (*k*-NN), centroid distance (SOM), and probabilistic (PAD) data structures. Low cell load leads to a high misdetection rate, what makes our outage detection approach disadvantageous for practical use. One of the ways for the further improvement is to combine the presented approach with statistical quality control techniques.

5 CONCLUSION

No exaggeration to refer self-management and MDT functionalities as the crucial steps of the radio networks evolution. These cutting edge functionalities are highly anticipated remedies for challenges that network operators are starting to face nowadays. Analytical companies unanimously forecast the exponential growth of mobile traffic consumption over the next five years. The densification of the network structure is one of the solutions to meet growing network capacity demands. The appearance of small cells and the development of radio access technologies are leading to sophisticated and heterogeneous network architecture. The manual management of such multi-layer and multi-RAT structures looks like a very expensive, error prone, sluggish, and sometimes even impossible process. Introduced by 3GPP, the SON concept is an automated solution for all stages of a radio network lifecycle: planning, deployment, maintenance, optimization, and fault management. The implementation of this self-organizing solution would be impossible without actual knowledge of a networks condition. Existing data collection tools such as drive tests are either expensive or not enough. MDT is an auxiliary function for the network management optimization. MDT enables network operators to collect location-aware measurements about radio environment or service quality from regular cell phones. Aside from substantial capabilities, the utilization of MDT measurements brings to network operators additional challenges, such as data flow optimization, data storage, and data processing. KDD seems to be one of the most efficient and promising tools to overcome the data processing problem. KDD enables network operators to leverage various dimensionality reduction and data mining techniques to extract valuable knowledge from MDT databases. The extracted knowledge can be employed in a number of use cases, for example, cell outage detection.

The malfunctioning of network equipment negatively influences on the quality of service, hence, on customer satisfaction. As a consequence, the revenue of a network operator significantly depends on the efficiency of fault management. The thesis considers sleeping cell problem, which is a type of hardly detectable cell outage. The degradation of a sleeping cell performance is invisible to network operators via traditional KPIs, may last for days, and is usually detected after multiple client complaints.

The main scientific contribution of the thesis is the design and development of cell outage detection framework based on MDT measurements and advanced data mining techniques. Paper PI investigates the feasibility of the original idea and employs advanced data mining techniques for the detection of a sleeping cell. Based on the results from Paper PI, Paper PV introduces a complete knowledge data discovery process for cell outage recognition. The essence of the proposed method is to analyze the sequences of network events reported by mobile terminals to serving base stations. A supervised data mining algorithm at first learns the profiles of normal event sequences. Then the algorithm – also called as anomaly detection algorithm – extracts reported event sequences, whose behavior deviates from the learned normal profile. The extracted sequences are employed to reveal the location of the problematic cell. The other included articles focus on the development of the proposed data mining framework.

The development of the original framework is carried out as follows. Paper PII demonstrates that the choice of 1-gram feature selection algorithm preserves high failure detection quality of the proposed framework. The result enables one to exclude the dimensionality reduction module from the KDD process and to significantly reduce overall computational complexity. Based on the improved KDD process, Paper PIII infers that the post-processing step mitigates the differences between the utilized anomaly detection algorithms. This particular result lets one to implement the cheapest anomaly detection algorithm in terms of computational complexity. The paper provides the theoretical computational complexities of distance based (k-NN), centroid distance based (SOM), and probabilistic based (LSH, PAD) algorithms for training and testing phases. Computational complexities are presented as functions of algorithms' parameters. From the theoretical comparison one may conclude that PAD beats others in the testing phase. Other papers investigate the impact of the positioning accuracy and the scarcity of MDT measurements on the performance of the cell outage detection framework. It turns out that the location error negatively affects the framework's performance. Paper PIV introduces a combined post-processing algorithm that demonstrates high reliability in the case of location errors. The combined algorithm merges location based and target cell based post-processing algorithms. Paper PVI exhibits the applicability of the proposed approach in suburban and rural areas, where the density of cell phone users is low. This factor is critically important for data-driven solutions. The result indicates the increase of a failure detection error, while the density of users decreases. However, performance can be improved by, for example, a change detection algorithm, such as a cumulative sum.

The author's vision on the future of the proposed cell outage detection approach is as follows. From an industrial point of view, the next step is the implementation and validation within a real network infrastructure. It will require the development of a mobile application to collection the necessary field measurements. On the network side, the main challenges are related to the effective utilization of tracing functionality, the optimization of computational complexity, and the improvement of failure detection performance. From an academic point of view, future studies should consider a more complicated simulation layout and network behavior. Hannover simulation scenario appears to be a good option because it replicates the layout of a real radio network. Besides, network behavior is characterized by day, week, and seasonal changes. These peculiarities bring additional challenges for cell outage detection approaches. One of the ways out is to leverage reactive and proactive concept drift methods. So far, there have already been a number of algorithms addressing different types of cell failures. To efficiently perform self-healing features of future radio networks, cell outage detection solutions should be merged under a common framework. The main findings of the actual thesis can be borrowed and incorporated into a unified SON system such as the one developed within the European project SEMAFOUR.

Automated cell outage management is a highly anticipated feature for network operators. Cell outage management covers cell outage detection, root cause analysis, and recovery or compensation actions. Cell outage detection is a popular topic among academic and industry researchers. It is worth pointing out that most of the existing studies deal with the complete inoperability of a cell. The identification of cells with degraded performance is a more challenging task, which lacks sufficient attention from the research community. Besides, the majority of the proposed solutions are examined in ideal simulation layouts and requires practical validation. Although, there are already studies regarding a unified approach for automated failure detection, a standardized common solution is a necessary step in this area. As for root cause analysis, researchers agree that this module should represent an artificial intelligence algorithm that automatically diagnoses a failure reason. The algorithm requires supervised learning based on numerous decisions of human experts. Even though network operators may already accumulate ’symptom’ – ’failure reason’ databases, it needs to be integrated into a unified framework and continuously updated.

Future radio networks will feature a heterogeneous structure. One of the major challenges of multi-layer networks is related to the wide spread of small cells deployed by regular users without any planning. With this regard, the malfunctioning of small cells because of misconfiguration is very likely. In most of the cases, the existing solutions for macro-cell outage detection are inappropriate because of the following reasons. Usually, macro-cell solutions require centralized statistical analysis, what may cause significant communication overhead in the case of the dense deployment of small cells. Also, small cells reside within the coverage of a macro-cell and lack the sufficient statistics of user-level measurements. Thus, the development of fully automated failure detection algorithms for small cells is the next step towards self-organizing radio networks.

Despite substantial opportunities that MDT provides to network operators, the technology requires the further development. For example, algthough MDT enables one to collect location-aware measurements the existing positioning techniques fail indoors. Hence, the accurate indoor positioning is a challenge for future technology. Besides, traditional drive tests cannot be completely excluded.

Drive test campaigns are still needed in cases like new network deployment, insufficient MDT statistics, the detailed investigation of coverage holes, and others.

Overall, SON is a vital concept for future radio networks and requires further development. So far, academic and industry researchers have developed a number of different SON solutions. Although, these solutions might be beneficial in the short term, the management of numerous unsystematically designed SON functions kills all the benefits in the long term. To prevent this effect a recently finished European project SEMAFOUR introduces a unified self-management system. The system performs network configuration and maintenance across different RATs and cell layers via the set of common SON functions. With regard to the unified self-management system, SEMAFOUR project elaborates on future SON challenges. The first is the conflict diagnosis and conflict resolution between different SON functions executed in parallel. The second is the development of a mapping process from high-level objectives of a network operation into particular SON functions. The third is the interpretation of global network policies through high-level objectives. The solutions to these challenges will bring the operation of radio networks to the next self-management level.

YHTEENVETO (FINNISH SUMMARY)

Mobiilidatan suosion ja kasvun ennustetaan kasvavan eksponentiaalisesti vielä usean vuoden ajan. Radioverkkorakenteen tiivistämisen, eli nk. pienten solujen käytön, nähdään olevan avainasemassa kapasiteetin kysyntään vastattaessa. Monitasoisen verkon manuaalinen hallinta on erittäin kallis, virhealtis ja hidas prosessi. Tämän lisäksi verkon toiminnan koko elinkaaren automatisointi lisää kustannuksia ja manuaalisen työn määrää edelleen. Tähän vastatakseen 3GPP on spesifioinut radioverkon automaattisen itsehallintakonseptin nimeltä SON (Self-Organizing Networks). Keskeinen osa SONia on tiedonkeruu päätelaitteilta, joka voidaan toteuttaa esimerkiksi MDT-mekanismin avulla. MDT mahdollistaa verkko-operaattorille signaalin ja palvelunlaadun mittaukset suoraan matkapuhelimista. Itsekorjautuvuus on SONin ominaisuus, joka hyödyntää radioverkkojen vianhallintaa. Viallisen solun automaattinen ja varhainen havainnointi on vieläkin eräs verkko-operaattoreiden isoista haasteista.

Väitöskirja tarkastelee itseorganisoituvia radioverkkoja ja esittelee vikaantuneen solun tunnistusmekanismin, joka perustuu MDT-mittauksiin sekä edistyneisiin tiedonlouhintatekniikoihin. Tutkimus perustuu LTE-verkon tapahtumien sekventiaaliseen analyysiin. Tutkimus demonstroi alkuperäisen hypoteesin toteutettavuutta sekä KDD-prosessin malleja soluhäiriöiden automaattista analyysiä varten. Tutkimuksen toinen osa parantaa ehdotetun ratkaisun laskennallista vaativuutta ja suorituskykyä. Tämän lisäksi tutkimuksessa saatiin selville MDT-mittausten sijaintitarkkuuden ja niukkuuden vaikutuksia vikaantuneen solun havainnointitarkkuuteen. Teorian validointiin käytettiin huipputason LTE/LTE-A simulaattoria. Tulokset osoittavat, että viallinen solu voidaan havainnoida luotettavasti ja hyvissä ajoin. Täten kehitettyä tunnistusmekanismia voidaan harkita käytännön validointiin ja käyttöönottoon.

REFERENCES

3GPP 2010. "Evolved Universal Terrestrial Radio Access (E-UTRA); Study on minimization of drive-tests in next generation networks". TR 36.805.

3GPP 2011. "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network use cases and solutions". TS 36.902.

3GPP 2014a. "Overview of 3GPP Release 10 V0.2.1".

3GPP 2014b. "Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements". TS 36.500.

3GPP 2014c. "Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio measurement collection for Minimization of Drive Tests (MDT); Overall description; Stage 2". TS 37.320.

3GPP 2015a. "Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall description; Stage 2". TS 36.300.

3GPP 2015b. "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification". TS 36.331.

3GPP 2015c. "Technical specification group services and system aspects; IP Multimedia Subsystem (IMS); Stage 2, (Release 13)". TS 23.228.

3GPP 2015d. "Telecommunication management; Subscriber and equipment trace; Trace concepts and requirements". TS 32.421.

3GPP 2015e. Telecommunication management; Subscriber and equipment trace; Trace control and configuration management. TS 32.422.

Altman, Z. 2006. "Project GANDALF". ⟨URL:https://www.celticplus.eu/project-gandalf⟩. (Accessed 30.10.2015).

Amirijoo, M., Jorguseski, L., Litjens, R. & Nascimento, R. 2011. "Effectiveness of cell outage compensation in LTE networks". In Consumer Communications and Networking Conference (CCNC). IEEE, 642 - 647.

Angiulli, F. & Pizzuti, C. 2002. "Fast outlier detection in high dimensional spaces". In Principles of Data Mining and Knowledge Discovery (PKDD) conference. London, UK: Springer-Verlag, 15–26.

Azevedo, A. & Santos, M. F. 2008. "KDD, SEMMA and CRISP-DM: a parallel overview". In IADIS conference on data mining. IADIS, 182-185.

Baumann, D. 2014. "Minimization of Drive Tests (MDT) in mobile communication networks". In Future Internet (FI) and Innovative Internet Technologies and Mobile Communications (IITM) seminars. The Technische Universität München, 9-16.

Blennerud, G. 2010. "Mobile broadband-busting the myth of the scissor effect". Ericsson Business Review 2.

Cassidian 2013. "5 things you need to know when investing in a new radio communication network".

Chernogorov, F., Turkka, J., Ristaniemi, T. & Averbuch, A. 2011. "Detection of sleeping cells in LTE networks using diffusion maps". In Vehicular Technology Conference (VTC) spring. IEEE, 1-5.

Cheung, B., Fishkin, S., Kumar, G. & Rao, S. 2006. "Method of monitoring wireless network performance". (US Patent App. 10/946,255).

Choi, J., Kim, H., Choi, C. & Kim, P. 2011. "Efficient malicious code detection using N-gram analysis and SVM". In Network-Based Information Systems (NBIS) conference. Washington, DC, USA: IEEE Computer Society, 618–621.

Ciocarlie, G., Lindqvist, U., Novaczki, S. & Sanneck, H. 2013. "Detecting anomalies in cellular networks using an ensemble method". In Network and Service Management (CNSM) conference, 171-174.

Ciocarlie, G., Lindqvist, U., Nitz, K., Novaczki, S. & Sanneck, H. 2014a. "DCAD: dynamic cell anomaly detection for operational cellular networks". In Network Operations and Management Symposium (NOMS). IEEE, 1-2.

Ciocarlie, G., Lindqvist, U., Nitz, K., Novaczki, S. & Sanneck, H. 2014b. "On the feasibility of deploying cell anomaly detection in operational cellular networks". In Network Operations and Management Symposium (NOMS). IEEE.

Cisco 2015. "Cisco visual networking index: global mobile data traffic forecast update, 2014–2019".

Coenen, F. 2011. "Data mining: past, present and future". The knowledge engineering review 26 (1), 25-29.

Cong, F., Nandi, K. A., He, Z., Cichocki, A. & Ristaniemi, T. 2012. "Fast and effective model order selection method to determine the number of sources in a linear transformation model". In European Signal Processing Conference (EUSIPCO), 1870–1874.

Davis, J. & Goadrich, M. 2006. "The relationship between precision-recall and ROC curves". In International Conference on Machine Learning (ICML). New York, NY, USA: ACM, 233–240.

Eisenblatter, A., Gonzalez Rodriguez, B., Gunnarsson, F., Kurner, T., Litjens, R., Sas, B., Sayrac, B., Schmelz, L. & Willcock, C. 2013. "Integrated self-management for future radio access networks: Vision and key challenges". In Future Network and Mobile Summit (FutureNetworkSummit), 1-10.

Ericsson 2011. "Positioning with LTE". Ericsson white paper.

Ericsson 2015. "On the pusle of the network society". Ericsson mobility report.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996a. "The KDD process for extracting useful knowledge from volumes of data". Communications of the ACM 39 (11), 27–34.

Fayyad, U. M., Piatetsky-Shapiro, G. & Smyth, P. 1996b. "Advances in knowledge discovery and data mining". In "Advances in knowledge discovery and data mining". Menlo Park, CA, USA: AAAI, 1–34.

Ganapathiraju, M., Weisser, D., Rosenfeld, R., Carbonell, J., Reddy, R. & Klein-Seetharaman, J. 2002. "Comparative N-gram analysis of whole-genome protein sequences". In Human Language Technology (HLT) conference. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 76–81.

Hahn, S. & et al. 2015. "D6.6 - Final report on a unified self-management system for heterogeneous radio access networks". FP7 SEMAFOUR. Public Deliverable.

Hand, D. J., Smyth, P. & Mannila, H. 2001. "Principles of data mining". Cambridge, MA, USA: MIT Press.

He, Z., Xu, X. & Deng, S. 2003. "Discovering cluster-based local outliers". Pattern Recogn. Lett. 24 (9-10), 1641–1650.

Holma, H. & Toskala, A. 2011. "LTE for UMTS: evolution to LTE-advanced" (2nd edition). Wiley Publishing.

Hwang, I., Song, B. & Soliman, S. 2013. "A holistic view on hyper-dense heterogeneous and small cell networks". IEEE Communications Magazine 51 (6), 20-27.

Hämäläinen, S., Sanneck, H. & Sartori, C. 2012. "LTE Self-Organising Networks (SON): network management automation for operational eefficiency" (1st edition). Wiley Publishing.

Indyk, P. & Motwani, R. 1998. "Approximate nearest neighbors: towards removing the curse of dimensionality". In ACM Symposium on Theory of Computing (STOC). New York, NY, USA: ACM, 604–613.

Johansson, J., Hapsari, W., Kelley, S. & Bodog, G. 2012. "Minimization of drive tests in 3GPP Release 11". Communications Magazine, IEEE 50 (11), 36-43.

Josse, J., Pagès, J. & Husson, F. 2011. "Multiple imputation in principal component analysis". Advances in Data Analysis and Classification 5 (3), 231–246.

KDnuggets 2014. "What main methodology are you using for your Analytics, data mining, or data science projects?". ⟨URL:http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html⟩. (Accessed 22.10.2015).

Kohonen, T., Schroeder, M. R. & Huang, T. S. (Eds.) 2001. "Self-organizing maps" (3rd edition). Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Kotu, V. & Deshpande, B. 2015. "Predictive analytics and data mining". Boston, USA: Morgan Kaufmann.

König, W. 2009. "End-to-end efficiency project". ⟨URL:https://ict-e3.eu⟩. (Accessed 30.10.2015).

Kürner, T. & et al. 2010. "Final report on self-organisation and its implications in wireless access networks". SOCRATES. D5.9. ⟨URL:http://www.fp7-socrates.eu⟩. (Accessed 30.10.2015).

Laiho, J., Wacker, A. & Müller, S. 2007. "Measurement based methods for WCDMA radio network design verification". In Spring Simulaiton (SpringSim) multiconference - Volume 1. San Diego, CA, USA: SCS, 234–239.

Laiho, J., Wacker, A. & Novosad, T. 2005. "Radio network planning and optimisation for UMTS" (1st edition). Chichester: Wiley.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. 2011. "Big data: the next frontier for innovation, competition, and productivity". McKinsey Global Institute.

Mueller, C. M., Kaschub, M., Blankenhorn, C. & Wanke, S. 2008. "A cell outage detection algorithm using neighbor cell list reports". In International Workshop on Self-Organizing Systems (IWSOS). Berlin: Springer-Verlag, 218–229.

Mutz, D., Valeur, F., Vigna, G. & Kruegel, C. 2006. "Anomalous system call detection". ACM Transactions on Information and System Security 9 (1), 61–93.

NGMN 2007. "Use cases related to self-organising network, overall description". NGMN.

NGMN 2008. "Recommendation on SON and O&M requirements". NGMN.

NGMN 2010. "NGMN top OPE recommendations". NGMN.

Nokia 2015. "Intelligent Self Organizing Networks (iSON)". ⟨URL:http://networks.nokia.com/be/portfolio/solutions/self-organizing-networks⟩. (Accessed 30.10.2015).

Nováczki, S. 2013. "An improved anomaly detection and diagnosis framework for mobile network operators". In Design of Reliable Communication Networks (DRCN) Conference, 234–241.

Nováczki, S. & Szilágyi, P. 2011. "Radio channel degradation detection and diagnosis based on statistical analysis". In Vehicular Technology Conference (VTC) Spring. IEEE, 1-2.

Nováczki, S. & Szilágyi, P. 2012. "An intelligent anomaly detection and diagnosis assistant for mobile network". In International Workshop on Self-Organizing Networks (IWSON). IWSON.

Olsson, M., Sultana, S., Rommer, S., Frid, L. & Mulligan, C. (Eds.) 2013. "EPC and 4G packet networks: driving the mobile broadband revolution" (2nd edition). Academic Press.

Pingping, X., Li, Y., Clara, Q. L., Panagiotis, D. & Andreas, G. 2013. "Multi-RAT network architecture". Wireless World Research Forum.

Rose, D., Jansen, T., Werthmann, T., Türke, U. & Kürner, T. 2013. "The IC 1004 urban Hannover scenario – 3D pathloss predictions and realistic traffic and mobility patterns". EURO-COST.

Sesia, S., Toufik, I. & Baker, M. 2011. "LTE, the UMTS long term evolution: from theory to practice" (2nd edition). Wiley Publishing.

Szilágyi, P. & Nováczki, S. 2012. "An automatic detection and diagnosis framework for mobile communication systems". IEEE Transactions on Network and Service Management 9 (2), 184-197.

Turkka, J., Chernogorov, F., Brigatti, K., Ristaniemi, T. & Lempiäinen, J. 2012. "An approach for network outage detection from drive-testing databases". Computer Networks and Communications 2012, 163184:1-163184:13.

Wang, W., Liao, Q. & Zhang, Q. 2014. "COD: a cooperative cell outage detection architecture for self-organizing femtocell networks". IEEE Transactions on Wireless Communications 13 (11), 6007-6014.

Wang, W., Zhang, J. & Zhang, Q. 2011. "Transfer .earning based diagnosis for configuration troubleshooting in self-organizing femtocell networks". In Global Communications (GLOBECOM) Conference. IEEE, 1-5.

Wikipedia 2015a. "Bluetooth". ⟨URL:https://en.wikipedia.org/wiki/Bluetooth⟩. (Accessed 22.10.2015).

Wikipedia 2015b. "Integrated services digital network". ⟨URL:https://en.wikipedia.org/wiki/Integrated_Services_Digital_Network⟩. (25.10.2015).

Wikipedia 2015c. "Wi-Fi". ⟨URL:https://en.wikipedia.org/wiki/Wi-Fi⟩. (Accessed 22.10.2015).

Willcock, C. 2015. "SEMAFOUR. Self-management for unified heterogeneous radio access networks". ⟨URL:http://fp7-semafour.eu⟩. (Accessed 30.10.2015).

Xiaojin, Z. 2012. "LTE-advanced air interface technology" (1st edition). CRC Press.

ORIGINAL PAPERS

PI

N-GRAM ANALYSIS FOR SLEEPING CELL DETECTION IN LTE NETWORKS

by

Fedor Chernogorov, Tapani Ristaniemi, Kimmo Brigatti, Sergey Chernov 2013

38th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, Canada

Reproduced with kind permission of IEEE.

N-GRAM ANALYSIS FOR SLEEPING CELL DETECTION IN LTE NETWORKS

Fedor Chernogorov^{1,2}, *Tapani Ristaniemi*^{1}, *Kimmo Brigatti*^{1}, *Sergey Chernov*^{1}

^{1}University of Jyväskylä, Department of Mathematical Information Technology, Jyväskylä, Finland
^{2}Magister Solutions Ltd., Jyväskylä, Finland

ABSTRACT

Sleeping cell detection in a wireless network means to find the cells which are not working properly due to various reasons. The research in the area has mostly focused on cell outage detection, e.g. due to hardware failures at the base station antennas or non-optimal network planning. In this paper we extend the research into a more challenging setting which is overlooked in the literature: the case where no outages occur in the network. The essence of the proposed method for detection of problematic cells is to analyze the sequences of the events reported by the mobile terminals to the serving base stations. The suggested n-gram analysis includes dimensionality reduction and classification of the data and ends up with providing a set of abnormal users, which at the end reveal the location of the problematic cell. We verify the proposed framework with simulated LTE network data and using the minimization of drive testing (MDT) functionality to gather the training and testing data sets.

1. INTRODUCTION

Self-healing, which is a part of self-organizing network concept, means automated detection of problems or malfunctioning in the radio network elements and actions to automatically recover from these problematic situations [1]. Most of the works considered so far have focused on cell outage detection (see e.g. [2–4] and references therein) and management [5, 6]. Reasons for outage situations are many, but the usual ones are hardware problems in base station antennas, improper radio network planning, erroneous antenna tilt or transmit power. Hence the usual approach for cell outage detection is to analyze several key performance indicator (KPI) measurements from both base stations and mobile terminals.

Latest works in this line of research were recently published by the authors in [7] and [8]. The approach in [7] was to analyze the data set of signal strength and quality measurements reported by mobile terminals. These measurements contained both serving and neighboring cell measurements in LTE network according to minimization of drive testing (MDT) functionality specified by 3GPP. The main finding was that advanced data mining and machine learning techniques, which rely on autonomous learning of network behavior, were able to reveal latent abnormal behavior in the high dimensional data set of RF measurements and can thus be used to pinpoint a problematic cell in the network. In [8] this approach was extended by targeting to find similarities between periodical measurement reports and reports related to failures happened before at the radio link. By this way one was able to substantially increase the number of samples (in addition to true failure reports) which indicated the existence of a problem in specific cells, resulting in more reliable and faster detection.

All the above-mentioned approaches rely on the measurements of the radio environment, which however, are able to reveal only radio related problems. In this paper we extend the scope of problem detection in radio networks by considering the case where radio coverage outages do not exist. This is a relevant case in practice e.g. when there exists hierarchical cells (pico/micro/macro) in the same area or when the real problem of a particular cell is not radio related at all. An example of the latter case is a software bug or a malfunctioning protocol. Detecting a cell having such problems is no longer doable by analyzing RF measurements, but calls for another approach. A relevant solution where the problematic cell was detected by investigating graphs constructed from the reported neighboring cell patterns can be found in [4]. The essence of this paper is to employ more generic approach by analyzing the sequences of events reported by mobile terminals to the serving base stations. Subsequently, the approach will end up with providing a set of abnormal users (or calls) in the networks, which can be utilized at the end to reveal the location of the problematic cell.

2. SLEEPING CELL PROBLEM

Sleeping cell is a special kind of cell degradation. A cell is called degraded in case if it is not 100% functional - its services are suffering in terms of quality what affects user ex-

This work was financially supported by Tekes (the Finnish Funding Agency for Technology and Innovation) under grant 40040/08 “Real-Time Management of Industrial Processes and Communication Networks (MIPCOM)” and Celtic-Plus (an industry-driven European research initiative) project “Cognitive Network Management Under Uncertainty (COMMUNE)”. Authors would like to thank colleagues from University of Jyväskylä and Magister Solution Ltd. for their professional advices and support.

perience. There exist a vague classification of degraded cells depending on how much they affect the network operation (partly based on [9]). The first type is *impaired* cell - which still carries some traffic, but the performance characteristics are slightly lower expected. The second kind of degradation is *crippled* cell, characterized by a severely decreased capacity. The last, clearly most critical type of sleeping cell is *catatonic* cell - kind of outage which leads to complete absence of service in the faulty area and cell does not carry any traffic and for that reason it is important to timely detect such cells and apply recovery actions.

Usual degraded cell produces fault alarms which are available to mobile network operator. In opposite, in sleeping cells degradation appears seamlessly and no direct notification to the service provider is given.

Different hardware or software failures can cause appearance of a sleeping cell and due to that it is considered to be a complex umbrella term. In this research we investigate catatonic sleeping cells with RACH (Random Access Channel) problem described in further details in Section 4.

3. DETECTION FRAMEWORK

3.1. N-Gram Analysis

An n-gram is defined as a subsequence of n terms. These terms can be e.g. letters or words from a sequence. The analysis results in statistics regarding the frequency of occurrence of n-grams within string sequence. Thus, feature vector of n-gram frequencies can be assembled from the string sequence.

N-gram analysis is widely used in spheres concerning data processing. It has been utilized e.g. for the analysis of whole-genome protein sequences [10], computer virus detection [11] and also in a wide variety of natural language processing applications.

In our research the terms are network events reported by the mobile terminal to the base station (in total 10 events listed in Table 1). [1] The data used for sleeping cell detection is a K by 10^n matrix containing the n-gram frequencies of each of K individual users (or call), where n is the number of terms in considered subsequences.

3.2. Dimensionality reduction and classification

The goal for data analysis here is first to identify abnormal calls. As a next step this information is used for the detection of the sleeping cell. To do that, one usually performs reduction of the dimensionality for the data and clusters the data in low dimensional space. Here we performed standard principal component analysis for dimensionality reduction and applied the FindCBLOF [14] algorithm for clustering and outlier detection part. The advantage of FindCBLOF is in its ability to find local outliers based on the clustering solution for training data.

Table 1. Network events triggering MDT log entry

PL PROBLEM - Physical Layer Problem [12].
RLF - Radio Link Failure [13].
RLF REESTABLISHMENT - Connection reestablishment after RLF.
A2 RSRP ENTER - RSRP goes under A2 enter threshold.
A2 RSRP LEAVE - RSRP goes over A2 leave threshold.
A2 RSRQ ENTER - RSRQ goes over A2 enter threshold.
A3 RSRP - A3 event, according to spec.
HO COMMAND RCVD - handover command received [13].
HO COMPLETE RCVD - handover complete received [13].
HO TO VOID - handover is done to one of the cells in outer tier.

3.3. Symmetry Analysis of 2-Gram Subsequences

Under symmetry we mean the following: if the first event of a 2-gram is located in cell A and the second event is located in cell B, we are interested in how many of those 2-grams originate from A and how many originate from B. In simulations, where the user movement is random, one expects any 2-grams to be somewhat balanced. Hence, the deviation from learned balance is to be used as an indication of problem in a particular cell.

4. SIMULATION ASSUMPTIONS AND GENERATED DATA

Dynamic system level LTE simulator with step resolution of one OFDM[2] symbol has been used as a platform for data generation in this research. The simulator is designed in accordance to specifications 3GPP E-UTRAN Release. 8 and beyond. Methodology for mapping link level SINR to system is presented in [15].

Network scenario utilized in the simulations for this study and shown on Fig. 1, is an extended version of 3GPP macro case 1, described in [16]. Scenario setup is such that outer tier of cells is used only for interference generation to make radio link conditions more realistic. On the other hand 21 center cells are utilized for statistical data collection. The main simulation parameters applied in this research are presented in Table 2.

[1] RSRP = Reference Signal Received Power; RSRQ = Reference Signal Received Quality; A2 = an event which triggers when the serving cell becomes worse than threshold; A3 = an event which triggers when a neighboring cell becomes an offset better than the serving cell.

[2] OFDM - Orthogonal Frequency-Division Multiplexing

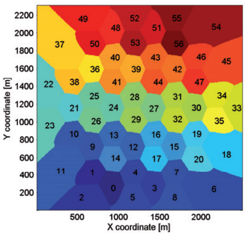
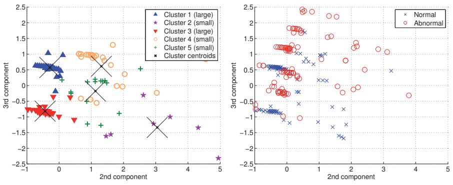


Fig. 1. Macro 57 network scenario layout

Table 2. General Simulation Parameters

Parameter	Value
Cellular layout	Homogeneous Macro 57
Number of cells	21 active and 36 interfering
Inter-Site Distance	500 m
Link direction	Downlink
Maximum BS TX power	46 dBm
Initial cell selection criterion	Strongest RSRP value
Simulation length	142 s (2000000 steps)
Simulation resolution	1 time step = 71.43 μs
Max number of UEs/cell	20
UE velocity	30 km/h
Duration of calls	Uniform 30 to 140 s
Traffic model	Constant Bit Rate 256 kbps
Reference case	Simulation without sleeping cell
Problematic case	Simulation with RACH problem in cell 28



(a) FindCBLOF training

(b) FindCBLOF testing

Fig. 2. FindCBLOF testing and training results

In this paper the sleeping cell was modelled through malfunctioning of the Random Access Channel (RACH). RACH is a channel used in connection establishment in the beginning of a call when establishment procedure is initiated, during handover to another cell or connection re-establishment after handover failure or RLF.

By simulating LTE network operation we generate a performance monitoring dataset using the principle of drive test minimization reporting. This principle implies addition of log entry by the mobile terminal to a global MDT log either periodically or at occurrence of a specified network events, presented in Table 1. Usually one sample includes values of different performance indicators, time stamp and location fingerprint. Depending on the type of the sleeping cell we might need different amount of information from the log. As far as in this research we are doing identification of random access sleeping cells the required MDT data limits to event type and location of this event occurrence. In more details the procedure of drive test minimization, variables of the MDT log utilized for detection of other kinds of sleeping cells are presented in [7, 8].

5. RESULTS

5.1. Analysis of Abnormal User Calls

In accordance to our detection framework the first step is the construction of n-gram subsequences as described in Section 3.1. Using network events shown in Table 1, we chose $n = 2$ for simplicity and generate the full set of 2-gram subsequences.

Reference data were used for creation of a normal network operation model. There were 264 users with sequences longer than 20 event-triggered MDT log entries, while shorter user sequences were filtered. On the basis of these reference user sequences corresponding matrix of 2-gram occurrences was constructed. After that same procedure was done with the problematic data, Resulting occurrence matrix of 2-grams was compared to the corresponding reference matrix in order to find anomalous users.

Reference data were clustered to five groups, among which there were two large and dense clusters (1 & 3) and three small (clusters 2, 4 & 5), as shown in Fig. 2a. In problematic data, users were clustered into normal and abnormal groups; red markers shown on Fig. 2b represent abnormal user, while the rest belong to normal users' group. Decision whether a certain point is abnormal or not was based on the value of CBLOF, where higher likelihood sample abnormality corresponds to high score values. From Fig. 2a and Fig. 2b it can be seen that points from the problematic dataset which belong to the areas of compact clusters in training data are marked as normal, while samples which are outside of dense clusters or are in low density areas are clustered as abnormal. In total 113 users were marked as abnormal and 205 users as normal.

After having detected the abnormal users, their movement can be traced by locating the events through dominance maps like in [7]. As can be seen from Fig. 3, abnormal users tend to

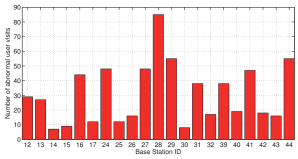


Fig. 3. Number of cell visits by abnormal users

be those who have visited the sleeping cell 28. Naturally due to mobility they visit other cells as well. In these simulations abnormal users on average camp in 6 different cells, while the range varies from 3 to 11 cells. From the histogram on Fig. 3, it can be observed that 85 abnormal users (75.2%) have visited cell 28. Next 8 most visited cells get from 38 to 55 visits from which 6 are neighbors of cell 28 (cells 24, 27, 29, 39, 41 & 44). On the basis of this information we can claim that there is some anomalous behavior in the area of cell 28.

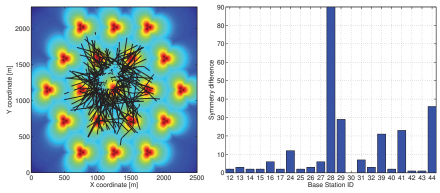
5.2. Analysis of Abnormal 2-Gram Sequences

Observation of the abnormal users gives only a rough idea of possible problem in the cell of interest. For that reason more detailed analysis of 2-gram subsequences of the abnormal users' calls needs to be employed. As further results demonstrate, this approach gives more reliable indication of problem existence. In particular, the knowledge of the most descriptive 2-grams, meaning that over 50% of abnormal users have this 2-gram occurred at least once, is taken into account. Nine 2-grams met this condition and as an example the characteristics of two of them are shown on Figs. 4 and 5.

Sequence "A2 RSRP LEAVE - A3 RSRP" is a common 2-gram all over the network and it should occur within all users who are on the move. In the group of abnormal users it exists for all users and the total number of occurrences is 869. However, in the dominance area of cell 28 this sequence occurs far less frequently than for the rest of the network, as it can be seen from Fig. 4a.

Sequence "HO COMMAND - A2 RSRP ENTER", on the other hand, is a direct consequence of Random Access problem. In normal network behavior "HO COMMAND" should be followed by "HO COMPLETE", but as it can be seen "A2 RSRP ENTER" appears instead. Thus this sequence happens only in the area of problem (in total 126 times), as it can be seen from Fig. 5a.

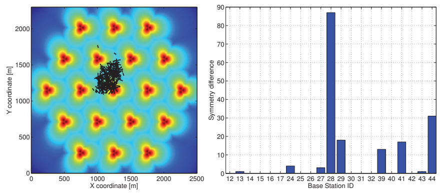
The described examples of abnormal 2-gram sequences are the only ones among the nine selected 2-grams. To select these subsequences in automatic manner, thus being able to detect sleeping cell, symmetry analysis based on their locations is employed, as described in Section 3.3. As it can be seen from Figs. 4b and 5b there exists a clear unbalance for each of these 2-grams in the dominance areas of cell 28 and also in its neighbor cells 29, 39, 41 and 44. Elsewhere in the network these 2-grams are more or less in better balance. In fact, the sequence "A2 RSRP LEAVE - A3 RSRP" completes in cell 28 more often than it starts from there and more often than it ends in one of its neighbor cells. Regarding "HO COMMAND - A2 RSRP ENTER" subsequence, it starts more often from cell 28 ending up in one of its neighbors than vice versa. Thus, the symmetry analysis demonstrates that the behavior of cell 28 is clearly abnormal.



(a) Locations of this 2-gram sequence

(b) Deviation from the 2-gram symmetry in each cell

Fig. 4. Characteristics of abnormal 2-gram sequence "A2 RSRP LEAVE - A3 RSRP", which is a common 2-gram for all the abnormal calls.



(a) Locations of this 2-gram sequence

(b) Deviation from the 2-gram symmetry in each cell

Fig. 5. Characteristics of abnormal 2-gram sequence "HO COMMAND - A2 RSRP ENTER", which is a common 2-gram for all the abnormal calls.

6. CONCLUSIONS

In this article advanced data mining framework for the network performance monitoring automation was presented. The considered problem of sleeping cell detection, is among highly complex identification problems as far as there is no direct alarm sent to the operator. A validation of the framework was given in this setting using the random access malfunction as an example for the sleeping cell root cause.

Suggested detection framework is based on such techniques as n-gram analysis, association-based clustering algorithm and dimensionality reduction. Altogether application of these methods in the proposed way on top of MDT data leads to a reliable detection of random access sleeping cell.

References

[1] *Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions (Release 11)*, 3GPP Std. TR 36.902, 2011.

[2] O. Sallent, J. Pérez-Romero, J. Sánchez-González, R. Agustí, M. Díaz-guerra, D. Henche, and D. Paul, “A roadmap from UMTS optimization to LTE self-optimization,” *Communications Magazine, IEEE*, vol. 49, no. 6, pp. 172–182, June 2011.

[3] J. Sánchez-González, O. Sallent, J. Pérez-Romero, R. Agustí, M. Díaz-guerra, J. Moreno, and D. Paul, “A new methodology for RF failure detection in UMTS networks,” in *Network Operations and Management Symposium, NOMS 2008, IEEE*, April 2008, pp. 718–721.

[4] C. M. Mueller, M. Kaschub, C. Blankenhorn, and S. Wanke, “A cell outage detection algorithm using neighbor cell list reports,” *Lecture Notes in Computer Science, Springer Berlin/Heidelberg*, 2008.

[5] M. Amirijoo, L. Jorguseski, T. Kürner, R. Litjens, M. Neuland, L. C. Schmelz, and U. Türke, “Cell outage management in LTE networks,” in *Proceedings of the 6th international conference on Symposium on Wireless Communication Systems*, ser. ISWCS’09. IEEE Press, 2009, pp. 600–604.

[6] M. Amirijoo, L. Jorguseski, R. Litjens, and R. Nascimento, “Effectiveness of cell outage compensation in LTE networks,” in *Consumer Communications and Networking Conference (CCNC), 2011 IEEE*, January 2011, pp. 642–647.

[7] F. Chernogorov, J. Turkka, T. Ristaniemi, and A. Averbuch, “Detection of sleeping cells in LTE networks using diffusion maps,” in *Vehicular Technology Conference (VTC Spring), 2011 IEEE 73rd*, May 2011, pp. 1–5.

[8] J. Turkka, F. Chernogorov, K. Brigatti, T. Ristaniemi, and J. Lempiäinen, “An approach for network outage detection from drive-testing databases,” *Journal of Computer Networks and Communications*, 2012.

[9] B. Cheung, S. Fishkin, G. Kumar, and S. Rao, “Method of monitoring wireless network performance,” Patent US 2006/0 063 521 A1, 2006.

[10] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy, and J. Klein-Seetharaman, “Comparative n-gram analysis of whole-genome protein sequences,” in *Proceedings of the Human Language Technologies conference*, 2002.

[11] J. Choi, H. Kim, C. Choi, and P. Kim, “Efficient malicious code detection using n-gram analysis and SVM,” in *Proceedings of the 2011 14th International Conference on Network-Based Information Systems*, ser. NBIS ’11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 618–621.

[12] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 10)*, 3GPP Std. TS 36.331, 2011.

[13] S. Sesia, M. Baker, and I. Toufik, *LTE - The UMTS Long Term Evolution: From Theory to Practice*. John Wiley & Sons, 2011.

[14] Z. He, X. Xu, and S. Deng, “Discovering cluster based local outliers,” *Pattern Recognition Letters*, vol. 2003, pp. 9–10, 2003.

[15] K. Brueninghaus, “Link performance models for system level simulations of broadband radio access systems,” in *PIMRC-2005*, Berlin, Germany, 2005.

[16] *Technical Specification Group Radio Access Network; Further Advancements for E-UTRA Physical Layer Aspects (Release 9)*, 3GPP Std. TR 36.814, 2010.

PII

DATA MINING FRAMEWORK FOR RANDOM ACCESS FAILURE DETECTION IN LTE NETWORKS

by

Sergey Chernov, Fedor Chernogorov, Dmitry Petrov, Tapani Ristaniemi 2014

25th IEEE Annual International Symposium on Personal, Indoor, and Mobile
Radio Communication (PIMRC), Washington DC, USA

Reproduced with kind permission of IEEE.

Data Mining Framework for Random Access Failure Detection in LTE Networks

Sergey Chernov
University of Jyväskylä
Jyväskylä, Finland
sergey.chernov@jyu.fi

Fedor Chernogorov
Magister Solutions Ltd.
Jyväskylä, Finland
fedor.chernogorov@magister.fi

Dmitry Petrov
Magister Solutions Ltd.
Jyväskylä, Finland
dmitry.petrov@magister.fi

Tapani Ristaniemi
University of Jyväskylä
Jyväskylä, Finland
tapani.ristaniemi@jyu.fi

***Abstract*—Sleeping cell problem is a particular type of cell degradation. There are various software and hardware reasons that might cause such kind of cell outage. In this study a cell becomes sleeping because of Random Access Channel (RACH) failure. This kind of network problem can appear due to misconfiguration, excessive load or software/firmware problem at the Base Station (BS). In practice such failure might cause network performance degradation, which is hardly traceable by an operator. In this paper we present a data mining based framework for the detection of problematic cells. In its core is the analysis of event sequences reported by a User Equipment (UE) to a serving BS. The choice of N in N-gram feature selection algorithm is considered, because of its significant impact on computational efficiency. Moreover, qualitative and heuristic performance metrics have been developed to assess the performance of the proposed detection algorithm. Sleeping cell detection framework is verified by means of dynamic LTE (Long-Term Evolution) system simulator, using Minimization of Drive Testing (MDT) functionality. It is shown that sleeping cell can be determined with very high reliability even using 1-gram algorithm.**

I. Introduction

Nowadays, reduction of operational expenses becomes one of the main requirements for constantly growing mobile cellular networks in addition to such needs as high throughput, low capital expenditures, etc. Maintenance of more and more advanced networks is significantly complex and an expensive issue. In order to automate network operations and reduce operators expenditures, the 3rd Generation Partnership Project (3GPP) has proposed Self-Organizing Network (SON) technology in Release 8 and subsequent specifications [1].

Self-healing is a part of SON paradigm, which is designed for automated detection of the malfunctioning and for the outage management in cellular networks. So far, there are many studies that are focused on the cell outage detection [2] and the mitigation of the malfunctioning cells [3]. The most common reasons for network faults are hardware and software failures, external failure of power supply or network connectivity, or even erroneous configuration.

Advanced data mining and machine learning techniques were used for the detection of cell outages in publications [4, 5]. The designed methods utilized MDT functionality as it is specified by 3GPP [6]. MDT records are reported by UEs and combine signal strength and quality measurements from serving and neighboring eNBs . The presented detection algorithms are based on radio frequency measurements. Therefore, they are able to pinpoint only problems in the radio part of the network.

In the paper [7] we considered the other kind of cell failure, which is not related to radio coverage outages. The radio access malfunctioning is a highly complex failure as long as there is no direct alarm for network operators. The application of data mining methods to MDT measurements resulted in a reliable detection of a sleeping cell.

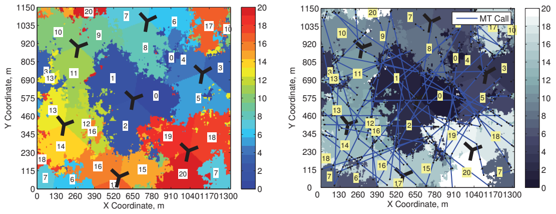
The target of this study is to provide more advanced and improved knowledge data discovery framework than in paper [7]. Besides, authors consider the choice of N in N-gram feature selection algorithm, as long as it has a great impact on computational efficiency of the overall framework. The other novelties are the automatic sleeping cell decision making and extended usage of both conventional and own-developed performance metrics.

The rest of this paper is organized as follows: in Section 2 we describe in more details a sleeping cell problem and consider RACH failure, which causes such network malfunctioning. Then description of the simulation tool used for this study and corresponding simulation assumptions are given in Section 3. After that we present our sleeping cell detection framework in Section 4. in Section 5 we demonstrate simulation results and discuss performance of the developed framework. Finally, Section 6 concludes the paper.

II. Sleeping cell problem

Sleeping cell is a special kind of cell degradation that means a malfunction resulting in a network performance decrease. In many cases it remains invisible for a network operator. Depending on the extent of performance degradation, sleeping cells can be roughly classified into three groups [8]. The first type of sleeping cell is called *impaired*. In this case the certain part of traffic is carried, but performance characteristics are slightly lower than expected. The second kind of degradation is *crippled* cell, characterized by a severely decreased capacity. The last, and clearly the most critical type of sleeping cell is a *catatonic* cell. This kind of outage leads to the complete absence of service in the faulty area and cell does not carry any traffic. For that reason it is important to timely detect such cells and apply recovery actions.

978-1-4799-4912-0/14/$31.00 ©2014 IEEE



(a) Slow faded dominance map

(b) Mobile terminal calls on slow faded dominance map

Fig. 1. Wrap around Macro 21 scenario

Sleeping cell problem can be caused by different hardware or software failures. In this study we consider sleeping cell problem caused by RACH failure. This kind of failure can appear due to RACH misconfiguration, excessive load or software/firmware problem at the eNB side [9]. RACH malfunction leads to inability of the affected cell to serve any new users. However, earlier connected UEs get served, because pilot signals are still transmitted. Thus, the problematic cell belongs to crippled sleeping cell type and tends to become catatonic with time. For a network operator in many cases RACH problem would become visible only after a long observation time or even due to user complains. Thus, it is easy to underestimate the importance of timely detection of RACH failures in LTE networks.

III. Simulation Assumptions

Verification of the developed data mining framework is based on MDT measurements generated by means of dynamic LTE system simulator. This simulator has been designed according to 3GPP specifications from Releases 8, 9, 10 and partially 11. Step resolution of the simulator is one Orthogonal Frequency-Division multiplexing (OFDM) symbol.

Simulation scenario is created on the basis of 3GPP macro case 1 [10]. Investigated LTE network scenario is wrap-around and consists of 7 base stations with inter site distance of 500 meters. Each eNB contains 3 directed antenna sectors, resulting in 21 cells network layout, as it can be seen from slow faded dominance map on Fig. 1a. On Fig. 1 colors and numerical labels represent cell IDs. Modeling of propagation and radio link conditions includes slow and fast fading. The set of the main configuration parameters of the simulated network is shown in Table I. Mobile Terminals (MT) travel throughout the whole network with a random walk mobility model, as it is shown of Fig. 1b.

The effects of RACH sleeping cell failure is as follows: whenever UE tries to initiate random access to cell 1, this attempt fails. The malfunction area covers around 5% of the overall network.

As an output from the simulations we get two kinds of data files: MDT log and dominance map information. MDT log includes a lot of variables, but in this study we employed the following ones: MDT triggering events (listed in Table II), UE IDs and target cell IDs.

TABLE I
General Simulation Parameters

Parameter	Value
Cellular layout	Macro 21 wrap-around
Number of cells	21
Inter-Site Distance	500 m
Link direction	Downlink
Maximum BS TX power	46 dBm
Initial cell selection criterion	Strongest RSRP value
Simulation length	572 s (9.5 min)
Simulation resolution	1 time step = 71.43 μs
Max number of UEs/cell	20
UE velocity	30 km/h
Duration of calls	Uniform 30 to 140 s
Traffic model	Constant Bit Rate 320 kbps
Normal and Reference case	Simulation without sleeping cell
Problematic case	Simulation with RACH problem in cell 1

TABLE II
Network events triggering MDT log entry

PL PROBLEM – Physical Layer Problem [10].
RLF – Radio Link Failure [11].
A2 RSRP ENTER – RSRP goes under A2 enter threshold.
A2 RSRP LEAVE – RSRP goes over A2 leave threshold.
A3 RSRP – A3 event, according to 3GPP specification.
HO COMMAND RECEIVED – handover command received [11].
HO COMPLETE RECEIVED – handover complete received [11].

IV. Theoretical Basis

In order to pinpoint an outage cell we have developed sleeping cell detection framework based on advanced data mining techniques. The exploited algorithms are described in the following subsections.

The flowchart of the framework is depicted on Fig. 2. In a nutshell, the detection of the sleeping cell is done in two steps. First, outlier sub-calls are extracted by means of k-nn anomaly score outlier detection algorithm. Then the target cell feature of the outlier sub-calls is cosidered and sleeping cell scores are assigned to cells. The necessary thresholds and values are discovered from the normal dataset during the training phase.

A. Sliding window approach

Sliding window approach is a technique for slicing data instances into smaller pieces, which can then be considered as separate units. The slicing is done in such way that adjacent pieces have overlapping parts. The depth of the overlapping and size of a piece are set by two self-explanatory parameters – sliding window step and size.

The bottom line of sliding window technique is as follows. If there is an abnormal behavior in one part of the long time sequence of events, then the abnormality will be hidden bacause the rest and major part of the sequence is normal. However, if we cut a long sequence to smaller subsequences and consider

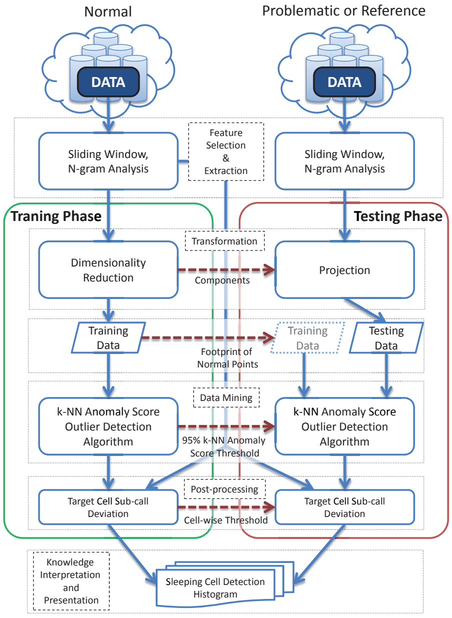

Normal
Problematic or Reference
DATA
DATA
Sliding Window, N-gram Analysis
Feature Selection & Extraction
Sliding Window, N-gram Analysis
Traning Phase
Testing Phase
Transformation
Dimensionality Reduction
Components
Projection
Training Data
Footprint of Normal Points
Training Data
Testing Data
Data Mining
k-NN Anomaly Score Outlier Detection Algorithm
95% k-NN Anomaly Score Threshold
k-NN Anomaly Score Outlier Detection Algorithm
Post-processing
Target Cell Sub-call Deviation
Cell-wise Threshold
Target Cell Sub-call Deviation
Knowledge Interpretation and Presentation
Sleeping Cell Detection Histogram

Fig. 2. Sleeping cell detection framework

them as single units, then the abnormal behavior of a small unit will be more noticeable. Therefore, the detection will be more robust. Besides, subsequences of events have the same lengths, what is in turn, eliminates the need of normalization.

B. N-gram analysis

N-gram analysis is one of the popular feature selection methods concerning sequential data. For instance, this method is utilized for the analysis of whole-genome protein sequences [12], for computer virus detection [13], and in many other fields.

N-gram analysis decomposes the original ordered dataset to small pieces. This new representation is stored in a feature vector. The feature vector conveys frequencies of how often each particular N-gram sub-sequence happens in the original dataset. Essentially, N-gram analysis is a sliding window with size equal to N and step is 1.

It is worth noting that the higher N is chosen, the more elaborate description of the original data is obtained. However, the increase of N raises the dimensionality of the feature space. Thus, the choice of N is a tradeoff between computational efficiency and the overall algorithm performance.

C. Dimensionality reduction and model order selection

Dimensionality reduction and model order selection method transform original matrices to the lower dimensional embedded space of new features.

In the current research Principal Component Analysis (PCA) and Minor Component Analysis (MCA) are utilized for the dimensionality reduction. Both these methods build the embedded space from the subset of eigenvectors of covariance matrix of an initial dataset. The eigenvectors corresponding to the biggest eigenvalues of the covariance matrix are referred to as principal components and taken by PCA. The eigenvectors with smallest eigenvalues are chosen by MCA. The formal and elaborate derivation of the components is proposed in [14].

The splitting point between biggest and smallest eigenvalues is defined by means of the method called Ratio of Adjacent Eigenvalues (RAE) [15]. RAE breaks the space composed of eigenvectors into the signal and noise subspaces. The maximum value of the ratio of adjacent eigenvalues corresponds to the splitting point.

D. k-NN anomaly score outlier detection algorithm

k-NN anomaly score outlier detection algorithm belongs to distance based anomaly detection methods. The anomaly score assigned to each point is the sum of distances to k nearest neighbors [16]. The rule of thumb is to set k equal to squared root of the number of points in the dataset.

Points with the largest anomaly scores are referred to as *outliers*. Anomaly score threshold is defined such that $n\%$ data points having the smallest anomaly scores are labeled as cluster and the rest are outliers. In other words, k-NN anomaly score threshold is $n\%$ quantile.

E. Post-processing

Post-processing phase extracts meaningful information out of the detected outlier sub-calls and assigns sleeping cell scores to each cell. These scores are represented as bars on a histogram called sleeping cell detection histogram.

In this study we have developed target cell sub-call deviation method. It is based on target cell ID feature of considered sub-calls. Finally, the influence of the order of N-gram analysis is evaluated in accordance to the suggested post-processing method.

At first, all normal sub-calls are stored during the training phase and are used while model is trained or tested. These normal sub-calls are considered to construct the vector of scores denoted by $a^{(N)}$. The number of entities is equal to the number of cells in the network. Entity is denoted by ${a_i}^{(N)}$. Each sub-call increments value of ${a_i}^{(N)}$ by one, if the corresponding sub-call's target cell feature value is equal to i, where i stands for cell ID. Finally, the calculated vector $a^{(N)}$ is normalize by the number of all normal sub-calls.

The very same procedure is implemented with the outlier sub-calls, and score vector $a^{(O)}$ is obtained.

The deviation of scores Δb is the absolute value of the difference of all normal and outlier sub-calls scores. Δb is defined by equation 1.

$$\Delta b = \left| a^{(N)} - a^{(O)} \right| \tag{1}$$

The post-processing method assigns high deviation scores to a sleeping cell and adjacent cells. The reason is the outlier sub-calls point to neighbors as well. In order to emphasize malfunctioning cell equation 2 is performed. It reduces the height of normal bars, while the height of the sleeping cell bar is increased.

$$c_i = \frac{\Delta b_i}{\sum\limits_{\substack{j \neq i \\ j \neq nbr(i)}} \Delta b_j} \tag{2}$$

where i and j are cell IDs, $nbr(i)$ contains the set of cell IDs neighboring to the cell i.

Equation 3 obtains sleeping cell scores scs_i, which characterize the behavior of each cell of the network. It is worth noting that cells are always assigned outage scores and their sum is equal to 100.

$$scs_i = \frac{c_i}{\sum\limits_{j} c_j} * 100 \tag{3}$$

The higher the score, the more chances for the cell to be detected as a sleeping cell. Since, sleeping cell scores are in the range from 0 to 100, it enables one to easily understand the degree of cell malfunction.

F. Performance Metrics

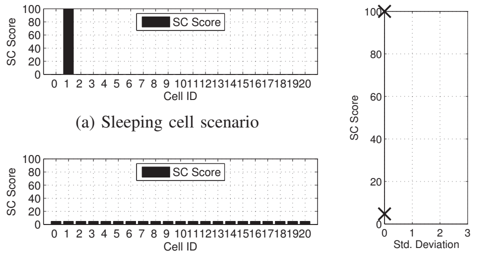
In order to assess the performance of sleeping cell detection algorithms we have developed qualitative and heuristic measures.

1) Qualitative measure: Qualitative performance measure estimates accuracy, precision, F-measure, True Negative Rate (TNR), True Positive Rate (TPR) and False Positive Rate (FPR) of a post-processing algorithm. This algorithm assigns sleeping cell scores to each cell. During the training phase the post-processing algorithm calculates a cell-wise normal threshold, which is mean plus three standard deviations of assigned sleeping cell scores. While the testing phase is run, if an assigned cell score is higher than the cell-wise normal threshold, then the cell is labeled as sleeping, otherwise it is normal. Afterwards, confusion matrix is drawn and above mentioned metrics are calculated.

2) Heuristic measure: Heuristic measure provides a quantitative estimation of sleeping cell detection histogram. The idea is the mapping of the histogram into a two dimensional heuristic space. It is taken into account that there is at most one sleeping cell in the network.

The heuristic mapping works as follows. The height of the highest bar of the sleeping cell detection histogram is referred to as *'Sleeping Cell Score'*. The standard deviation of heights of other bars is *'Standard Deviation'*. Thus, the histogram or post-processing algorithm is depicted by a point on *('Standard Deviation'; 'Sleeping Cell Score')* space.

The ideal sleeping cell detection algorithm for the network with 21 cells works as follows, see Fig. 3. Let us say the base station 1 has RACH failure. Then its sleeping cell score is going to be 100, while the rest cells' scores are equal to 0. In opposite, for the normal scenario all cells should be characterized by the same outage values. Therefore, sleeping cell histograms for the considered cases are mapped to points [0; 100] and [0; 100/21] on the heuristic space.



(a) Sleeping cell scenario

(b) Normal network scenario

(c) Heuristic space

Fig. 3. Heuristic performance measure

The heuristic performance of any algorithm is measured as Euclidean distance from the corresponding point to either point [0; 100] or point [0; 100/21] depending on the network scenario. The smaller the distance, the better the algorithm performance.

V. Experimental Results

The essence of supervised machine learning process is to train a decision making model by means of a labeled training dataset. Given training data as anomaly free dataset, we extract the normal behavior of the network. Training and testing processes are described in the following subsections and shown on Fig. 2.

Three independent datasets are obtained from the simulator. Normal and reference datasets are collected from the normally functioning network. The scenario with a malfunctioning cell 1 corresponds to the problem dataset. In order to get statistically valid results, each of three datasets is split into 6 sub-datasets.

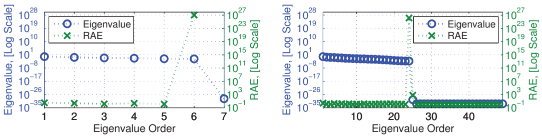
Simulator provides raw data, therefore preprocessing phase is performed. Calls are sliced to sub-calls by sliding window with size 15 and step 10, as it is described in Section IV-A. One more benefit of this step is that outlier sub-calls reveal the area of the sleeping cell more precisely, because they interact with fewer number of cells than the original calls.

1-gram and 2-gram feature selection algorithms are employed to construct 7 dimensional and 49 dimensional feature matrices respectively, see Section IV-B. 1-gram features are actual MDT events (Tabel II), whereas 2-gram features are ordered pairs of events.

A. Training phase

Sleeping cell detection model learns regular network behavior during the training phase. Basis vectors for the embedded space, footprint of normal sub-call in the low dimensional space, 95% k-NN anomaly score threshold and cell-wise sleeping cell detection threshold are extracted from the normal sub-datasets.

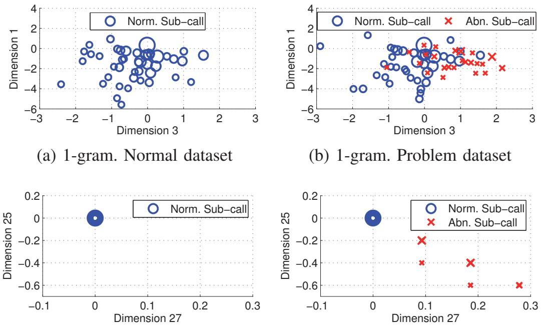
The embedded space is composed of eigenvectors of covariance matrix of normal data. Major components are utilized for 1-gram analysis, whereas minor components are for 2-gram



(a) 1-gram analysis

(b) 2-gram analysis

Fig. 4. Eigenvalues and RAEs of training dataset



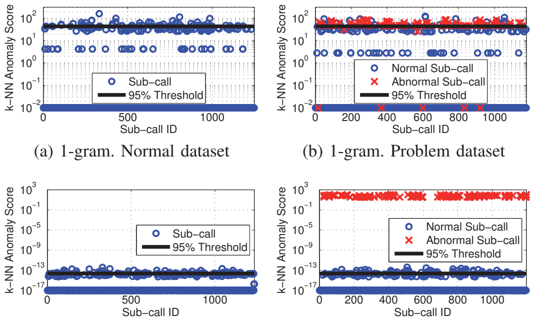
(a) 1-gram. Normal dataset

(b) 1-gram. Problem dataset

(c) 2-gram. Normal dataset

(d) 2-gram. Problem dataset

Fig. 5. Data in the embedded space



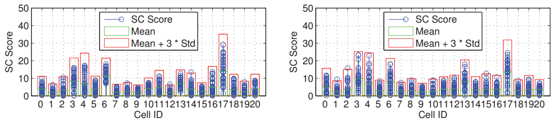
(a) 1-gram. Normal dataset

(b) 1-gram. Problem dataset

(c) 2-gram. Normal dataset

(d) 2-gram. Problem dataset

Fig. 6. k-NN anomaly score



(a) 1-gram analysis

(b) 2-gram analysis

Fig. 7. Cell-wise sleeping cell detection threshold. Target cell sub-call deviation method

analysis. Although, this aspect makes the comparison between 2 orders of N-grams different, there is a clear reason for it.

As long as 1-gram features are mobile terminal events, then, roughly speaking, all of them are evenly expected in both malfunctioning and properly functioning cellular networks. Thus, PCA is employed to simply cut off the noise space and to reduce the dimensionality.

2-grams are ordered pairs of network events. Some of them, for instance 'HO COMMAND RECEIVED - A2 RSRP ENTER', are uncommon for normal sub-calls and happen in problematic scenario. Hence, embedded space based on MCA is able to distinguish between normal and abnormal sub-calls.

Eigenvectors chosen by PCA and MCA are determined by RAE, see Fig. 4. Although, the dimensionality reduction has been implemented 1-gram PCA embedded space is four time smaller than 2-gram MCA embedded space.

The pattern of normal sub-calls in the subspace is depicted on Fig. 5a and 5c. Marker size shows the density of points.

The next step is the outlier detection based on k-NN anomaly scores, Section IV-D. As long as each sub-dataset contains about 1300 sub-calls, then k is set to 35. Besides, we define the anomaly score threshold as a value that cuts off 5% of normal sub-calls, having the highest k-NN anomaly score. In case of 1-gram analysis, the k-NN scores and 95% anomaly threshold of normal dataset are shown on Fig. 6a and 6c.

Finally, target cell sub-call deviation method is employed to obtain sleeping cell detection histograms. Having multiple normal sub-datasets, we capture the normal pattern for each cell. The detection threshold is shown on Fig. 7 and equal to the cell-wise mean plus three standard deviations of the sleeping cell scores.

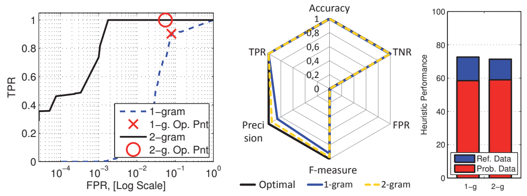
B. *Testing phase*

During the testing phase the sleeping cell detection framework is fed by problem or reference dataset as it is shown on Fig. 2. Decisions of the model are based on the normal patterns and thresholds extracted during the training phase.

The detection of a sleeping cell is performed in two steps. At first, outlier sub-calls are assigned by means of k-NN anomaly detection method. Afterwards, the post-processing method considers these outliers and provides sleeping cell detection histogram. The performance of sleeping cell detection method is defined in two ways, as it is written in Section IV-F.

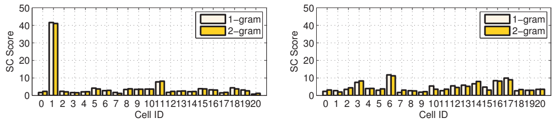
To measure the classification accuracy of k-NN, sub-calls are labeled as abnormal if they experienced a handover failure, otherwise normal. However, this labeling is not utilized to train the model. Abnormal sub-calls of problem dataset deviate from the normal instances in the embedded space, as it can be seen on Fig. 5. That is why k-NN assigns higher anomaly scores to abnormal sub-calls. ROC curves of k-NN outlier detection method are presented on Fig. 8a. As it can be seen 2-gram feature selection method provides better separation between normal and abnormal sub-calls, than 1-gram analysis. However, areas under ROC curves are high in both cases: 0.93 and 0.99 for 1-gram and 2-gram analyses respectively.

Having assigned outlier sub-calls, the framework executes the post-processing algorithm. Sleeping cell detection histograms of target cell sub-call deviation method for 1 and 2-gram analyses are compared on Fig. 9. Although, 2-gram analysis obtains better separation between normal and abnormal sub-calls, there is no a significant difference between the final histograms.



(a) ROC curves (b) Qualitative performance (c) Heuristic performance

Fig. 8. Performances. (a) k-NN anomaly score outlier detection algorithm, (b, c) target cell sub-call deviation method



(a) Testing is problem dataset (b) Testing is reference dataset

Fig. 9. Sleeping cell detection histogram. Target cell sub-call deviation method

From the performance point of view, the usage of 2-gram analysis provides slightly better results for the considered method, see Fig. 8b. However, this difference is very insignificant. For instance, heuristic performance, see Fig. 8c, increases by about 1% if 2-gram analysis is chosen instead of 1-gram analysis.

VI. Conclusion

In this paper we presented the data mining framework for the detection of cell outages, such as sleeping cell. The developed model has been verified by data gathered from dynamic LTE system simulator. Sleeping cell was modelled by means of random access channel failure.

The detection framework is based on data mining techniques, and performance metrics are suggested. Outlier sub-calls are assigned with help of N-gram analysis, dimensionality reduction and k-NN anomaly score algorithms. Afterwards, outlier sub-calls are allocated on the map, and outage cell is defined. The choice of N in N-gram feature selection algorithm is considered, as long as it has a great impact on computational efficiency of the overall framework.

References

[1] “Telecommunication management; Self-Organizing Networks (SON); Self-healing concepts and requirements,” 3GPP, TS 32.541, December 2012.

[2] G. F. Ciocarlie, U. Lindqvist, S. Nováczki, and H. Sanneck, “Detecting anomalies in cellular networks using an ensemble method,” in *CNSM*, 2013, pp. 171–174.

[3] M. Amirijoo, L. Jorguseski, R. Litjens, and L.-C. Schmelz, “Cell outage compensation in LTE networks: Algorithms and performance assessment.” in *VTC Spring*. IEEE, 2011, pp. 1–5.

[4] F. Chernogorov, J. Turkka, T. Ristaniemi, and A. Averbuch, “Detection of sleeping cells in LTE networks using diffusion maps.” in *VTC Spring*. IEEE, 2011, pp. 1–5.

[5] J. Turkka, F. Chernogorov, K. Brigatti, T. Ristaniemi, and J. Lempiäinen, “An approach for network outage detection from drive-testing databases.” *Journal Comp. Netw. and Communic.*, 2012.

[6] “UTRA and E-UTRA; Radio measurement collection for MDT; Overall description; Stage 2,” 3GPP, TS 37.320, March 2014.

[7] F. Chernogorov, T. Ristaniemi, K. Brigatti, and S. Chernov, “N-gram analysis for sleeping cell detection in LTE networks.” in *ICASSP*. IEEE, 2013, pp. 4439–4443.

[8] B. Cheung, S. Fishkin, G. Kumar, and S. Rao, “Method of monitoring wireless network performance,” Patent US App. 10/946,255, March, 2006.

[9] O. N. C. Yilmaz, J. Hämäläinen, and S. Hämäläinen, “Self-optimization of random access channel in 3rd generation partnership project long term evolution,” *Wireless Communications and Mobile Computing*, vol. 11, no. 12, pp. 1507–1517, 2011.

[10] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 10)*, 3GPP Std. TS 36.331, 2011.

[11] S. Sesia, M. Baker, and I. Toufik, *”LTE - The UMTS Long Term Evolution: From Theory to Practice”*. John Wiley & Sons, 2011.

[12] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy, and J. Klein-Seetharaman, “Comparative n-gram analysis of whole-genome protein sequences,” in *Proceedings of the Human Language Technologies conference*, 2002.

[13] J. Choi, H. Kim, C. Choi, and P. Kim, “Efficient malicious code detection using n-gram analysis and SVM,” in *Proceedings of the 2011 14th International Conference on Network-Based Information Systems*, ser. NBIS ’11. Washington, DC, USA: IEEE Computer Society, 2011, pp. 618–621.

[14] M. Timmerman, “Principal Component Analysis (2nd Ed.). I. T. Jolliffe,” *Journal of the American Statistical Association*, vol. 98, pp. 1082–1083, January 2003.

[15] F. Cong, A. K. Nandi, Z. He, A. Cichocki, and T. Ristaniemi, “Fast and Effective Model Order Selection Method to Determine the Number of Sources in a Linear Transformation Model,” in *Proceedings of the 2012 European Signal Processing Conference (EUSIPCO-2012), Bucharest, Romania*, August 27-31 2012.

[16] F. Angiulli and C. Pizzuti, “Fast outlier detection in high dimensional spaces,” in *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery*, ser. PKDD ’02. London, UK: Springer-Verlag, 2002, pp. 15–26.

PIII

ANOMALY DETECTION ALGORITHMS FOR THE SLEEPING CELL DETECTION IN LTE NETWORKS

by

Sergey Chernov, Michael Cochez, Tapani Ristaniemi 2015

81st IEEE Vehicular Technology Conference (VTC) Spring, Glasgow, Scotland

Reproduced with kind permission of IEEE.

Anomaly Detection Algorithms for the Sleeping Cell Detection in LTE Networks

Sergey Chernov, Michael Cochez and Tapani Ristaniemi
Department of Mathematical Information Technology
University of Jyväskylä, Finland
sergey.chernov@jyu.fi michael.cochez@jyu.fi tapani.ristaniemi@jyu.fi

***Abstract*—The Sleeping Cell problem is a particular type of cell degradation in Long-Term Evolution (LTE) networks. In practice such cell outage leads to the lack of network service and sometimes it can be revealed only after multiple user complains by an operator. In this study a cell becomes sleeping because of a Random Access Channel (RACH) failure, which may happen due to software or hardware problems. For the detection of malfunctioning cells, we introduce a data mining based framework. In its core is the analysis of event sequences reported by a User Equipment (UE) to a serving Base Station (BS). The crucial element of the developed framework is an anomaly detection algorithm. We compare performances of distance, centroid distance and probabilistic based methods, using Receiver Operating Characteristic (ROC) and Precision-Recall curves. Moreover, the theoretical comparison of the methods' computational efficiencies is provided. The sleeping cell detection framework is verified by means of a dynamic LTE system simulator, using Minimization of Drive Testing (MDT) functionality. It is shown that the sleeping cell can be pinpointed.**

I. Introduction

Ever increasing user demands to wireless communication services and the inevitable growth of mobile cellular network complexity pose ambitious challenges for mobile operators and the telecommunication research community. The high popularity of tablet computers and smartphones has led to a fierce competition among wireless network providers. To attract customers, the operators are forced to offer high quality services and more network traffic for less money. On the other hand, the configuration and maintenance of network equipment is a rather expensive process, which in many cases requires manual work of qualified specialists. One of the reasons is the heterogeneous nature of wireless networks, i.e. they combine different radio access technologies (HSPA, OFDM) and cell layers (macro, micro, pico). In order to cope with the high level of network complexity and to avoid an increase in capital expenditures, 3GPP has been developing the Self-Organizing Network (SON) concept since Release 8 [1].

The Self-Organizing Network concept introduced by 3GPP consists of three solutions, namely Self-Configuration, Self-Optimization, and Self-Healing [2]. Self-Configuration refers to the automated configuration of the newly established networks, while Self-Optimization includes changing the network parameters in order to meet the current demands on the network. The purpose of the Self-Healing mechanism is to detect and address problems automatically, avoiding significant impact on subscribers' experience and reducing operational expenses. In one of the earliest research efforts [3] neighbor cell list reports are considered in order to evaluate the network functionality. Although, the algorithm has shown good performance in failure detection, there is no possibility to detect problems with a low number of active users. In a recent work the authors of [4] propose anomaly detection and a diagnosis framework for mobile network operators.

Our current study is related to the sleeping cell detection problem caused by RACH failure. RACH failures can occur when there is no radio coverage outage. Therefore, instead of using radio environment measurements, we analyze sequences of events reported by Mobile Terminals (MTs). In [5] we proposed a data mining approach that enables automatic detection of malfunctioning cells. In a further study [6] we considered the effect of the size of the N-grams for the feature selection algorithm. Based on that research, we concluded that the influence of the size is minimal, and hence we choose N-grams of size 1 in this article.

In the current article we focus on another part of the designed data mining structure. The quality of the sleeping cell detection depends on the utilized data mining algorithm. Therefore, we compare, against each other, the application of different classification methods in our cell outage detection framework. The methods used are based on distance (k-Nearest Neighbors, k-NN), centroid distance (Self-Organizing Map, SOM), and probabilistic data structures (Local-Sensitive Hashing, LSH and Probabilistic Anomaly Detection, PAD). The practical comparison involves analysis of ROC and Precision-Recall curves. Finally, we present theoretical bounds of computational complexities of the considered algorithms.

The rest of the paper is organized as follows. We provide the description and simulation details of the sleeping cell detection problem and RACH failure as its particular case in Section 2. A short overview of the utilized simulation tool and related assumptions is given in the same section. Afterwards, the cell outage detection framework and its functional parts are presented in Section 3. Section 4 demonstrates the main results and compares performances of the considered anomaly detection methods. Finally, Section 5 concludes the paper.

II. Sleeping Cell Problem and Simulation

A. Sleeping Cell Problem

Sleeping cell is a situation whereby a base station's malfunctioning is not detected by the operator as there is no alarm triggered. There could be multiple reasons for this to happen, including misconfiguration, excessive load or software/firmware problem at the base station side. The main danger is that the failure remains invisible for the network

978-1-4799-8088-8/15/$31.00 ©2015 IEEE

maintainers and would be revealed only after a number of complaints from subscribers, in other words unsatisfied customers. There are three types of sleeping cells. *Impaired* and *crippled* cells show, respectively, slightly and significantly lower performances than expected. The most critical type is a *catatonic* cell, which leads to the complete absence of service.

In our study we consider sleeping cell caused by RACH failure. This type of failure makes it impossible for users to establish a connection or to make a handover to the malfunctioning cell. However, MTs previously connected to the sleeping cell do not experience any lack of service. Hence, the cell starts out being impaired since only few users notice problems, but over time it becomes catatonic since it does not serve any MTs. Typically, a RACH problem would be detected only after long observation time or after multiple users complaints. Thus, a timely detection of RACH failures is an important issue in LTE networks.

B. Simulation

The simulator used in our experiments has been designed in accordance with LTE 3GPP specifications and is utilized by the Nokia Network research group. The data are MDT measurements generated using an advanced LTE system level simulator.

The simulations have been run several times for a duration of 572 simulated seconds each. The first simulation was conducted without cell outages and corresponds to the normal operation of the network. The normal patterns of the MDT logs are learned and captured by the data mining framework. Cell 1 happened to have a RACH failure in the beginning of the second run. In spite of the failure, Cell 1 was able to provide services to the connected MTs. However, it finally became a *catatonic* sleeping cell.

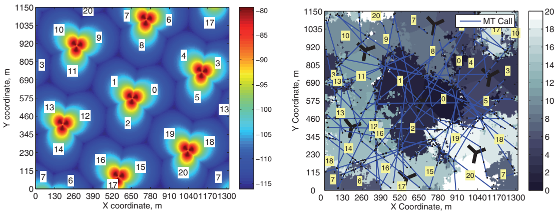
The utilized scenarios consist of 7 sites, respectively 21 cells, with a BS distance of 500 meters, see fig. 1(a). The model of the radio propagation environment includes slow and fast fading conditions. MTs were initialized uniformly random on the map and traveled under a random walk mobility model. Table I contains the whole list of the simulation parameters.

The simulations provided us with two kinds output: MDT logs and signal strength information. MDT logs are described by timely ordered sequences of variables, but we have employed MDT triggering events (listed in table II), MT IDs and target cell IDs. Signal strength information reports the distribution of the cells' signal strengths on the whole map. Hence, the dominance areas of the cells can be drawn, see fig. 1(b).

III. Cell Outage Detection Framework

The whole sleeping cell detection framework involves two separate procedures, see fig. 2. At first, the model learns a "normal" network behavior. It practically means that the footprint of normal sub-calls in a feature space and necessary thresholds are extracted. Afterwards, the current network state is compared against the previously trained model, and cell outage predictions are made.

In contrast to our previous work [6] we have not employed dimensionality reduction. Actually, there is no high need for it, since data are fully described by seven 1-gram features.



(a) ***Network map. SINR*** **(b)** ***Mobile terminal calls on slow faded dominance map***

Fig. 1. Wrap around Macro 21 scenario

TABLE I. General Simulation Parameters

Parameter	Value
Cellular layout	Macro 21 wrap-around
Number of cells	21
UEs per cell	15
Inter-Site Distance	500 m
Link direction	Downlink
User distribution in the network	Uniform
Maximum BS TX power	46 dBm
Initial cell selection criterion	Strongest RSRP value
Simulation length	572 s (≈ 9.5 min)
Simulation resolution	1 time step = 71.43 μs
Max number of UEs/cell	20
UE velocity	30 km/h
Duration of calls	Uniform 30 to 140 s
Traffic model	Constant Bit Rate 320 kbps

TABLE II. Network events triggering MDT log entry

A2 RSRP ENTER — RSRP goes under A2 enter threshold.
A2 RSRP LEAVE — RSRP goes over A2 leave threshold.
A3 RSRP — A3 event, according to 3GPP specification.
HO COMMAND RECEIVED — handover command received [7].
HO COMPLETE RECEIVED — handover complete received [7].
PL PROBLEM — Physical Layer Problem [8].
RLF — Radio Link Failure [7].

In the current research we focus on the choice of an anomaly detection algorithm as the main functional block of the scheme. The utilized methods are based on distance (k-NN), centroid distance (SOM), and probabilistic data structures (LSH, PAD). The practical comparison is carried out by means of ROC and Precision-Recall curves. The description of the other functional parts is provided in the following sub-sections.

A. Sliding Window

Each MT causes an ordered sequence of events in the LTE network. Due to the random nature of calls, the sequences have different lengths. In order to do length-wise normalization and localize calls, we slice the sequences to sub-sequences, or sub-calls, by means of a sliding window. The chosen window step and size allow a sub-call to visit three cells on average.

B. N-Gram Analysis

In order to extract features from sub-calls, which are sequential data, we make use of N-gram analysis. An N-gram is an ordered set of N terms. The number of N-grams

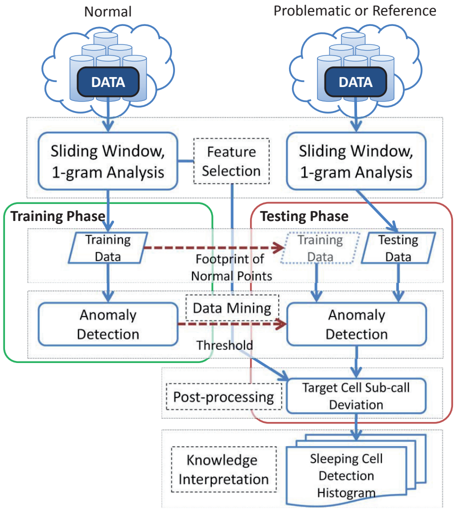


Fig. 2. Sleeping cell detection framework

is defined as all possible permutations of unique instances in the considered data. In our research, the instances are network events, see table II. Thus, sub-calls can be transformed into feature vectors, which indicate how often each particular N-gram occurs in these subsequences.

C. Anomaly Detection Algorithms

We define *abnormal* sub-call as sub-call that moved into or travelled through a sleeping cell, otherwise the sub-call is referred as *normal*. An *outlier* sub-call is a sub-call detected as abnormal and a *non-outlier* sub-call is one detected as normal. Next, we introduce binary classification methods, which split the sub-calls into the *outlier* and *non-outlier* classes.

1) k-NN anomaly score outlier detection algorithm: this algorithm belongs to the family of distance based anomaly detection methods. The sum of distances to the k nearest neighbors of a point is referred to as k-NN anomaly score of that point [9]. In practice, k is usually chosen to be squared root of the number of points in the dataset.

The higher the k-NN anomaly score the more chances for a point to be assigned to the *outlier* class. The exact decision boundary is introduced as $n\%$ quantile of normal dataset. Thus, the defined threshold splits the considered dataset into two classes: *non-outlier* class, having low anomaly scores, and *outlier* class, having high anomaly scores.

2) Self-organizing map: SOM is a set of connected nodes which "memorizes" the shape of a larger set of points [10]. SOM can be used for binary classification problems by considering points which are located in the interior of the map being member of the one class and points which fall outside it in the other.

For a given set of points, in our case the set of normal sub-calls, the map is created by morphing a set of interconnected nodes such that it covers the points. The morphing is an iterative process and happens by moving the best matching unit, i.e. closest node of the map, closer towards given random training point. This also affects the connected nodes to a smaller extend. We utilized a two dimensional, linearly initialized self-organizing map with a hexagonal lattice. [1] SOM anomaly scores are assigned by mean of k-NN.

3) Probabilistic anomaly detection: PAD calculates density functions of features of the normal dataset [11]. The untypical or abnormal behavior can be defined as an event with a small probability. There are two types of exploited density functions, referred to as consistency checks. First and second order consistency checks estimate, respectively, unconditional and conditional probabilities of elements. The Friedman-Singer estimator was used for the probability calculations. It allowed us to estimate the probabilities for already known as well as previously unknown elements.

During the training phase, PAD is fed by data representing normal network behavior, and consistency checks are calculated. A testing record is labeled as abnormal if it has not passed one of the consistency checks, i.e. the probability of one of its elements is lower than a predefined threshold.

4) Locality-sensitive hashing: LSH first introduced as a method for finding approximate nearest neighbors, given a distance measure, d, and a threshold for the error, ϵ [12]. In order to use the method, one needs a family of hash functions, which likely map elements with a high similarity to the same value and elements with low similarity to different ones. For our experiment, the Jaccard distance measure ($d\left(A, B\right) = 1 - sim\left(A, B\right)$) and its associated family of locality sensitive hash functions, namely min-hash, were used.

Concrete, to test whether a sub-call is abnormal, we use LSH and try to find near neighbors in our training data set. If these are not found, we conclude that the sub-call is abnormal.

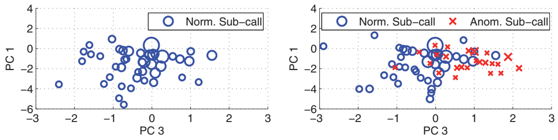
D. ROC and Precision-Recall curves

Common ways to represent the performance of binary classifiers are ROC and Precision-Recall curves.

- *Precision* is the number of *abnormal* sub-calls, classified as *outliers*, divided by the total number of *outlier* sub-calls.
- *Recall* or *True Positive Rate* (TPR) is the number of *abnormal* sub-calls, classified as *outliers*, divided by the number of *abnormal* sub-calls.
- *False positive rate* (FPR) is the number of *normal* sub-calls, classified as *outlier*, divided by the number of *normal* sub-calls.

ROC curve represents TPR in function of FPR, while Precision-Recall shows the relation between precision and recall. A better algorithm will have a higher precision and recall and a lower FPR. We used 6-fold cross validation and included information about the standard deviation in the curves. The ROC curves also contain the operating point which shows the actual algorithm's performance. This point corresponds to the threshold used to split the sub-calls.

[1] We used the SOM Toolbox from http://www.cis.hut.fi/projects/somtoolbox/



(a) ***Normal dataset*** **(b)** ***Problem dataset***

Fig. 3. Data in the feature space. Marker size represents the density of points

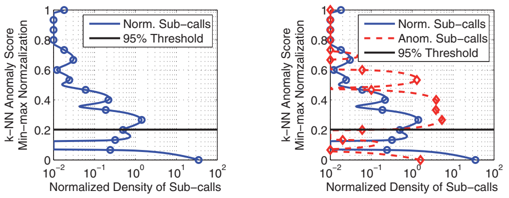
E. Post-processing

Post-processing is implemented by the *target cell deviation of sub-calls* algorithm, which makes use of target cell feature of sub-calls. Two groups of sub-calls are taken into account. The first one consists of all instances of the training dataset. The second one is composed of the sub-calls assigned to the *outlier* class by the anomaly detection algorithm during the testing phase. Next, two histograms representing the distribution of target cell feature among cells are calculated and normalized by the corresponding number of sub-calls. The deviation histogram is obtained as the absolute difference between testing and training histograms.

The deviation histogram is not the final evaluation of the network performance. A sub-call in itself hits several cells. Hence, the *outlier* sub-calls can indicate as malfunctioning not only the actually sleeping cell, but also its neighbors. The proposed amplification procedure increases a sleeping cell's bar on the deviation histogram, while decreasing the neighboring bars. It is worth to point out that if there is no cell outage then the deviation histogram is not significantly changed by the amplification. Finally, the amplified deviation histogram is normalized and introduced as sleeping cell detection histogram, see fig. 6.

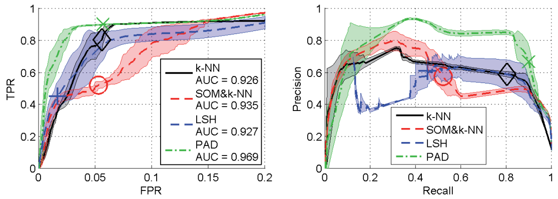
In accordance to our detection framework the ordered event sequences generated by MTs are sliced into pieces by sliding window with size 15 and step 10. 1-gram analysis construct the set of features, which are basically network events listed in table II. Normal and problem datasets projected on two principal components of the problem dataset is plotted in fig. 3. This first analysis of the data set shows that it is not trivial to separate the *abnormal* sub-calls from *normal* ones.

An anomaly detection algorithm both assigns anomaly scores to dataset's points and separates them into *outlier* and *non-outlier* classes. In fig. 4 we plotted the normalized density of sub-calls in function of the k-NN anomaly score. Note that in the figures we swapped the X and Y axis and that the density is plotted on a logarithmic scale. A black line, located around anomaly score 0.2, separates the figure in two areas. This decision threshold is determined during the training phase, such that the anomaly scores below the line represent 95% of the points in the normal dataset, while the remaining 5% are located above the line. The chosen threshold is mapped to ROC and Precision-Recall curves as an operating point.



(a) ***Normal dataset*** **(b)** ***Problem dataset***

Fig. 4. Min-max normalization of k-NN anomaly scores



(a) ***ROC curves*** **(b)** ***Precision-Recall curves***

Fig. 5. Performances of anomaly detection algorithms. Colored areas show 3 standard deviations of the lines. Makers represent operating points: "◇" — k-NN, "◯" — SOM&k-NN, "×" — PAD, "+" — LSH

Figure 4(b) shows how the decision threshold works on *i*) *normal* sub-calls *ii*) *anomalous* sub-calls from the problem dataset. Note that many of the anomalous points obtain a high anomaly score and will hence be correctly classified as *outliers*. The points of curve of *anomalous* sub-calls under the threshold indicate sub-calls which will be erroneously classified as *non-outliers*. The figure also shows the source of false positives. These are sub-calls which are represented by the normal curve above the threshold. Similar figures can be plotted for the other anomaly detection algorithms, but we left them out for brevity.

We compare the performance of the different anomaly detection methods visually using ROC curves in fig. 5(a) and Precision-Recall curves in fig. 5(b). Note that we show only the most interesting part of the ROC curves. Points with a FPR greater than 0.2 result in pretty much the same TPR for all methods. For interpretation of these figures, we remind that an optimal method would find a TPR of 1 and a FPR of 0 and hence the ROC curve would have a point in the top-left corner of the chart. For the Precision-Recall curve the optimal method would find a precision and recall equal to one corresponding to a point in the top-right corner.

From the curves, we see that the PAD method performs better as the other methods. SOM&k-NN, despite the fact that it does not have the smallest AUC, could be considered the weakest method. In order to get the TPR at an acceptable level, one has to accept a high false positive rate. k-NN as such scores fairly strong. Its main benefit is that is very stable which can be seen from the small standard deviations. The LSH method scores similar to k-NN in the regions of interest, i.e. these with high precision and recall. However, the method seems somewhat unstable. In other words, when performing analysis using LSH on a limited dataset, the result might be unreliable.

Another consideration regarding the used anomaly detection methods is their computational complexity. In a real-life scenario the resource budget allocated for a Self-Healing algorithm might be of great importance. In table III we listed the computational complexities of the training and testing

TABLE III. COMPUTATIONAL COMPLEXITIES OF ANOMALY DETECTION ALGORITHMS

Algrithm	Training phase	Testing phase
k-NN [13]	$O(dN_{trn}^2)$	$O(dN_{trn}N_{tst})$
SOM & k-NN [14]	$O(duN_{trn})$	$O(d\sqrt{u}N_{tst})$
LSH	$O(brN_{trn})$	$O(brN_{tst})$
PAD [15]	$O(v^2N_{trn}^2)$	$O(N_{tst})$

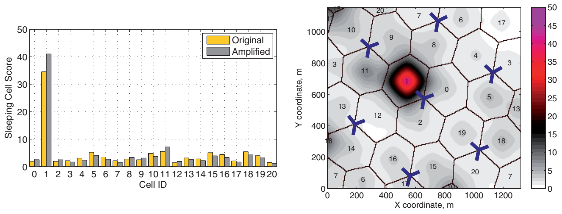
N_{trn}, N_{tst} — number of instances in the training and testing dataset respectively

d — number of N-gram features or dimensions

u — number of unique instances in N_{trn}

v — number of unique record values for each record component

b and r — number of bands and rows used for LSH



(a) ***Sleeping cell detection histogram*** **(b)** ***Heat map***

Fig. 6. Sleeping cell detection

phases. As we can see, LSH is the only method which, with fixed values for b and r, remains linear for both phases. Other methods show a roughly quadratic complexity during training. During testing, k-NN loses out with again a quadratic complexity compared to a linear complexity for PAD. SOM&k-NN outperforms k-NN and shows slightly better result than during its training. The PAD method seems to utilize the least computation resources to label testing dataset.

Figure 6 shows the anomaly scores obtained using k-NN anomaly detection method. Note that the bars at Cell 1 are significantly higher as what is found for the other cells. From the histogram it is obvious that we are able to detect the sleeping cell correctly.

Even though we observe differences in the anomaly detection methods, we note that all of them were able to determine the sleeping cell correctly. Hence, we find that the performance of the overall detection framework is not strongly dependent on the choice of the classifier. Likely, the post-processing method cancels out the small differences between the considered anomaly detection methods.

IV. CONCLUSION

In this paper we introduced a data mining framework for sleeping cell detection, caused by RACH failure. One of the crucial parts of the framework is the anomaly detection algorithm. We compared computational complexity as well as practical performances of several algorithms based on a simulation. The considered algorithms have shown different detection rates on the simulated scenario. PAD showed the best result for the detection of abnormal sub-calls, but it is computationally expensive to train. Finally, we noticed that the utilized post-processing algorithm mitigated the differences between methods' performances.

ACKNOWLEDGMENT

The authors would like to thank Fedor Chernogorov from Magister Solutions Ltd. for providing the simulation data. Further, the work is made financially possible by the Graduate School in Electronics, Telecommunications and Automation. Also authors thank the Foundation of Nokia Corporation for their support.

REFERENCES

[1] 3GPP, "Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements," 3rd Generation Partnership Project (3GPP), TS 32.500, Sep. 2014.

[2] S. Hämäläinen, *LTE Self-Organising Networks (SON)*. Wiley,, 2012.

[3] C. M. Mueller, M. Kaschub, C. Blankenhorn, and S. Wanke, "A cell outage detection algorithm using neighbor cell list reports," in *IWSOS*. Springer-Verlag, 2008, pp. 218–229.

[4] S. Nováczki, "An improved anomaly detection and diagnosis framework for mobile network operators," in *DRCN'13*, 2013, pp. 234–241.

[5] F. Chernogorov, T. Ristaniemi, K. Brigatti, and S. Chernov, "N-gram analysis for sleeping cell detection in lte networks." in *ICASSP*. IEEE, 2013, pp. 4439–4443.

[6] S. Chernov, F. Chernogorov, D. Petrov, and T. Ristaniemi, "Data mining framework for random access failure detection in. lte networks." in *PIMRC*. IEEE, 2015.

[7] S. Sesia, M. Baker, and I. Toufik, *"LTE - The UMTS Long Term Evolution: From Theory to Practice"*. John Wiley & Sons, 2011.

[8] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification (Release 10)*, 3GPP Std. TS 36.331, 2011.

[9] F. Angiulli and C. Pizzuti, "Fast outlier detection in high dimensional spaces," ser. PKDD. Springer-Verlag, 2002, pp. 15–26.

[10] T. Kohonen, M. R. Schroeder, and T. S. Huang, Eds., *Self-Organizing Maps*, 3rd ed. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2001.

[11] P. Macioek, P. Krl, and J. Kolak, "Probabilistic anomaly detection based on system calls analysis," *Computer Science*, vol. 8, 2007.

[12] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, 1998, pp. 604–613.

[13] Nearest neighbors. [Online]. Available: http://scikit-learn.org/stable/modules/neighbors.html

[14] H. Kusumoto and Y. Takefuji, "Self-organizing map algorithm without learning of neighborhood vectors." *IEEE Transactions on Neural Networks*, vol. 17, no. 6, pp. 1656–1661, 2006.

[15] S. J. Stolfo, F. Apap, E. Eskin, K. Heller, S. Hershkop, A. Honig, and K. Svore, "A comparative evaluation of two algorithms for windows registry anomaly detection," *J. Comput. Secur.*, vol. 13, no. 4, pp. 659–693, Jul. 2005.

PIV

LOCATION ACCURACY IMPACT ON CELL OUTAGE DETECTION IN LTE-A NETWORKS

by

Sergey Chernov, Dmitry Petrov, Tapani Ristaniemi 2015

11th IEEE International Wireless Communications & Mobile Computing Conference (IWCMC), Dubrovnik, Croatia

Reproduced with kind permission of IEEE.

Location Accuracy Impact on Cell Outage Detection in LTE-A Networks

Sergey Chernov
Dept. of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
sergey.chernov@jyu.fi

Dmitry Petrov
Magister Solutions Ltd.
Jyväskylä, Finland
dmitry.petrov@magister.fi

Tapani Ristaniemi
Dept. of Mathematical Information Technology
University of Jyväskylä
Jyväskylä, Finland
tapani.ristaniemi@jyu.fi

***Abstract*—Automated and timely detection of malfunctioning cells in Long-Term Evolution (LTE) networks is of high importance. Sleeping cell is a particular type of cell degradation hardly detectable by traditional network monitoring systems. Recent introduction of Minimization of Drive Test (MDT) functionality enables to collect user-level statistics from regular user devices without expensive and time-consuming drive-test and measurement campaigns. In this study data mining techniques are used to process MDT measurements to detect efficiently a sleeping cell. The developed earlier data mining framework is briefly overviewed in the paper. Special attention is devoted to post-processing stage as one of the key elements of the detection scheme. In practice, location information of collected measurements might contain considerable errors. This factor impacts the precision of malfunctioning cell detection. Therefore several post-processing algorithms are proposed, where location accuracy is taken into account. The performance of the algorithms is compared based on the results of thorough system-level LTE network simulations. Combined post-processing method shows the best reliability against location errors in terms of Root Mean Squared Error (RMSE) and percent gain.**

***Keywords*—LTE, SON, Self-healing, cell outage, data mining, anomaly detection.**

I. Introduction

The complexity of modern and constantly evolving mobile cellular networks implies difficulties in their configuration and maintenance. This task usually requires considerable capital and operational expenses. For that reason, there is a strong need for the automation of the network management functions. The research activities related to this field are referred as Self-Organizing Networks (SON). So far, 3GPP has already included SON concept, requirements and solutions in the series of cellular network standards [1, 2].

Automatized and prompt detection of outages in current Long Term Evolution (LTE) and LTE-Advanced (LTE-A) networks is a challenging problem. For cellular networks operators it is important to:

- be aware about the problems in the network as fast as possible;
- identify the problem with minimal human involvement, for instance without the visit of maintenance personnel or without extensive measurement campaign;
- minimize amount of false and missed alarms;
- make the ”set-up” or training phase of the algorithms as short as possible;
- minimize the amount of manually set parameters of network performance monitoring system.

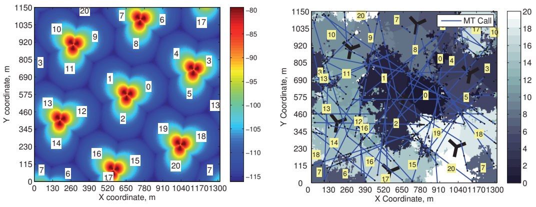
Presented targets can be combined only with the use of intelligent network performance monitoring system. The accuracy and efficacy of such analytic system directly depends on timely and precise information about the network.

Traditionally, cell-level key performance indicators (KPI) are used in network monitoring. Such approach for analyzing failures in cellular networks is introduced in [3]. Two Bayesian estimators are derived and suggested to be adopted by network operators as robust cell KPIs. In paper [4] adaptive ensemble method framework is presented and used to detect partial and complete cell degradation. The base station performance status is determined by cell-level KPIs.

Recently with the introduction of MDT functionality it became possible to collect user-level statistics from regular user devices. Such approach can replace expensive and time-consuming drive-test and measurement campaigns. It also provides much larger volume of data for analysis. For that reason more advanced data processing techniques based on data mining should be implemented. In particular, these measurements can be used to detect outages in the network. Corresponding anomaly detection algorithms have been proposed and analyzed in the number of papers [5–7]. Using statistical classification techniques and neighbor cell lists reported by mobile terminals (MT) authors of [5] developed an algorithm for the detection of cell outages. In paper [6] diffusion maps and nearest neighbors classification algorithm are utilized to pinpoint a malfunctioning cell. The designed method is based on certain features extracted from periodical MDT logs.

In our papers [7, 8] we have proposed complete data mining framework for cell outage detection. Malfunctioning was caused by a failure in Random Access Channel (RACH) procedure. Due to this outage MTs were not able to establish a connection or to make a handover to the sleeping cell. The developed machine learning model was based on MDT measurements and successfully pinpointed the malfunctioning cell. Our following articles [9, 10] have improved certain components of the initial framework. In particular, the influence

978-1-4799-5344-8/15/$31.00 ©2015 IEEE



(a) Network map. SINR

(b) Mobile terminal calls on slow faded dominance map

Fig. 1. Network map

of N in N-gram feature selection algorithm on the overall framework performance was discovered in [9]. In [10] the application of different types of anomaly detection algorithms was compared to our model.

In this paper we consider the proposed earlier detection framework [7, 8], RACH failure, and tackle a new topic - precision of measurements. In the previous studies it was assumed that location of MT is known exactly. However, due to the natural reasons the error in MT location estimation may change from few up to hundreds of meters. At the post-processing stage locations of abnormal MTs are used to pin out faulty cell. MTs can be localized by means of either target cell feature or by their geographical location. No superior post-processing method can be found out, because methods' performances depend differently on positioning accuracy. For that reason a new combined post-processing methods is proposed. It demonstrates higher reliability in case of errors in MT and measurements position.

The paper is arranged in the following way: In the next section simulation scenario and the scope of MDT measurements are introduced. It is also described how MT can be localized and what positioning error model is used. Section III overviews a framework that leverages knowledge data discovery process and detects a sleeping cell in the simulated scenarios. As for post-processing, there is a detailed description of three independent and one combined methods. The comparison metrics are introduced and results are analyzed in Section IV. The final section concludes the paper.

II. Outage simulation scenario

Computer simulation is a widely adopted approach to study big enough wireless networks, whereas real experiments are usually much more expensive or even impossible. In our research we made use of a system-level simulator, which meets LTE-A 3GPP specifications and is utilized by the Nokia Networks research group. The considered cellular network is composed of 7 base stations as shown on Figure 1(a). There are 3 sector antennas deployed on each base station. Thus, the whole network consists of 21 cells. Besides, the considered map and propagation model is endless due to wrap-around technique. Specifically, if MT during its random walk goes behind the network border, then it immediately appears on the other side of the map without any interruption. Other general simulation parameters are listed in Table I.

TABLE I. General Simulation Parameters

Parameter	Value
Cellular layout	Macro 21 wrap-around
Number of cells	21
UEs per cell	15
Inter-Site Distance	500 m
Link direction	Downlink
User distribution in the network	Uniform
Maximum BS TX power	46 dBm
Initial cell selection criterion	Strongest RSRP value
Simulation length	572 s (≈ 9.5 min)
Simulation resolution	1 time step = 71.43 μs
Max number of UEs/cell	20
UE velocity	30 km/h
Duration of calls	Uniform 30 to 140 s
Traffic model	Constant Bit Rate 320 kbps

TABLE II. Network events reported in MDT

Event	Comment
A2 RSRP ENTER	RSRP goes under A2 enter threshold.
A2 RSRP LEAVE	RSRP goes over A2 leave threshold.
A3 RSRP	A3 event, according to 3GPP specification.
HO COMMAND RECEIVED	handover command received [12].
HO COMPLETE RECEIVED	handover complete received [12].
PL PROBLEM	Physical Layer Problem [13].
RLF	Radio Link Failure [12].

In the study two types of network scenarios have been considered: normal and problem. As for normal ones, the cellular system worked in a proper way, i.e. eNBs and MTs did not experience any malfunctioning during the run time. The distinctive feature of the problematic scenario is the existence of a faulty eNB. Sleeping cell is introduced in the problematic scenario. It is a special type of a cell failure that cannot not be detected by traditional methods based on cell-level KPIs monitoring. In our study a cell turned to be sleeping because of RACH failure. This kind of malfunctioning prevents MTs from the ability to connect to or to make a handover to impaired cell, while previously connected MTs are still serviced.

A. MDT Measurements

MDT was introduced as a part of coverage and capacity optimization in self-organizing networks [11]. One of the major goals of MDT concept is to replace manual and expensive drive testing data collection process. MDT measurements are reported by MTs and provide actual information about network quality in the target area.

In the study MDT measurements with corresponding time and location stamps have been periodically reported by the simulated MTs. MDT logs are composed of signal strength information, target cell and MDT triggering events, presented in Table II. Collected MDT events are ordered in time and space sequences, which lay at the core of the actual sleeping cell detection framework. These measurements can be contributed by the propagation map corresponding to each of eNBs. Hence dominance cell maps can be derived, showing eNB with maximal reception power in each geographical point. Figure 1(b) shows an example of such a map, where path loss and slow fading are taken into account.

B. Localization

MT localization is of importance for sleeping cell detection. In case of network faults the current serving cells cannot always be used as a reliable indicator of the position. For example, MT might be still connected to eNB with very low signal strength, because it just cannot handover any other cell. Therefore, in our algorithms it is proposed to use two different ways in order to find relation between events and network layout. First, MT can be matched with a cell in accordance with a target cell of an event. For example, target cell can be easily defined for such events as A3, handover, etc. Second, MT localization can be related to the dominance areas of eNBs, where the signal power of the cell exceeds the signal power of the neighboring cells. Therefore, if we know the coordinates of a terminal and network prorogation map, we are able to define dominance cell for a MT. Dominance area is equal to the dominance cell ID.

The geographical location of a MT in practice can be determined by a number of methods, for example listed in [14]. Although, Assisted-General Positioning System (A-GPS) measurements provide the highest accuracy for outdoor positioning, it is not the superior approach. Since A-GPS receivers are usually installed in modern smartphones and cause high device battery consumption, they might be not present in MT or switched off. Therefore, we cannot expect, that is possible to get high location accuracy for all MTs in cell outage detection algorithms.

The distribution of horizontal positioning error is well approximated with Rayleigh noise, as it is shown in the report submitted to Federal Aviation Administration [15]. In order get realistic result, it was decided to add Rayleigh noise to actual MTs coordinates in our simulations. Rayleigh probability density function is

$$f(x, \sigma) = \frac{x}{\sigma^2} e^{-x^2/(2\sigma^2)}, x \geq 0, \tag{1}$$

where noise mean value and variance are defined by scale parameter σ.

III. Anomaly Detection Framework

In this section we present the framework, see Figure 2, that leverages knowledge data discovery process and detects a sleeping cell in the simulated scenarios. Leaving aside a data collection process, the developed framework can be broken into several parts: preprocessing and transformation, data mining and post-processing. The more detailed description of these steps is given in the following subsections.

The detection of the sleeping cell is carried out in two phases. The first phase is referred as training. During the training process, the overall scheme is fed by normal dataset, see Figure 2. Normal data is the data that reflects the regular, failure free, network behavior. Training phase enables the detection framework to learn the profile of normal functioning of the network.

The second phase is for testing. This time the anomaly detection framework processes previously unseen datasets. Records of these datasets assigned to an outlier class are considered by post-processing algorithms. Finally, a sleeping cell detection histogram provides outage scores for each cell and, hence, explains the current network state.

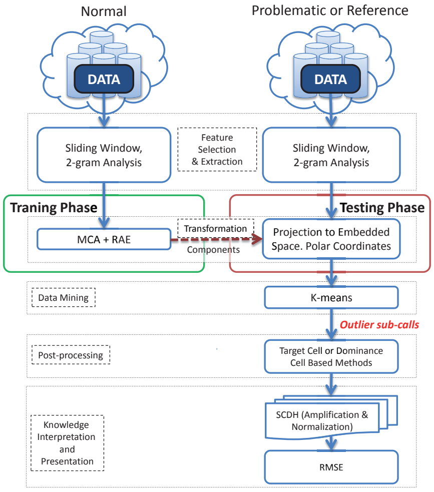

Normal
Problematic or Reference
DATA
DATA
Sliding Window, 2-gram Analysis
Feature Selection & Extraction
Sliding Window, 2-gram Analysis
Traning Phase
Testing Phase
MCA + RAE
Transformation
Components
Projection to Embedded Space. Polar Coordinates
Data Mining
K-means
Outlier sub-calls
Post-processing
Target Cell or Dominance Cell Based Methods
Knowledge Interpretation and Presentation
SCDH (Amplification & Normalization)
RMSE

Fig. 2. Sleeping cell detection framework

A. Preprocessing

Preprocessing is a data preparation process, which is employed during both training and testing phases. Preprocessing is carried out by sliding window and N-gram analysis methods.

Calls have different duration and consist of the different number of MDT events. Having records of different scale is a common situation in the data mining world. One usually applies normalization procedures, such as min-max or z-score. In our study we slice calls into equal pieces named sub-calls by sliding window with a fixed size and step. The size is chosen such that one sub-call visits 2 to 3 cells in average.

N-gram analysis is a feature selection procedure utilized for the analysis of sequential data. One can think about N-gram method as a sliding window of size N and step 1. Every record is broken into smaller pieces by this sliding window. These pieces or N-grams characterize a record by means of a feature vector. The feature vector preserves frequencies of each N-gram in the considered record. Thus, all sub-calls can be described by feature vectors, which form a feature matrix.

B. Transformation

High dimensional data require extra computational resources. That is why we have employed Minor Component Analysis (MCA) in pair with Ration of Adjacent Eigenvalues (RAE) methods that map the original manifold of points into a low dimensional space.

MCA projects dataset to an embedded space composed of minor components. Minor components are learned from normal dataset during the training phase as follows. At first, eigenvectors and eigenvalues of the covariance matrix of training dataset are calculated. The formal derivation is described in [16]. Afterwards, eigenvectors corresponding to the lowest eigenvalues are chosen as minor components. Normal points in such embedded space are located in the vicinity of the origin.

The number of extracted minor components is defined by Ration of Adjacent Eigenvalues (RAE) [17]. The maximum value of the ration of adjacent eigenvalues is referred as *splitting point*. Splitting point breaks eigenvectors into two sets. The set of eigenvectors with lowest eigenvalues constitute the low dimensional space.

Points from normal dataset slightly deviate from the origin of the embedded space. It means that the chosen minor components are perhaps the worst way to capture the nature of the normal sub-calls. However, sub-calls visited the sleeping cell reside comparatively far from the origin. Obviously, sub-calls can be described by just one coordinate, which is a polar radius. Thus, only the polar radius is taken into account for the further processing.

C. Data Mining

In our previous research [9] we made use of k-NN anomaly detection algorithm to label sub-calls as normal and outliers. k-means separates normal and abnormal sub-calls in a better way that is why it has been chosen for this study.

k-means is fed by testing dataset and aims at partition considered instances into k clusters. At first, k random points from testing dataset are taken and called as centroids. The considered space can be partitioned by the centroids into Voronoi cells. Then the coordinates of each centroid are recalculated as the average point of observed instances in centroid's Voronoi cell. The updated centroids split the considered space again. This iterative process continues until some stop condition is met.

In our implementation we initialize k-means by setting k equal to 2. Having run the algorithm, big cluster is considered as *normal* and sub-calls from other clusters are referred as *outliers*. Sometimes there are few outlier points returned by the algorithm. This small amount of outlier sub-calls is not enough to diagnose the current status of the whole network. That is why k-means is rerun with higher order of k, until enough number of sub-calls are labeled as outliers. Thus, the threshold should be introduced.

The minimum number of sub-calls needed to diagnose a cellular network is determined by the following reasoning. If there is a sleeping cell in a network and a malfunctioning occurred from the beginning of the observation, then number of abnormal sub-calls may be inferred from the average number of sub-calls per cell and the average number of cells visited by a sub-call. It is worth noting that in our simulation cells had equal popularities, but in other case the threshold should reflect possible unevenness.

D. Post-processing

Post-processing step infers the current state of a mobile cellular network based on the given outlier sub-calls. Sub-calls are localized on the map in accordance with their dominance cell or target cell features. We have developed three distinct post-processing methods that assign *outage scores* to each cell.

In our previous article [7] symmetry property of 2-grams was discovered. It says that the number of incoming and outgoing 2-grams of the same kind is roughly equal. However, this symmetry property is broken in the presence of the sleeping cell. Therefore, one type of anomaly score can be introduced as a sum of absolute differences between the number of incoming (I) and outgoing (O) 2-grams taken separately. *Target Cell O-I* and *Dominance Cell O-I* methods utilize the property of symmetry.

In this research we have discovered that the number of 2-grams that completely reside (C) in a cell depends on whether the cell is properly functioning or broken. The best effect of this property is exposed when localization is defined with help of dominance cell feature. Thus, *Dominance Cell C* method is introduced as a sum of 2-grams that completely reside in cells.

Post-processing methods assign higher outage scores to the actual sleeping cell and its neighbors in contrast with the rest cells. *Amplification* procedure increases the outage score of the sleeping cell at the expense of its neighbors, while scores of other cells remain barely touched. Practically, outage score of each cell is divided by the sum of scores of other cell, but not neighbors.

Normalization of calculated estimates enables us to compare post-processing methods. Each outage score is divided by the sum of scores. The received histogram is referred as *sleeping cell detection histogram*.

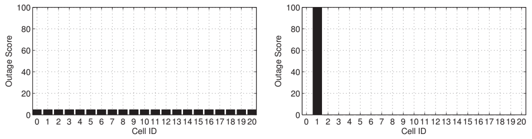
Combined method is a weighted sum of sleeping cell detection histograms derived by other methods. Weights are equal to the portion of sub-calls allocated to post-processing methods.

IV. Simulation Results

The main focus of this paper is on the post-processing algorithms. For that reason the other parts of the sleeping cell detection framework are not discussed in details here, but can be found in the following publications [7–10]. It is worth noting that the chosen feature space, composed of 2-grams, enables to distinguish explicitly normal and abnormal sub-calls. The utilization of k-means clustering technique is an important novelty of this study. k-means carries out the most accurate separation of sub-calls. The other new approach concerns the estimation and comparison of the utilized post-processing algorithms. These terms are referred as RMSE and percentage gain and are describe in the following sub-section.

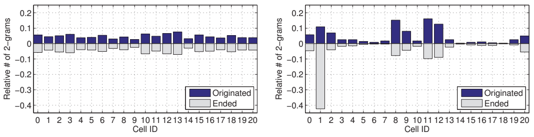
A. Comparison Metrics

The proposed post-processing methods are compared in terms of *root mean squared error* (RMSE) and *percent gain*. The following discussion clarifies the way of usage of these metrics. The outcome of the failure detection framework is outage scores assigned to each cell. There were no hot spots, such as markets or other public places, in the considered simulations. Hence, for the normal scenario an ideal post-processing algorithm would assign equal outage scores to each cell, see histogram given by Figure 3(a). Figure 3(b) depicts



(a) Normal scenario (b) Problem scenario

Fig. 3. Ideal sleeping cell detection histograms



(a) Outlier sub-calls from the normal scenario are considered (b) Outlier sub-calls from the problem scenario are considered

Fig. 4. Number of 2-grams "A2 RSRP LEAVE - A3 RSRP" originated and ended in each cell

outage scores produced by the ideal algorithm in the problem case. Here the sleeping cell gets the highest anomaly score and the rest cells are not suspected at all. The deviation of estimated outage scores of a proposed post-processing algorithm from the ideal one is determined by RMSE. The less RMSE, the closer a proposed algorithm to the ideal one.

The advantage of algorithm 1 over algorithm 2 is expressed by *percent gain*:

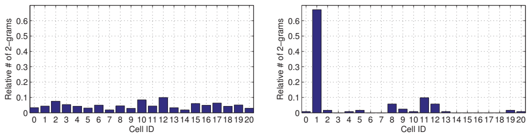
$$PercentGain = \frac{|RMSE_{alg1} - RMSE_{alg2}|}{RMSE_{alg2}} \cdot 100 \qquad (2)$$

These two metrics are an objective way to compare the efficacy of proposed algorithms.

B. Results Analysis

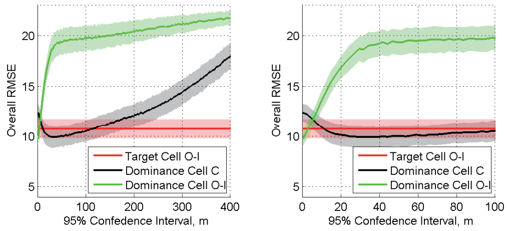
The diagnosis of network cells is drawn from the outlier sub-calls. Post-processing methods leverage 2-gram properties of outlier sub-calls and infer the outage scores. For instance, the number of times "A2 RSRP LEAVE - A3 RSRP" event pair originates and ends in a given cell is rather balanced for normal scenario, see Figure 4(a). However, the symmetry is broken in the presence of sleeping cell. Note that the same 2-gram ends in the malfunctioning cell, which is Cell 1, several times more often than originates, see Figure 4(b). The given example considers dominance cell feature of sub-calls. The same dependencies are hold for the target cell feature. Thus, the difference between the frequencies of incoming and outgoing 2-grams may be interpreted as outage score. Based on the discussed property we propose 'Target Cell O-I' and 'Dominance Cell O-I' methods.

Figure 5 shows the number of times "A2 RSRP LEAVE - A3 RSRP" event pair from outlier sub-calls completely resides in cells. The distribution of this 2-gram remains rather even on Figure 5(a), which corresponds to normally functioning network. In contrast, in problem scenario the histogram of the same distribution emphasizes abnormal behavior of the sleeping cell and its neighbors. This property motivated us to propose 'Dominance Cell C' method.



(a) Outlier sub-calls from the normal scenario are considered (b) Outlier sub-calls from the problem scenario are considered

Fig. 5. Number of 2-grams "A2 RSRP LEAVE - A3 RSRP" completely presented in each cell



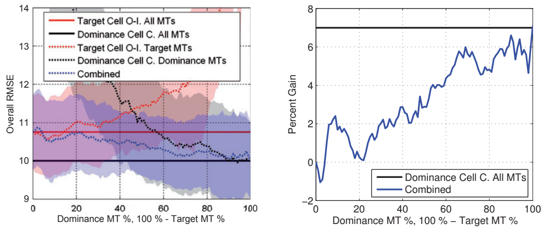
(a) 95% confidence interval in range 0:400 meters (b) 95% confidence interval in range 0:100 meters

Fig. 6. Overall RMSE of 'Target Cell O-I', 'Dominance Cell C' and 'Dominance Cell O-I' methods as a function of 95% confidence interval for location error. Colored areas show standard deviations of the lines

The presented figures are drawn from dominance cell feature of outlier sub-calls. The same properties are hold for target cell feature. Remember, that dominance cell feature exploits MT positioning information and significantly depends on its accuracy. Positioning error conforms to the model of Rayleigh noise, which is defined by σ. From the practical perspective, one should understand the location error as follows. Given sigma set to σ, 95% of measurements are within 2.45 σ radius. This value is referred as *95% confidence interval* in statistics. In our simulations we varied 95% confidence interval within a range from 0 up to 400 meters. It is worth noticing that the inter-site distance is equal to 500 meters.

Overall RMSE is the sum of RMSEs corresponding to normal and problem scenarios. Figure 6 depicts *Overall RMSEs* of considered post-processing methods as a function of Rayleigh sigma. Dominance cell methods overtake target cell symmetry method if relatively high positioning accuracy is achieved. Further increase of sigma leads to the degradation of dominance cell based approaches, while target cell one remains stable and becomes superior. It is worth noting, 'Dominance Cell O-I' fails to assign the highest outage score to the sleeping cell when sigma is greater than 10. Thus, 'Dominance Cell O-I' method remains reliable only if A-GPS or other equally accurate positioning technology is available.

'Target Cell O-I' and 'Dominance Cell C' methods have been chosen to generate combined estimate. Figure 7(a) shows Overall RMSEs of these post-processing methods. These curves depend on the number of sub-calls allocated for each of the methods. 'Dominance Cell C' outperforms 'Target Cell O-I' algorithm if accurate MDT measurements of all MTs are



(a) Overall RMSE of combined method. Colored areas show standard deviations of the lines

(b) Percentage gain of 'Dominance Cell C' and Combined methods with respect to 'Target Cell O-I' method

Fig. 7. Comparison metrics in the presence of location error. 95% confidence interval is equal to 50 meters

available. However, it might be the case that a mobile operator is able to figure out accurate enough coordinates only for the portion of MTs. Therefore, the combination of the results of dominance and target cell methods is the best option. RMSE of the combined estimate decreases as the number MTs with available dominance cell feature increases.

'Dominance Cell C' outperforms 'Target Cell O-I' method by 7% in terms of percent gain as shown in Figure 7(b). The more Combined method rely on the 'Dominance Cell C' algorithm the higher percent gain is achieved.

V. Conclusion

Our current research exploits MDT measurements and introduces machine learning framework for sleeping cell detection, caused by RACH problem. This article mainly considers the impact of precision of MT location on the post-processing algorithms' performances. In the real life MTs' locations are estimated with different accuracies. As long as there is no a post-processing algorithm that would exhibit the best performance at any level of location noise, the combined method was introduced. The combined post-processing method demonstrates higher reliability in a case of location errors of MTs. All the measurements were carried out by system-level LTE network simulator, which is utilized by the Nokia Networks research group. In our future work we are going to study how the amount of testing and training data influences on the performance of the detection framework.

Acknowledgment

The authors would like to thank Fedor Chernogorov from Magister Solutions Ltd. for providing the initial simulation data. Further, the work is made financially possible with the support of the Graduate School in Electronics, Telecommunications and Automation (GETA). Also authors thank the Foundation of Nokia Corporation for the support.

References

[1] 3GPP, "Telecommunication management; Self-Organizing Networks (SON); Concepts and requirements," TS 32.500, Sep. 2014.

[2] ——, "Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Self-configuring and self-optimizing network (SON) use cases and solutions," TS 36.902, Jul. 2011.

[3] Coluccia, Angelo and Ricciato, Fabio and Romirer-Maierhofer, Peter, "Bayesian Estimation of Network-wide Mean Failure Probability in 3G Cellular Networks," in *Proceedings of the 2010 IFIP WG 6.3/7.3 International Conference on Performance Evaluation of Computer and Communication Systems: Milestones and Future Challenges*, ser. PERFORM'10. Springer-Verlag, 2011, pp. 167–178.

[4] G. F. Ciocarlie, U. Lindqvist, S. Nováczki, and H. Sanneck, "Detecting Anomalies in Cellular Networks Using an Ensemble Method," in *Proceedings of the 9th International Conference on Network and Service Management, CNSM*, 2013, pp. 171–174.

[5] C. M. Mueller, M. Kaschub, C. Blankenhorn, and S. Wanke, "A Cell Outage Detection Algorithm Using Neighbor Cell List Reports," in *IWSOS*. Springer-Verlag, 2008, pp. 218–229.

[6] J. Turkka, F. Chernogorov, K. Brigatti, T. Ristaniemi, and J. Lempiinen, "An Approach for Network Outage Detection from Drive-Testing Databases," *Journal Comp. Netw. and Communic.*, 2012.

[7] F. Chernogorov, T. Ristaniemi, K. Brigatti, and S. Chernov, "N-gram Analysis for Sleeping Cell Detection in LTE Networks," in *ICASSP*. IEEE, 2013, pp. 4439–4443.

[8] F. Chernogorov, S. Chernov, K. Brigatti, and T. Ristaniemi, "Data Mining Approach to Detection of Random Access Sleeping Cell Failures in Cellular Mobile Networks," *arXiv:1501.03935 [cs.NI]*, 2015.

[9] S. Chernov, F. Chernogorov, D. Petrov, and T. Ristaniemi, "Data Mining Framework for Random Access Failure Detection in. LTE Networks," in *PIMRC*. IEEE, 2015.

[10] S. Chernov, M. Cochez, and T. Ristaniemi, "Anomaly Detection Algorithms for the Sleeping Cell Detection in LTE Networks," in *VTC*. IEEE, 2015.

[11] *Universal Terrestrial Radio Access (UTRA) and Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Masurement Collection for Minimization of Drive Tests (MDT); Overall Description; Stage 2*, 3GPP Std. TS 37.320, 2014.

[12] S. Sesia, M. Baker, and I. Toufik, *"LTE - The UMTS Long Term Evolution: From Theory to Practice"*. John Wiley & Sons, 2011.

[13] *Technical Specification Group Radio Access Network; Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol Specification (Release 10)*, 3GPP Std. TS 36.331, 2011.

[14] "Positioning with LTE," Ericsson White paper, Tech. Rep. 284 23-3155 Uen, Sep. 2011.

[15] W. J. H. T. Center, "Global Positioning System (GPS) Standard Positioning Service (SPS) Performance Analysis Report," Federal Aviation Administration, Tech. Rep., Jul. 2014.

[16] M. Timmerman, "Principal Component Analysis (2nd Ed.). I. T. Jolliffe."

[17] F. Cong, A. K. Nandi, Z. He, A. Cichocki, and T. Ristaniemi, "Fast and Effective Model Order Selection Method to Determine the Number of Sources in a Linear Transformation Model," in *Proceedings of EUSIPCO-2012*, aug 2012.

PV

SEQUENCE-BASED DETECTION OF SLEEPING CELL FAILURES IN MOBILE NETWORKS

by

Fedor Chernogorov, Sergey Chernov, Kimmo Brigatti, Tapani Ristaniemi 2015

Wireless Networks: The Journal of Mobile Communication, Computation and Information

Reproduced with kind permission of Springer US.

Wireless Netw
DOI 10.1007/s11276-015-1087-9

Sequence-based detection of sleeping cell failures in mobile networks

Fedor Chernogorov[1,2] · Sergey Chernov[2] · Kimmo Brigatti[2] · Tapani Ristaniemi[2]

© Springer Science+Business Media New York 2015

Abstract This article presents an automatic malfunction detection framework based on data mining approach to analysis of network event sequences. The considered environment is long term evolution (LTE) of Universal Mobile Telecommunications System with sleeping cell caused by random access channel failure. Sleeping cell problem means unavailability of network service without triggered alarm. The proposed detection framework uses N-gram analysis for identification of abnormal behavior in sequences of network events. These events are collected with minimization of drive tests functionality standardized in LTE. Further processing applies dimensionality reduction, anomaly detection with K-Nearest Neighbors, cross-validation, postprocessing techniques and efficiency evaluation. Different anomaly detection approaches proposed in this paper are compared against each other with both classic data mining metrics, such as F-score and receiver operating characteristic curves, and a newly proposed heuristic approach. Achieved results demonstrate that the suggested method can be used in modern performance monitoring systems for reliable, timely and automatic detection of random access channel sleeping cells.

Keywords Data mining · Sleeping cell problem · Anomaly detection · Performance monitoring · Self-healing · LTE networks

1 Introduction

Modern cellular mobile networks are becoming increasingly diverse and complex, due to coexistence of multiple Radio Access Technologies (RATs), and their corresponding releases. Additionally, small cells are actively deployed to complement the macro layer coverage, and this trend will only grow. In the future this situation is going to evolve towards even higher complexity, as in 5th Generation (5G) networks there will be much more end-user devices, served by different technologies, and connected to cells of different types [21, 33, 59, 62, 63]. New applications and user behavior patterns are daily coming into play. In such environment, network performance and robustness are becoming critical values for mobile operators. In order to achieve these goals, efficient flow of Quality and Performance Management (QPM) [36], which is a sequence of fault detection, diagnosis and healing, should be developed and applied in the network in addition to other optimization functions.

The concept of self-organizing network (SON) [57, 58] has been proposed to automate and optimize the most tedious manual tasks in mobile networks, including QPM. Automation is the key idea in SON and it has been proposed for self-configuration, self-optimization and self-healing in LTE and UMTS networks [27, 36, 68]. In traditional systems detection, diagnosis and recovery of

✉ Fedor Chernogorov
fedor.chernogorov@magister.fi; fedor.chernogorov@jyu.fi

Sergey Chernov
sergey.a.chernov@jyu.fi

Kimmo Brigatti
kimmobrigatti@gmail.com

Tapani Ristaniemi
tapani.e.ristaniemi@jyu.fi

[1] Magister Solution Ltd., Sepankatu 14 C, 40720 Jyväskylä, Finland

[2] Department of Mathematical Information Technology, University of Jyvaskyla, P.O. Box 35, 40014 Jyväskylä, Finland

Published online: 07 October 2015

network failures is mostly manual task, and it is heavily based on pre-defined thresholds, aggregation and averaging of large amounts of performance data—so called Key Performance Indicators (KPIs). Self-healing [32, 67] automates the functions of QPM process to improve reliability of network operation. Though, self-healing is still among the least studied functions of SON at the moment, and the developed solutions and use cases require improvement prior to application in the real networks. This is especially important for non-trivial network failures such as sleeping cell problem [13, 14, 36]. This is a special term used to denote a breakdown, which causes partial or complete degradation of network performance, and which is hard to detect with conventional QPM within reasonable time. Thus, in the research and standardization community automatic fault detection and diagnosis functions, enhanced with the most recent advancements in data analysis, are seen as the future of self-healing. Thus, development of improved self-healing functions for detection of sleeping cell problems, through application of anomaly detection techniques is of high importance nowadays. This article presents a novel framework based on N-gram analysis of MDT event sequences for detection of random access channel sleeping cells.

The rest of this paper is organized as follows. Section 2 describes common practices of quality and performance management in mobile networks, including MDT functionality, and advanced methods based on knowledge mining algorithms. Section 3 defines the concept of sleeping cell and its possible root cause failures. In Sect. 4 simulation environment, assumptions and random access channel problem are presented. Also Sect. 4 describes the generated and analyzed performance MDT data. Sect. 5 concentrates on the suggested sleeping cell detection knowledge mining framework. It includes overview of the applied anomaly detection methods: KNN anomaly outlier scores, N-gram, minor component analyses, postprocessing and data mining performance evaluation techniques. Section 6 is devoted to the actual research results. Data structures at different stages of analysis are shown, and efficiency of different postprocessing methods is compared. In Sect. 7 the concluding remarks regarding the findings of the presented research are given.

2 Quality and performance management in cellular mobile networks

Performance management in wireless networks includes three main components: data collection, analysis and results interpretation. Data gathering can be done either by aggregation of cell-level statistics—collection of KPIs, or collection of detailed performance data with drive tests. The main weaknesses in analysis of KPIs are that a lot of statistics is left out at the aggregation stage, due to averaging over time, element and because fixed threshold values are applied. Even though drive test campaigns provide far more elaborate information regarding network performance, they are expensive to carry out and do not cover overall area of network operation. Root cause analysis is done manually in majority of cases, and because of that there is a room for more intelligent approaches to detection and diagnosis of network failures, e.g. with data mining and anomaly detection techniques. This would provide possibility to automate performance monitoring task furthermore.

2.1 Minimization of drive tests

Yet another way to improve network QPM is to collect a detailed performance database. This is enabled with MDT functionality standardized in 3rd Generation Partnership Project (3GPP) [28]. MDT is designed for automatic collection and reporting of user measurements, where possible complemented with location information. Collected data is then reported to the serving cell, which in turn sends it to MDT server [39]. Thus, large amount of network and user performance is available for analysis. This is where the power of data mining and anomaly detection can be applied.

Specification describes several use cases for MDT: improvement of network coverage, capacity, mobility robustness and end user quality of service [36]. According to the standard, MDT measurements and reporting can be done both in idle and connected Radio Resource Control (RRC) modes. In logged MDT, User Equipment (UE) stores measurements in memory, and reporting is done at the next transition from idle to connected state. In immediate MDT, measurements are reported as soon as they are done through existing connection. In turn, there are two measurement modes in immediate MDT: periodic and event-triggered [39]. Periodic measurements are very useful for initial network deployment coverage and capacity verification as they provide detailed map of network performance, say in terms of signal propagation or throughput. The main disadvantage of periodic measurements is that they consume a lot of network and user resources. In contrast, event-triggered approach provides less information regarding the network status, but can be very efficient for mobility robustness and resource savings. In our study, immediate event-triggered MDT is used for collection of performance database. Table 1 presents the list of network events which triggered MDT measurements and reporting.

2.1.1 Location estimation in MDT

One of the important features of MDT is collection of geo-location information at the measurement time

Table 1 Network events triggering MDT measurements and reporting

PL PROBLEM—Physical Layer Problem [30].

RLF—Radio Link Failure [69].

RLF REESTAB.—Connection reestablishment after RLF.

A2 RSRP ENTER—RSRP goes under A2 enter threshold.

A2 RSRP LEAVE—RSRP goes over A2 leave threshold.

A2 RSRQ ENTER—RSRQ goes over A2 enter threshold.

A3 RSRP—A3 event, according to 3GPP specification.

HO COMMAND—handover command received [69].

HO COMPLETE—handover complete received [69].

moments. Whenever UE location is provided in MDT report there are several ways to associated it with particular cell, such as: serving cell ID, dominance maps and a new approach based on target cell ID information.

Serving cell ID is available with MDT event-triggered report, even for early releases of LTE. However, in case of coverage hole or problems with new connection establishment, this approach can lead to mistakes in UE location association, because the faulty cell would never become serving in the worst case scenario. This limits the usage of serving cell method for sleeping cell detection. To overcome the problem presented above, a dominance maps method can be used. This is a map, which demonstrates the E-UTRAN Node B (eNB)[1] with the strongest radio signal in each point of the network, see Fig. 1. The main advantage of dominance maps is that mapping of cell ID to location coordinate of UE MDT measurement is very precise, and this results in higher accuracy of sleeping cell detection. The downside of dominance maps approach is that it requires a lot of detailed input measurement information. Though, MDT functionality is one of the efficient ways to create such maps [25].

The last method for cell ID and UE report location association uses target cell ID feature. This information is available in the network events A3 RSRP, HO COMMAND, HO COMPLETE and RLF REESTABLISHMENT. The strong side of this method is that detection of sleeping cell becomes possible with a very limited amount of information, as it is shown in Sect. 6.

The key aspects which should be taken into account when selecting a location association method are accuracy and amount of information to create mapping between cell and user location.

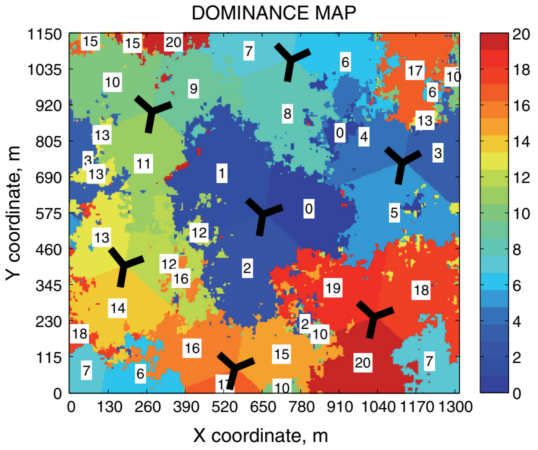


Fig. 1 Wrap around Macro 21 slow faded dominance map

2.2 Advanced data analysis approaches in QPM

Studies in advanced data analysis for QPM can be divided to several groups. In certain studies, the data reported by the users is used for the analysis. For instance, in [55] authors suggest a method for detection of sleeping cells, caused by transmitted signal strength problem, on the basis of neighbor cell list information. Application of non-trivial preprocessing and different classification algorithms allowed to achieve relatively good accuracy in detection of cell hardware faults. However, the proposed anomaly detection system is prone to have relatively high false rate. In [71] a method based on analysis of TRACE-based user data with diffusion maps is presented. More extensive application of diffusion maps for network performance monitoring can also be found in [49].

Even though, user level statistics is more detailed, still majority of studies devoted to improvement of QPM rely on cell-level data. The first proposals of sleeping cell detection automation using statistical methods of network monitoring are presented in [13, 14]. Preparation of normal cell load profile and evaluation of the deviation in observed cell behavior is suggested as a way for identification of problematic cells. The idea of statistical approach has been further studied in [61, 70], where a profile-based system for performance monitoring is proposed. The strong side of this study is that real data from 3rd Generation (3G) network has been analyzed. Moreover a complete system for fault detection and diagnosis is developed. However, the disadvantage of the proposed method is substantial time needed for training the algorithm (about 4 days of observation), and the necessity to manually input diagnosis options. Though in [60], the latter drawback is overcome using Kolmogorov–Smirnov two-sample test [48] for

[1] Evolved Universal Terrestrial Radio Access Network (E-UTRAN)

automatic creation of diagnosis profile database. Bayesian networks have also been applied for diagnosis and root cause probability estimation, given certain KPIs [3–5, 51]. The complications here are preparation of correct probability model and appropriate KPI threshold parameters. More advanced data mining methods are applied to analysis of cell-level performance statistics, and novel method of using an ensemble of classification algorithms is proposed [17, 20]. The idea is to use multiple algorithms for fault detection. At the training phase classification is done with real, manually labeled data, and the best performing methods are prioritized with higher weight. One of the core drawbacks of this approach is that rather extensive set of data is needed to achieve reliable detection. Data collection has been done for 2 months of network operation and 12 KPIs are observed. Labeling of the collected dataset is also a tedious manual task. In [18] application of classification and clustering methods for detection and diagnosis of strangely behaving network regions is presented. For this study a huge data collection campaign has been done: 4000 cells have been monitored for over 2.5 months, and 11 KPIs gathered. Authors manage to create a complete detection and diagnosis system. The largest achievement of this study is that no training or error free data is needed to find the anomalous/problematic cells. However, the critical question is the applicability of the presented method with a smaller input data set, both in terms of geographical scope and time scale. The continuation study [19] makes initial attempt to address changes in the network behavior, through adjustment of data mining model parameters on the fly. Some studies also consider neural network algorithms for detection of malfunctions [53, 65].

Among the drawbacks of the reviewed studies on advanced performance monitoring is that collection of appropriate statistical base takes substantial amount of time (from days to months). This increases reaction time in case of outages and does not completely solve the problems of operators in optimization of their QPM. For some of the proposed methods a very large geographical scope of data collection is also required.

In order to overcome weaknesses of the traditional QPM systems and advanced approaches described above we propose a sequence-based analysis method. The scope of our study is concentrated at the analysis of the user-level data, collected with immediate MDT functionality [40, 46]. Efficient detection and localization of the faulty cell is achieved through application of the knowledge mining framework based on N-gram analysis. Data collection for training and testing phases of the framework can be done within minutes of network operation. This becomes possible by configuring and running a compact management-based MDT campaign. Overall detection execution, together with initial learning stage, is going to take in the order of tens of minutes, but not days or even weeks. Subsequent detection, where training is not involved must be even faster.

In the early works cell outage detection caused by signal strength problems (antenna gain failure) is studied [10, 11, 72]. This area matches the 3GPP use case called "cell outage detection" [32]. Identification of the cell, in malfunction condition is done by means of analysis of numerical properties of multidimensional dataset. Each data point represents either periodic or event-triggered user measurement. Such methods as diffusion maps dimensionality reduction algorithm, k-means clustering and k-nearest neighbor classification methods are applied.

To increase robustness of the proposed solutions in MDT data analysis and make the developed detection system suitable for application in real networks, a more sophisticated experimental setup is considered. Sleeping cell caused by malfunction of random access channel, discussed in Sect. 3, does not produce coverage holes from perspective of radio signal, but still makes service unavailable to the subscribers. This problem, which is an instance of physical channel malfunction, is considered to be one of the most complex for mobile network operators, as detection of such failures may take days or even weeks, and negatively affects user experience [36]. To make fault detection framework more flexible and independent from user behavior, such as variable mobility and traffic variation, analysis of numerical characteristics of MDT data is substituted with processing of *network event sequences* with N-gram method. Network events can include different mobility or signaling related nature, such as A2, A3 or handover complete message [44]. Initial results in this area are presented in [12].

3 Sleeping cell problem

Sleeping cell is a special kind of cell service degradation, which leads to network performance decrease, invisible for the operator, but affecting user Quality of Experience (QoE). On one hand, detection of sleeping cell problem with traditional monitoring systems is complicated, as in many cases KPI thresholds do not indicate the failure. On the other hand fault identification can be very sluggish, as creation of cell behavior profile requires long time, as it is discussed in the previous section. Regular, less sophisticated types of failures usually produce cell level alarms to performance monitoring system of mobile network operator. In contrast, for sleeping cells degradation occurs seamlessly and no direct notification is given to the service provider.

In general, any cell can be called degraded in case if it is not 100% functional, what negatively affects user experience. Classification of sleeping cells, depending on the extent of performance degradation from the lightest, to the most severe [13, 15]: *impaired* or *deteriorated*—smallest negative impact on the provided service, *crippled*—characterized by a severely decreased capacity, and *catatonic*—kind of outage which leads to complete absence of service in the faulty area, such cell does not carry any traffic.

Degradation can be caused by malfunction of different hardware or software components of the network. Depending on the failure type, different extent of performance degradation can be induced. In this study the considered sleeping cell problem is caused by Random Access Channel (RACH) failure. This kind of problem can appear due to RACH misconfiguration, excessive load or software/firmware problem at the eNB side [1, 73]. RACH malfunction leads to inability of the affected cell to serve any new users, while earlier connected UEs still get served. This problem can be classified to crippled sleeping cell type, and with time the affected cell tends to become catatonic. In many cases RACH problem becomes visible for the operator only after a long observation time or even due to user complains. For this reason, it is very important to timely detect such cells and apply recovery actions.

3.1 Random access sleeping cell

Malfunction of RACH can lead to severe problems in network operation as it is used for connection establishment in the beginning of a call, during handover to another cell, connection re-establishment after handover failure or Radio Link Failure (RLF) [69]. Malfunction of random access in cell with ID 1, is caused by erroneous behavior of T304 timer [30], which expires before random access procedure is finished. Modeling of this failure is done so that at certain moment of network operation cell 1 loses capability to successfully go through random access procedure. Thus, whenever UE tries to initiate random access to this cell, this attempt fails. Malfunction area covers around 5 % of the overall network (1 out of total 21 cells).

4 Experimental setup

4.1 Simulation environment

Experimental environment is dynamic system level simulator of LTE network, designed according to 3GPP Releases 8, 9, 10 and partly 11. Throughput, spectral efficiency and mobility-related behavior of this simulator are validated against results from other simulators of several companies in 3GPP [31, 50, 52]. Step resolution of the simulator is one Orthogonal Frequency-Division Multiplexing (OFDM) symbol. Methodology for mapping link level Signal to Interference plus Noise Ratio (SINR) to the system level is presented in [7]. Simulation scenario is an improved 3GPP macro case 1 [29] with wrap-around layout, 21 cells (7 base stations with 3-sector antennas), and inter-site distance of 500 meters. Modeling of propagation and radio link conditions includes slow and fast fading. Users are spread randomly around the network, so that on average there are 15 dynamically moving UEs per cell. The main configuration parameters of the simulated network are shown in Table 2.

4.2 Generated performance data

Generated performance data includes dominance map information and MDT log, which contains the following fields:

- MDT triggering event ID. The list of possible events is presented in Table 1. This is a categorical (nominal) and sequential data, i.e. sequences of events are meaningful from data mining perspective;
- UE ID. This is also categorical data;
- UE location coordinates [m]. It is numerical, spatial data;
- Serving and target cell ID – spatial, categorical data.

It is important to know the type of the analyzed data to construct efficient knowledge mining framework [9, 37].

Simulations done for this study cover three types of network behavior: “normal” – network operation *without* random access sleeping cell; “problematic” – network *with* RACH failure in cell 1; “reference” – *no sleeping cell*, but different slow and fast fading maps, i.e. if compared to “normal” case, propagation-wise it is a different network. The latter case is used for validation purposes. All three of these cases have different mobility random seeds, i.e. call start locations and UE traveling paths are not the same. Each of these 3 cases is represented with 6 data chunks. The training and testing phases of sleeping cell detection are done with pairs of MDT logs by means of K-fold approach [37]. For example, “normal”-“problematic”, or “normal”-“reference” cases are considered. Thus, in total there are 72 unique combinations of analyzed MDT log pairs, which is rather statistically reliable data base.

5 Sleeping cell detection framework

The core of the presented study is sleeping cell detection framework based on knowledge mining, Fig. 2. Both training and testing phases are done in accordance to the process of Knowledge Discovery in Databases (KDD), which includes the following steps [24, 37]: data cleaning, integration from different sources, feature selection and

Table 2 General simulation configuration parameters

Parameter	Value	Parameter	Value
Cellular layout	Macro 21 Wrap-around	Number of cells	21
UEs per cell	17	Inter-Site Distance	500 m
Link direction	Downlink	RRC IDLE mode	Disabled
User distribution in the network	Uniform	Maximum BS TX power	46 dBm
Initial cell selection criterion	Strongest RSRP value	Handover margin (A3 margin)	3 dB
Handover time to trigger	256 ms	Hybrid Adaptive Repeat and reQuest (HARQ)	Enabled
Slow fading standard deviation	8 dB	Slow fading resolution	5 m
Simulation length	572 s (9.5 min)	Simulation resolution	1 time step = 71.43 μs
Network synchronicity mode	Asynchronous	Max number of UEs/cell	20
UE velocity	30 km/h	Call duration	90 s
Traffic model	Constant Bit Rate 320 kbps	Normal and Reference cases	Simulation without sleeping cell
Problematic case	Simulation with RACH problem in cell 1		
A2 RSRP Threshold	-110	A2 RSRP Hysteresis	3
A2 RSRQ Threshold	-10	A2 RSRQ Hysteresis	2

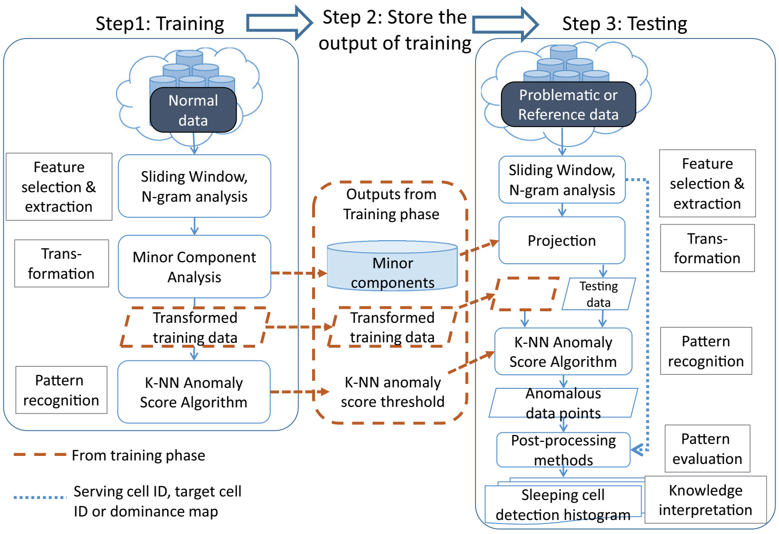

Step1: Training
Step 2: Store the output of training
Step 3: Testing
Normal data
Feature selection & extraction
Sliding Window, N-gram analysis
Trans-formation
Minor Component Analysis
Transformed training data
Pattern recognition
K-NN Anomaly Score Algorithm
Outputs from Training phase
Minor components
Transformed training data
K-NN anomaly score threshold
Problematic or Reference data
Sliding Window, N-gram analysis
Feature selection & extraction
Projection
Trans-formation
Testing data
K-NN Anomaly Score Algorithm
Pattern recognition
Anomalous data points
Post-processing methods
Pattern evaluation
Sleeping cell detection histogram
Knowledge interpretation
From training phase
Serving cell ID, target cell ID or dominance map

Fig. 2 Sleeping cell detection framework

extraction, transformation, pattern recognition, pattern evaluation and knowledge presentation. The constructed data analysis framework for sleeping cell detection is semi-supervised, because unlabeled error-free data is used for training of the data mining algorithms. The analysis can be logically separated to two parts: identification of the anomalous data points in MDT data and localization of these points in the real network and assignment of the real sleeping cell score to each cell (can be treated as extent of cell performance abnormality). The first problem is solved with preprocessing and pattern recognition, while the latter is more a task of pattern evaluation and postprocessing. In testing phase problematic data is analyzed to detect abnormal behavior. Reference data is used for testing in order to verify how much the designed framework is prone to make false alarms.

5.1 Feature selection and extraction

Feature selection and extraction is the first step of sleeping cell detection. At this stage, input data is prepared for further analysis. Preprocessing is needed as reported UEs MDT event sequences have variable lengths, depending on the user call duration, velocity, traffic distribution and network layout.

5.1.1 Sliding window preprocessing

Sliding window approach [64] allows to divide calls to *sub-calls* of constant length, and by that to unify input data. There are two parameters in sliding window algorithm: window size m and step n. After transformation, one sequence of N events (a call) is represented by several overlapping (in case if $n < m$) sequences of equal sizes, except for the last sub-call, which is the remainder from N modulo n.

In the presented results overlapping sliding window size is 15, and the step is 10 events. Such setup allows to maintain the context of the data after processing [49]. The number of calls and sub-calls for all three data sets are shown in Table 3.

Table 3 Number of calls and sub-calls in analyzed data

Amount / Dataset	Normal	Problem	Reference
Calls (all)	2530	1940	2540
Sub-calls (all)	7230	7134	7201
Normal sub-calls	6869	5932	6821
Abnormal sub-calls	361	1202	380

5.1.2 N-gram analysis

When input user-specific MDT log entries are standardized with sliding window method, the data is transformed from sequential to numeric format. It is done with N-gram analysis method, widely used e.g. for natural language processing and text analysis applications such as speech recognition, parsing, spelling, etc. [6, 8, 35, 45, 56]. In addition, N-gram is applied for whole-genome protein sequences [26] and for computer virus detection [16, 23].

N-gram is a sub-sequence of N overlapping items or units from a given original sequence. The items can be characters, letters, words or anything else. The idea of the method is to count how many times each sub-sequence occurs. This is the transformation from sequential to numerical space.

Here is an example of N-gram analysis application for two words: ‘performance’ and ‘performer’, $N = 2$, and a single unit is a character. The resulting frequency matrix after N-gram processing is shown in Table 4. In case of sequence analysis of MDT data, a letter from this example corresponds to an MDT event given in Table 1. Thus, for 2-gram analysis pairs of network events are considered, such “PL PROBLEM - RADIO LINK FAILURE”, or “A3 RSRP—HO COMMAND”.

Table 4 Example of N-gram analysis per character, $N = 2$.

Analyzed word	pe	er	rf	fo	or	rm	ma	me	an	nc	ce
performance	1	1	1	1	1	1	1	0	1	1	1
performer	1	2	1	1	1	1	0	1	0	0	0

5.2 Dimensionality reduction with minor component analysis

Dimensionality reduction is applied to convert high-dimensional data to a smaller set of derived variables. In the presented study Minor Component Analysis (MCA) method is applied [54]. This algorithm has been selected on the basis of comparison with other dimensionality reduction methods such as Principal Component Analysis (PCA) [47] and diffusion maps [22]. MCA extracts components of covariance matrix of the input data set and uses minor components (eigenvectors with the smallest eigenvalues of covariance matrix). 6 minor components are used as a basis of the embedded space. This number is defined by means of Second ORder sTatistic of the Eigenvalues (SORTE) method [42, 43].

5.3 Pattern recognition: K-NN anomaly score outlier detection

In order to extract abnormal instances from the testing dataset K-NN anomaly outlier score algorithm is applied. In contrast with K-NN classification, method is not supervised, but semi-supervised, as the training data does not contain any abnormal labels. In general, there are two approaches concerning the implementation of this algorithm; anomaly score assigned to each point is either the sum of distances to k nearest neighbors [2] or distance to k-th neighbor [66]. The first method is employed in the presented sleeping cell detection framework, as it is more statistically robust. Thus, the algorithm assigns an anomaly score to every sample in the analyzed data based on the sum of distances to k nearest neighbors in the embedded space. Euclidean metric is applied as similarity measure. Points with the largest anomaly scores are called outliers. Separation to normal and abnormal classes is defined by threshold parameter T, equal to 95th percentile of anomaly scores in the training data.

Configuration parameters of data analysis algorithms in the presented sleeping cell detection framework are summarized in Table 5.

5.4 Pattern evaluation

The main goal of pattern evaluation is conversion of output information from K-NN anomaly score algorithm to knowledge about location of the network malfunction, i.e. RACH sleeping cell. This is achieved with postprocessing of the anomalous data samples through analysis of their correspondence to particular network elements, such as UEs and cells. For this purpose we developed 4 post-processing methods: Dominance Cell Sub-Call Deviation, Dominance Cell 2-Gram Deviation, Dominance Cell 2-Gram Symmetry Deviation, and Target Cell Sub-Calls. The essence of these methods, discussed throughout this section, is reflected in their names. The first part describes which geo-location information is used for mapping data samples to cells, e.g. dominance map information, target or serving cell ID. The second part denotes what is used as feature space for postprocessing. It can be either "sub-calls", when rows of the dataset are used as features or "2-gram", when individual event pair combinations, i.e. columns of the dataset are used as features. The last, third part of the method name describes analysis considers the difference between training and testing data ("deviation" keyword), or whether only information about testing set is used to build sleeping cell detection histogram.

Output from the postprocessing methods described above is a set of values—sleeping cell scores, which correspond to each cell in the analyzed network. High value of this score means higher abnormality, and hence probability of failure. To achieve clearer indication of problematic cell presence, additional non-linear transformation is applied. It is called amplification, as it allows to emphasize problematic areas in the sleeping cell histogram. Sleeping cell score of each cell is divided by the sum of Sleeping Cell (SC) scores of all non-neighboring cells. Sleeping cell scores, received after postprocessing and amplification are then normalized by the cumulative SC score of all cells in the network. Normalization is necessary to get rid of dependency on the size of the dataset, i.e. number of calls and users.

Table 5 Parameters of algorithms in sleeping cell detection framework

Parameter	Value
Number of chunks in K-fold method per dataset	6
Sliding window size	15
Sliding window step	10
N in N-gram algorithm	2
Number of nearest neighbors (k) in K-NN algorithm	35
Number of minor components	6

5.5 Knowledge interpretation and presentation

The final step of the data analysis framework is visualization of the fault detection results. It is done with construction of a sleeping cell detection histogram and network heat map. However, sleeping cell histogram does not show how cells are related to each other: are they neighbors or not, and which area of the network is causing problems. Heat map method shows more anomalous network regions with darker and larger spots, while normally operating regions are in light grey color. The main benefit of network heat map is that mobile network topology and neighbor relations between cells are illustrated.

5.5.1 Performance evaluation

To apply data mining performance evaluation metrics labels of data points must be known. Cell is labeled as abnormal if its SC score deviates more than 3σ (standard deviation of sleeping cell scores) from the mean of SC score in the network. Mean value and standard deviation of the sleeping cell scores are calculated altogether from 72 runs produced by K-fold method for "normal"-"problematic", and "normal"-"reference" dataset pairs. Availability of the labels and the outcomes of different postprocessing methods enables application of such data mining performance metrics as accuracy, precision, recall, F-score, True Negative Rate (TNR) and False Positive Rate (FPR) [34]:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} = TP_{rate} \tag{3}$$

$$FP_{rate} = \frac{FP}{FP + TN} \tag{4}$$

$$Fscore = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

In these equations TP, TN, FP, FN denote elements of confusion matrix [37, 38, 41], and represent correspondingly the number of true positive, true negative, false

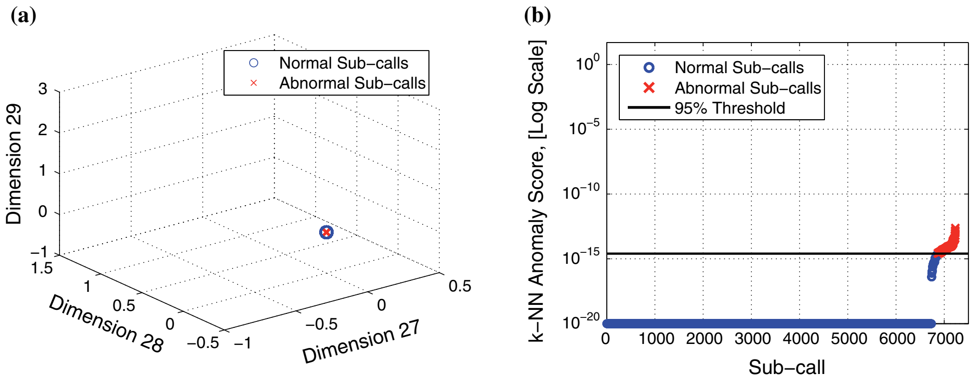


Fig. 3 Normal dataset used for training of the sleeping cell detection framework. **a** Normal training dataset in the embedded space. **b** Sorted outlier scores of normal training dataset

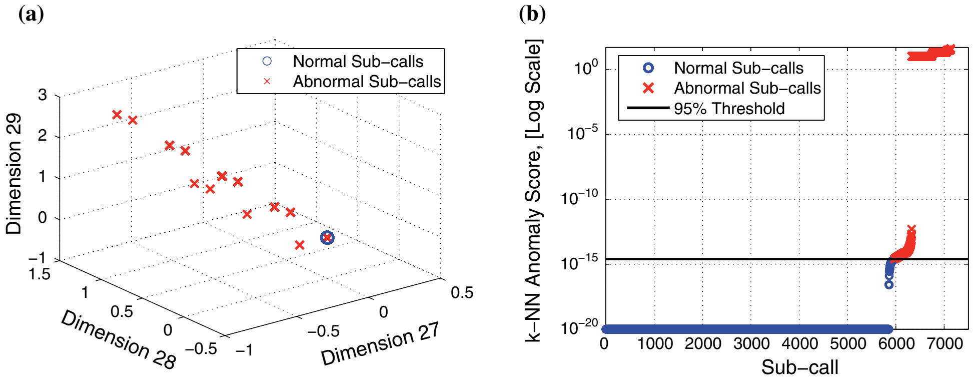


Fig. 4 Problematic dataset used at the testing phase of the sleeping cell detection framework. **a** Problem testing dataset in the embedded space. **b** Sorted outlier scores of problem testing dataset

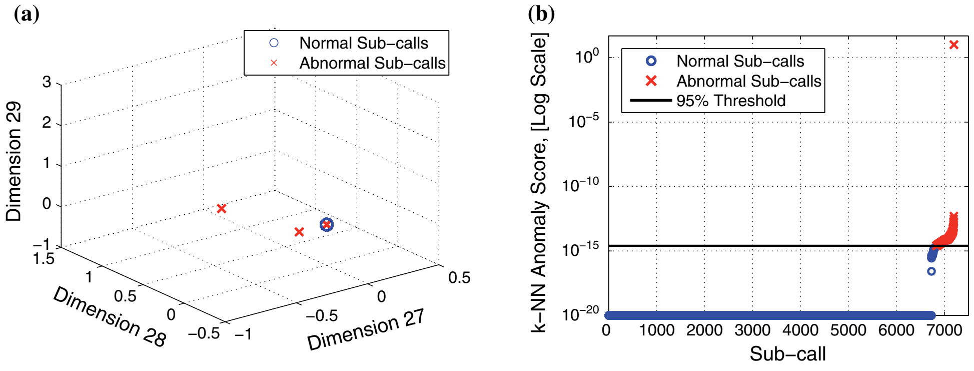


Fig. 5 Reference dataset used at the testing phase of the sleeping cell detection framework. **a** Reference testing dataset in the embedded space. **b** Sorted outlier scores of reference testing dataset

positive and false negative points. On the basis of these scores ROC curves are plotted.

In addition to the conventional performance evaluation metrics described above, a heuristic method is applied to complement the analysis. This approach measures how far is the achieved performance from the a priori known ideal solution. Performance of the sleeping cell detection algorithm can be described by a point in the space "cumulative standard deviation"-"sleeping cell magnitude". "Sleeping cell magnitude" is the highest SC score, which can reach value 100 due to normalization. "Cumulative standard deviation" coordinate equals to the standard deviation of SC scores of all other cells. This plane contains two points of interest: [0; 100] and $[0; 100/N_{\text{cellsinthenetwork}}]$. In case of malfunctioning network, the ideal sleeping cell detection algorithm assigns 100 value of SC score to the broken cell and zero values to the rest cells. Thus, the corresponding point [0; 100] is calculated. In case of error-free network, the ideal performance is mapped to the point $[0; 100/N_{\text{cellsinthenetwork}}]$, because all the cells have even SC scores equal to $100/N_{\text{cellsinthenetwork}}$. Thus, the smaller the Euclidean distance between the achieved and ideal sleeping cell histograms, the better the performance of the sleeping cell detection algorithm.

6 Results of sleeping cell detection

This section presents the results of sleeping cell detection for different post-processing algorithms. In addition, the data at different stages of the detection process is illustrated. Then performance metrics are used to compare effectiveness of the developed SC identification algorithms.

6.1 Preprocessing and K-NN anomaly score calculations

After preprocessing with sliding window and N-gram methods we get a so-called 2-gram popularity matrix. The size of this matrix equals to data chunk size and has 32 features—the number of non-zero 2-grams. This popularity matrix is transformed with MCA. The output of dimensionality reduction with MCA has 6 features—coordinates of points in 6-dimensional embedded space based on eigenvectors with the smallest eigenvalues. Then training MDT data is processed with K-NN anomaly score algorithm. As it is discussed in Sect. 5.3, the anomaly score threshold, used for separation of data points to normal and abnormal classes, is selected to be 95th percentile of outlier score in training data. Shape of normal training dataset in

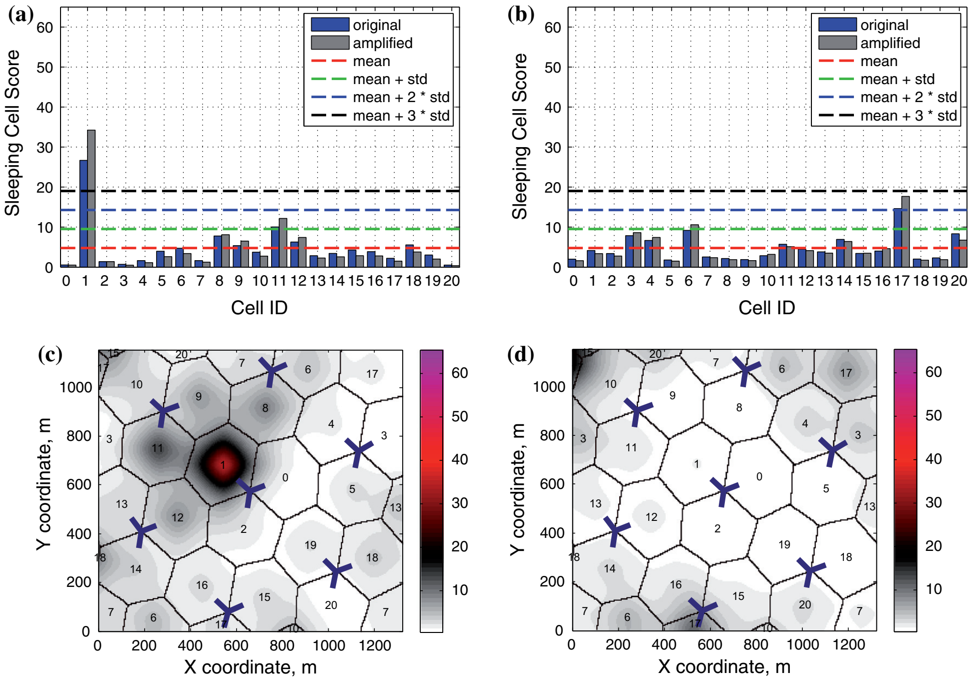


Fig. 6 Results of sleeping cell detection for dominance cell sub-call deviation method. **a** Problematic dataset sleeping cell detection histogram. **b** Reference dataset sleeping cell detection histogram, **c** Problematic dataset heat map, **d** Reference dataset heat map

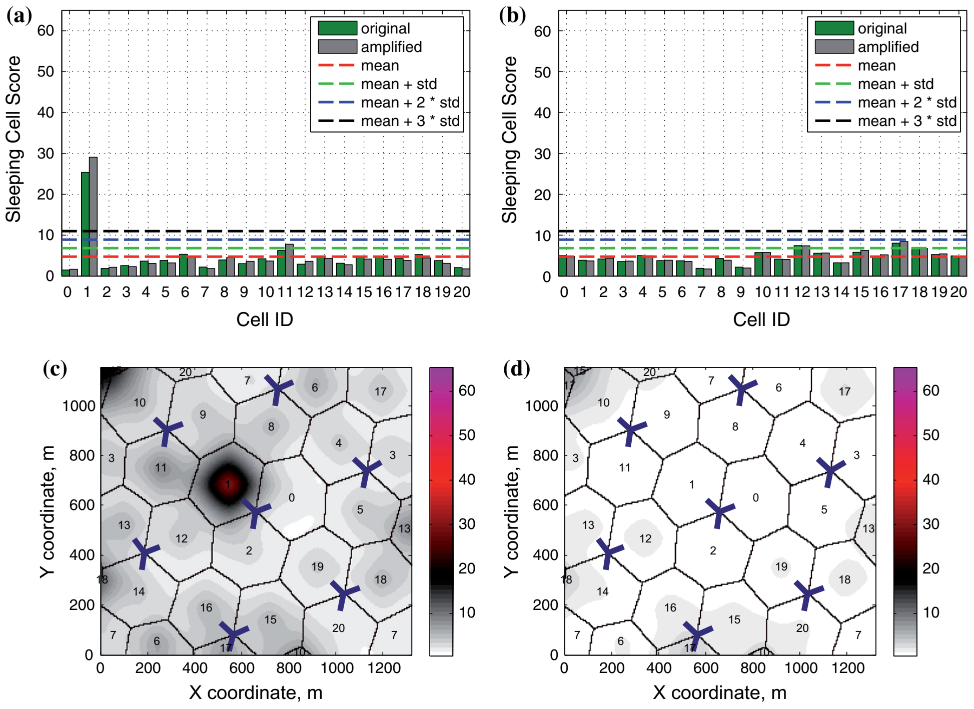


Fig. 7 Results of sleeping cell detection for dominance cell 2-gram deviation method **a** Problematic dataset sleeping cell detection histogram. **b** Reference dataset sleeping cell detection histogram. **c** Problematic dataset heat map. **d** Reference dataset heat map

the embedded space is shown in Fig. 3a. In this Fig. 3 dimensions are selected on the basis of visual inspection to demonstrate best the distribution of data points. Sorted anomaly outlier scores are presented in Fig. 3b. It can be seen that data points are very compact in the embedded space, and because of that there is no big difference in the anomaly score values. The main goals of analyzing testing dataset are to find anomalies, detect sleeping cell, and keep the false alarm rate as low as possible. At the testing phase either problematic or reference data are analyzed. After the same preprocessing stages as for training, the testing data is represented in the embedded space. When testing data is problematic dataset some of the samples are significantly further away from the main dense group of points, Fig. 4. These abnormal points are labeled as outliers, and the corresponding anomaly scores for these samples are much higher, as it can be seen from Fig. 4b. On the other hand, some of the points with relatively low anomaly score are above the abnormality threshold. This means that there is still certain percentage of false alarms, i.e. some "good" points are treated as "bad". The extent of negative effect caused by false alarms is discussed further in Sect. 6.4. Though, there is no opposite behavior referred to as "miss-detection"—none of the anomalous points are treated as normal.

Validation of the data mining framework is done by using error-free reference dataset as testing data. No real anomalies are present in the network behavior. Reference testing data in the embedded space and corresponding anomaly outlier scores are shown in Fig. 5. Only few points can be treated as outliers, and in general the shapes of normal (Fig. 3a) and reference (Fig. 5a) datasets in the embedded space are very similar. Anomaly outlier scores of the reference testing data are low for all points, except 2 outliers.

6.2 Application of postprocessing methods for sleeping cell detection

After training and testing phases certain sub-calls are marked as anomalies. The next step is conversion of this information to knowledge about location of malfunctioning cell or cells, and this is done through postprocessing described in Sect. 5.4.

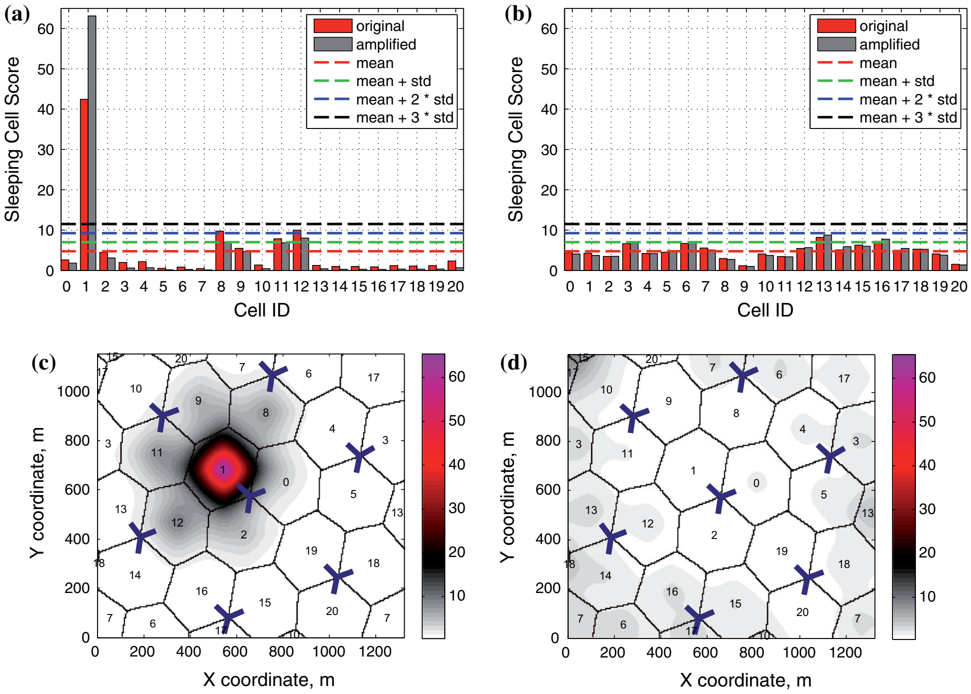


Fig. 8 Results of sleeping cell detection for dominance cell 2-gram symmetry deviation method. **a** Problematic dataset sleeping cell detection histogram. **b** Reference dataset sleeping cell detection histogram. **c** Problematic dataset heat map. **d** Reference dataset heat map

6.2.1 Detection based on dominance cell sub-call deviation

In our earlier study [12] postprocessing based on dominance cells and call deviation for sleeping cell detection is presented. One problem of using calls as samples is that, in case if the duration of the analyzed user call is long, the corresponding number of visited cells is large, especially for fast UEs. Hence, even if certain call is classified as abnormal, it is very hard to say which cell has anomalous behavior. To overcome this problem, analysis is done for sub-calls, derived with sliding window method, see Sect. 5.1.1. Sub-calls contain the same number of network events, and the length of the analyzed sequence is short enough to identify the exact cell, with problematic behavior. Deviation measures the difference between training and testing data, and it is used to sleeping cell detection histogram, presented in Fig. 6a. From this figure, it can be seen that abnormal sub-calls are encountered more frequently in the area of dominance of cell 1, which has the highest deviation. One can see that there are 2 types of bars—colorful (in this case blue) and grey. The second variant implies additional postprocessing step—amplification, described in Sect. 5.4. In addition to cell 1, its neighboring cells 8, 9, 11 and 12 also have increased deviation values, as it can be seen from the network heat map in Fig. 6c. Sleeping cell detection histogram and network heat map for reference dataset used as testing are shown in Fig. 6b, d correspondingly. Even though cells 6 and 17 have higher SC scores than other cells, they are not marked as abnormal, because their abnormality does not reach mean + 3σ level.

6.2.2 Detection based on dominance cell 2-gram deviation

In this method problematic network elements are found through the comparison of 2-gram frequencies in different areas of dominance map. For this purpose we consider all sub-calls from training data set against sub-calls assigned to abnormal class from testing dataset. In case there is a big increase or decrease, the cell associated with these changes is marked as abnormal. From sleeping cell detection histogram in Fig. 7a it can be that cell 1 has a clear difference in number of 2-gram occurrences in testing data, if compared to training data. This happens because handovers toward this cell fail. Due to this fact 2-gram sequence with events related to handovers become imbalanced in testing data if compared to training data. For instance, 2-grams

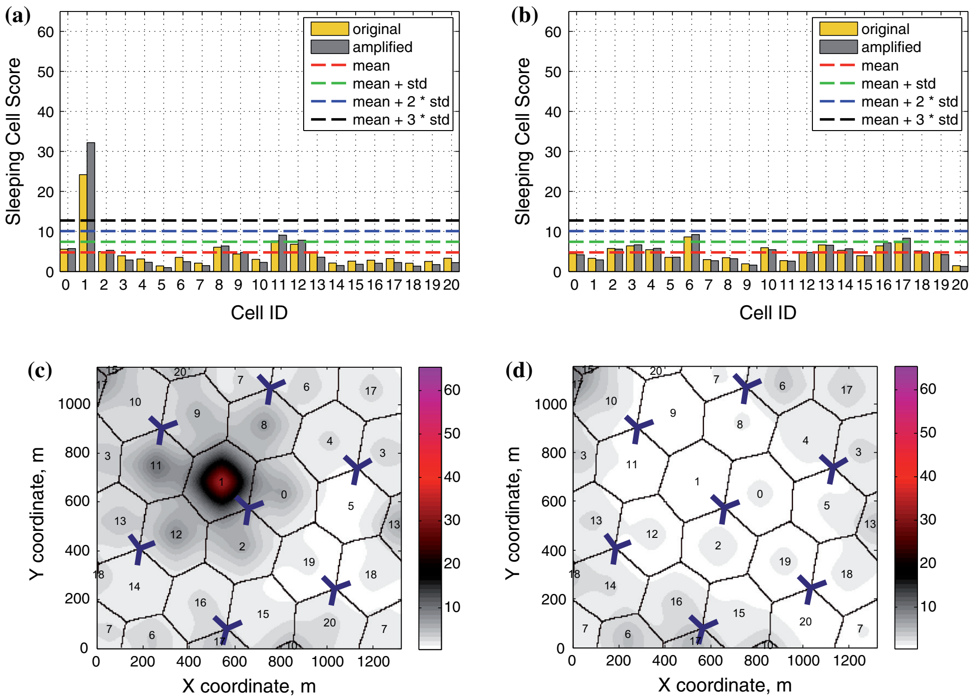


Fig. 9 Results of sleeping cell detection for target cell sub-calls method. **a** Problematic dataset sleeping cell detection histogram. **b** Reference dataset sleeping cell detection histogram. **c** Problematic dataset heat map. **d** Reference dataset heat map

like Handover (HO) Command—HO Complete and HO Complete—A2 RSRP ENTER, become very rare. On the other hand, 2-gram HO Command—A2 RSRP ENTER, which can be treated as indication of unsuccessful handovers, in opposite becomes very popular in testing data, while in training data it does not exist at all. Among the neighbors of problematic cell 1, only cell 11 has slightly increased sleeping cell score. Testing sleeping cell detection framework with reference data and postprocessing with Dominance Cell 2-Gram Deviation method demonstrates lower false-alarm rate than Dominance Cell Sub-Call Deviation, as it can be seen from Fig. 7b, d.

6.2.3 *Detection based on dominance cell 2-gram symmetry deviation*

This postprocessing method analyzes the symmetry imbalance of network event 2-grams. The symmetry imbalance is evaluated based on all sub-calls from training data set and sub-calls assigned to abnormal class from testing dataset. Information about the number of 2-grams directed to the cell, and from the cell is extracted from the training set. The considered 2-grams consist of events which sequentially occur in the dominance areas of 2 cells. It means that if in the training data, the number of handovers from Cell A to Cell B, and from Cell B to Cell A, is roughly the same, and in the testing set it is not, it can be concluded that symmetry of this particular 2-gram is skewed. Most common types of 2-grams which are analyzed with this method are related to handovers, e.g. A3—HO COMMAND sequences.

From Fig. 8 it can be seen that Dominance Cell 2-Gram Symmetry Deviation finds sleeping cell 1, while its neighboring cells 8, 9, 11 and 12 have suspiciously high sleeping cell score, if compared to other cells in the network.

Comparison of symmetry analysis method with two previously described postprocessing approaches shows that this method is very efficient in detecting sleeping cell and its neighbors. At the same time stability, i.e. false alarm rate, of this method is also very good, as it can be seen from Fig. 8b.

6.2.4 *Detection based on target cell sub-calls*

As it is discussed in Sect. 5.4, deviation between training and testing data is not calculated in this method. Extensive location information, like dominance map information, is

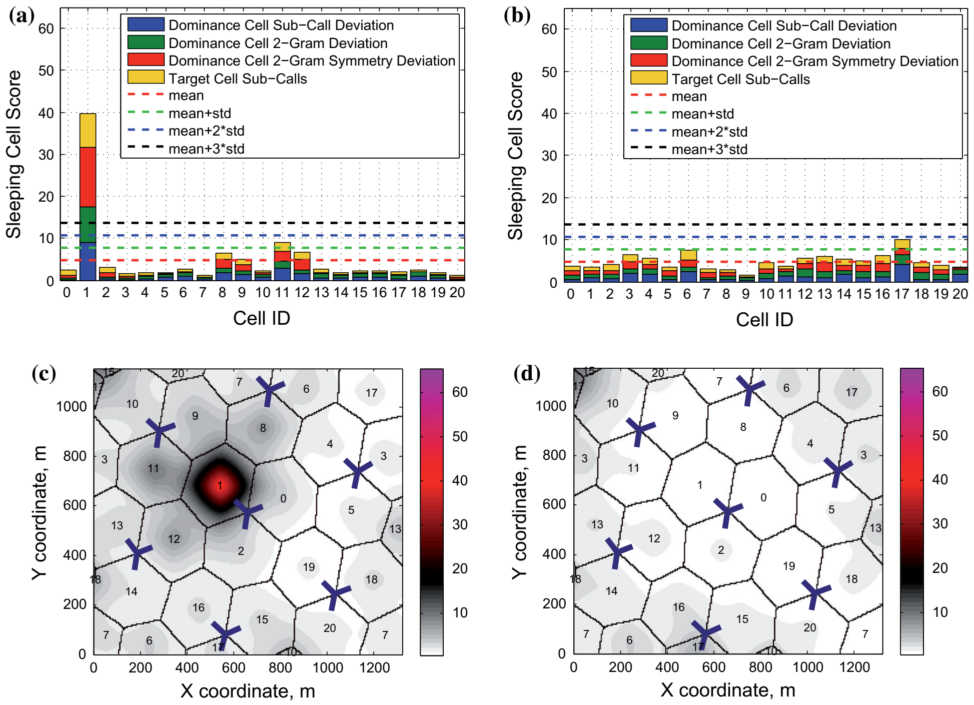


Fig. 10 Results of sleeping cell detection for amplified combined method. **a** Problematic dataset sleeping cell detection histogram. **b** Reference dataset sleeping cell detection histogram. **c** Problematic dataset heat map. **d** Reference dataset heat map

not required for sleeping cell detection with target cell sub-call method. The sleeping cell detection histogram, presented in Fig. 9, is constructed by counting all unique target cell IDs for each anomalous sub-call. It can be clearly seen that cell 1 is successfully detected. Neighboring cells 8, 9, 11 and 12 also contain indication of malfunction in this area, as it can be noticed from heat map, shown in Fig. 9b. For this method, the SC score of cell 1 is slightly lower than for the postprocessing methods, based on dominance cell deviation. On the other hand, target cell sub-call method is much simpler, and requires significantly less information about user event occurrence location.

6.3 Combined method of sleeping cell detection

The idea of this method is to create a cumulative sleeping cell detection histogram based on the results from all 4 postprocessing methods described above. The resulting amplified SC histogram is shown in Fig. 10. Cell 1 has sleeping cell score well over $\mu + 3 * \sigma$ threshold. Neighboring cells 8, 9, 11, 12 also have increased sleeping cell scores comparing to other cells, though they do not exceed the $\mu + 3 * \sigma$ threshold. Reference data used as testing also demonstrates stability of the combined approach – no false alarms are triggered. Though, it can be seen that usage of target cell sub-call method introduces some noise. It is important to note that postprocessing methods are applied with equal weights. However, it is possible to emphasize more accurate method by increasing its weight, and penalize the unreliable, by reducing its weight. Though, selection of optimal weights is a matter of a separate study and is not discussed in this article.

6.4 Comparison of algorithms and performance evaluation

The postprocessing methods discussed above have their own advantages and disadvantages. Traditional data mining metrics, discussed in Sect. 5.5.1, are applied for quantitative comparison of sleeping cell detection methods, Fig. 11a. Ideal performance is presented with the solid double black line, and corresponds to the maximum area of the hexagon. K-fold cross validation method is utilized to obtain statistically significant results. Figures 6, 7, 8, 9 and 10 show averaged values of the sleeping cell scores from all runs of K-fold separation. In some of the runs certain neighbors of cell 1 demonstrated sleeping cell scores

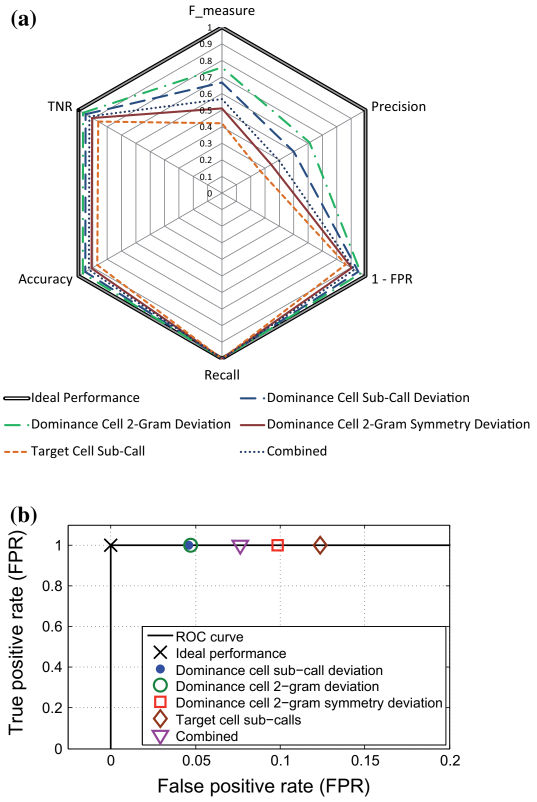


Fig. 11 Performance measures for comparison of sleeping cell detection algorithms. **a** Performance measures of algorithms. **b** ROC curve of sleeping cell detection framework

higher than 3σ threshold. This results in non-zero false positive rates. Formally, according to the values of the metrics, Dominance Cell 2-gram Deviation and Dominance Cell Sub-call Deviation methods, demonstrate better performance than other postprocessing techniques. However, high false positive rate for Dominance Cell 2-gram Symmetry Deviation and Target Cell Sub-call methods does not necessarily mean that these methods are worse. The reason is that neighbors of cell 1 exceed the 3σ threshold. This happens because adjacent cells are not completely independent, and are affected by malfunction in one of the neighbors. Thus, Dominance Cell 2-gram Symmetry Deviation and Target Cell Sub-call methods can be treated as more sensitive than the others. The ROC curves of the designed sleeping cell detection algorithms are presented in Fig. 11b. True positive rate equals 1 for all postprocessing methods. This is not always the case for many classification applications in real world systems. However, the proposed framework is able to create such projection of the input MDT data, that all anomalous points are correctly classified as abnormal. The latter can be seen from sorted anomaly scores of problematic data sets, shown in Fig. 4b. The negative side is that some methods mistakenly classify some normal points as abnormal, and this is reflected in false positive rate. Thus, the proposed framework is able to create such a projection of the MDT data, that in the new space normal data and anomalous data points are fully separable and do not overlap. Hence, the suggested data mining framework for sleeping cell detection is successful, and for reduction of false alarm rate it is necessary to invent a better separation rule, than 3σ deviation from mean SC score, see Figs. 6, 7, 8, 9 and 10.

Another method for comparison of postprocessing algorithms is a heuristic approach described in Sect. 5.5.1. According to this method, more accurate postprocessing algorithm is the one, which has the smallest distance to the ideal solution point for either problematic or error-free case. Cumulative distances for different algorithms in non-amplified and amplified cases are presented in Fig. 12a, b correspondingly. Also the coordinates of different postprocessing methods in heuristic performance measure plane are shown in Fig. 12c, d. It can be seen that Dominance Cell 2-Gram Symmetry Deviation method has the smallest distance from the ideal detection case. Thus, from perspective of the heuristic performance evaluation approach this method outperforms other postprocessing methods. Regarding the same performance metric we may conclude that amplified histograms show better results than non-amplified ones, which holds for all postprocessing methods.

7 Conclusions

This article presents a novel sleeping cell detection framework based on knowledge mining paradigm. MDT reports are used for the detection of a random access channel malfunction in one of the network cells. Experimental setup implements a simulated LTE network, used to generate a diverse statistics base with several thousands of user calls and tens of thousands of MDT samples. Investigated failure case is a sleeping cell caused by RACH malfunction. Even though the studied problem is rather specific, the proposed framework does not consider any properties or peculiarities of the random access failure for the detection. Moreover, analysis of event sequences makes the presented method applicable to data collected with MDT, TRACE functionality, mobile quality agents, and any other method, which is capable to gather the user specific sequences of network events. The studied type of sleeping cell problem is rather complex, and detection of this problem has never been done before. The applicability of our sequential analysis to other network failures might be beneficial, but it has to be studied.

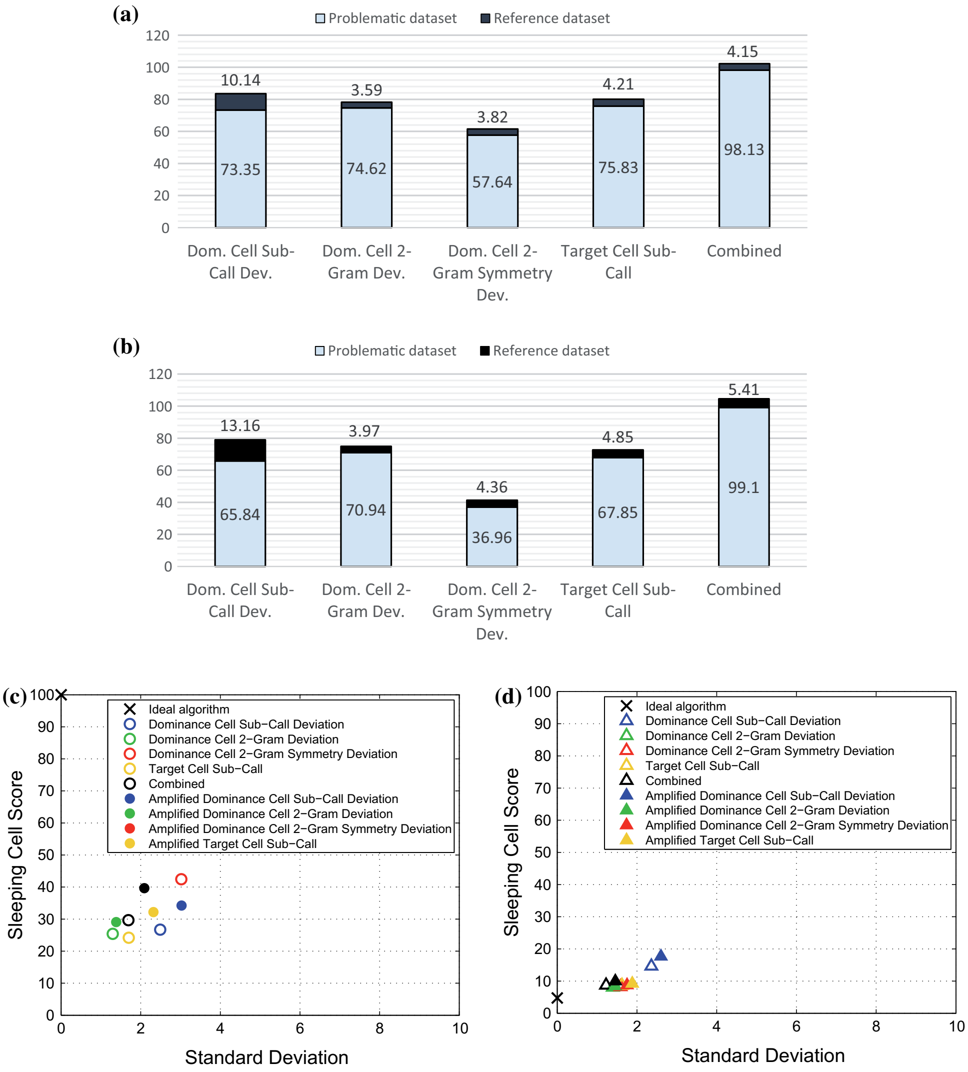


Fig. 12 Heuristic performance comparison of algorithms **a** Distances in original—non amplified approach. **b** Distances in amplified approach. **c** Heuristic performance distances for problematic case. **d** Heuristic performance distances for reference case.

The designed knowledge mining framework is semi-supervised. From the perspective of SONs the proposed system has centralized architecture, but it can also be hybrid, with preprocessing and transformation stages done in distributed manner. The heart of the developed detection framework is the analysis of sequences with N-gram method in the series of user event-triggered measurement MDT reports. Data preprocessing with sliding window transformation method allows to make the statistics base more reliable through standardization of the input event sequences. 2-gram analysis is used to convert sequential data to numeric format in the new feature space. To simplify analysis of the data in the new space, dimensionality reduction with minor component analysis method is applied. K-NN anomaly score detection algorithm is used to find the outliers in the data. Using this information, anomalous data points are converted with postprocessing to the knowledge about location of the problematic regions in the network. Comparison of different location mapping postprocessing methods is done. Additionally, so called amplification is used to take into account neighbor relations between cells and network topology, for improvement of sleeping cell detection performance. As it can be seen, amplified sleeping cell score of truly problematic cell is higher than corresponding non-amplified score.

Results demonstrate, that the developed suggested framework, based on sequence analysis, allows for efficient detection of the random access sleeping cell problem in the network. The projection of the data in the new space is such that accurate separation of normal and abnormal data points becomes possible. Evaluation shows that postprocessing method named Dominance Cell 2-Gram Symmetry Deviation demonstrates the best combination of results, with respect to heuristic performance measure. According to the same metric, the proposed amplification approach, improves the detection quality of postprocessing methods. However, this approach is an additional element of the developed nontrivial framework and is not the most important outcome of our research.

Results of this work lay grounds and suggest exact methods for building advanced performance monitoring systems in modern mobile networks. One of the possible directions in this area is extensive usage of data mining techniques in general, and anomaly detection in particular. New systems of network maintenance would allow to address growing complexity and heterogeneity of modern mobile networks, and will help to meet the requirements of 5G.

Future work in this field includes validation of the developed system in more complex scenarios, detection of several or different types of malfunctions, and substitution of semi-supervised approach with unsupervised. The ultimate goal is to achieve accurate and timely detection of different sleeping cell types in highly dynamic mobile network environments. Obviously, low level of false alarms must be supported, and at the same time significant increase of computational complexity should be avoided.

Acknowledgments Authors would like to thank colleagues from Magister Solutions, Nokia and University of Jyväskylä for collaboration, their valuable feedback regarding this research, and peer reviews. Work on this study has been partly funded by MIPCOM project, Graduate School in Electronics, Telecommunications and Automation (GETA), and Doctoral Program in Computing and Mathematical Sciences (COMAS).

References

1. Amirijoo, M., Frenger, P., Gunnarsson, F., Moe, J., & Zetterberg, K. (2009). On self-optimization of the random access procedure in 3g long term evolution. In Integrated network management-workshops, 2009. IM '09. IFIP/IEEE international symposium on, pp 177–184.
2. Angiulli, F., & Pizzuti, C. (2002). Fast outlier detection in high dimensional spaces. In Proceedings of the 6th European conference on principles of data mining and knowledge discovery, Springer-Verlag, London, UK, UK, PKDD '02, pp 15–26.
3. Barco, R., Lazaro, P., Diez, L., & Wille, V. (2008). Continuous versus discrete model in autodiagnosis systems for wireless networks. *Mobile Computing, IEEE Transactions on*, *7*(6), 673–681. doi:10.1109/TMC.2008.23.
4. Barco, R., Lazaro, P., Wille, V., Diez, L., & Patel, S. (2009). Knowledge acquisition for diagnosis model in wireless networks. *Expert Systems with Applications*, *36*(3, Part 1), 4745–4752.
5. Barco, R., Wille, V., Diez, L., & Toril, M. (2010). Learning of model parameters for fault diagnosis in wireless networks. *Wireless Networks*, *16*(1), 255–271. doi:10.1007/s11276-008-0128-z.
6. Brown, P. F., deSouza, P. V., Mercer, R. L., Pietra, V. J. D., & Lai, J. C. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, *18*, 467–479.
7. Brueninghaus, K., Astely, D., Salzer, T., Visuri, S., Alexiou, A., Karger, S., & Seraji, G. A. (2005). Link performance models for system level simulations of broadband radio access systems. In IEEE 16th international symposium on personal, indoor and mobile radio communications, 2005. PIMRC 2005., vol 4, pp 2306–2311 Vol. 4, doi:10.1109/PIMRC.2005.1651855.
8. Cavnar, W. B., & Trenkle, J. M. (1994). N-gram-based text categorization. In Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information Retrieval, pp 161–175.
9. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Survey*, *41*(3), 15:1–15:58.
10. Chernogorov, F. (2010). Detection of sleeping cells in long term evolution mobile networks. Master's thesis, University of Jyväskylä, Finland.
11. Chernogorov, F., Turkka, J., Ristaniemi, T., & Averbuch, A. (2011). Detection of sleeping cells in LTE networks using diffusion maps. In Vehicular technology conference (VTC Spring), 2011 IEEE 73rd, pp 1–5.
12. Chernogorov, F., Brigatti, K., Ristaniemi, T., & Chernov, S. (2013). N-gram analysis for sleeping cell detection in LTE networks. In Proceedings of the 38th international conference on acoustics, speech, and signal processing (ICASSP).
13. Cheung, B., Kumar, G. N., & Rao, S. A. (2005). Statistical algorithms in fault detection and prediction: Toward a healthier network. *Bell Labs Technical Journal*, *9*(4), 171–185.
14. Cheung, B., Fishkin, S. G., Kumar, G. N., & Rao, S.A. (2006a). Method of monitoring wireless network performance. Tech. rep., Los Angeles, CA, uS Patent 2006/0063521 A1, CN1753541A, EP1638253A1.
15. Cheung, B., Fishkin, S. G., Kumar G.N., & Rao, S. A. (2006b). Method of monitoring wireless network performance. US Patent 2006/0063521 A1, CN1753541A, EP1638253A1
16. Choi, J., Kim, H., Choi, C., & Kim, P. (2011). Efficient malicious code detection using n-gram analysis and svm. In L. Barolli, F. Xhafa, & M. Takizawa (Eds.), *NBiS* (pp. 618–621). : IEEE Computer Society.
17. Ciocarlie, G., Lindqvist, U., Novaczki, S., & Sanneck, H. (2013). Detecting anomalies in cellular networks using an ensemble method. In Network and service management (CNSM), 2013 9th international conference on, pp 171–174, doi:10.1109/CNSM.2013.6727831.
18. Ciocarlie, G., Cheng, C. C., Connolly, C., Lindqvist, U., Nitz, K., Novaczki, S., Sanneck, H., & Naseer-ul Islam, M. (2014a). Anomaly detection and diagnosis for automatic radio network verification. In 6th international conference on mobile networks and management, MONAMI 2014.
19. Ciocarlie, G., Cheng, C. C., Connolly, C., Lindqvist, U., Novaczki, S., Sanneck, H., & Naseer-ul Islam, M. (2014b). Managing scope changes for cellular network-level anomaly detection. In Wireless communications systems (ISWCS), 2014 11th international symposium on, pp 375–379, doi:10.1109/ISWCS.2014.6933381.

20. Ciocarlie, G., Lindqvist, U., Nitz, K., Novaczki, S., & Sanneck, H. (2014c). On the feasibility of deploying cell anomaly detection in operational cellular networks. In Network operations and management symposium (NOMS), 2014 IEEE, pp 1–6, doi:10.1109/NOMS.2014.6838305.
21. Cisco Systems (2014) Cisco visual networking index: Global mobile data traffic forecast update 20142019 white paper. https://gsmaintelligence.com/research/2014/12/understanding-5g/451/
22. Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, *21*(1), 5–30.
23. David, G. (2009). Anomaly detection and classification via diffusion processes in hyper-networks. PhD thesis, Tel-Aviv University, Tel-Aviv, Israel.
24. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Widener, T. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, *39*, 27–34.
25. Federal Communications Commission (2011) Small entity compliance guide: Wireless E911 location accuracy requirements. Federal communications commission: Report and order FCC 10-176 PS docket No 07-114 p 3
26. Ganapathiraju, M., Weisser, D., Rosenfeld, R., Carbonell, J., Reddy, R., & Klein-Seetharaman, J. (2002). Comparative n-gram analysis of whole-genome protein sequences. In Proceedings of the second international conference on Human Language Technology Research, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, HLT '02, pp 76–81.
27. 3rd Generation Partnership Project. (2009a). Evolved universal terrestrial radio access network (e-utran); self-configuring and self-optimizing network (SON) use cases and solutions (release 9). Tech. Rep. TR 36.902, 3GPP
28. 3rd Generation Partnership Project. (2009b). Technical specification group radio access network; study on minimization of drive-tests in next generation networks (release 9). Tech. Rep. TR 36.805, 3GPP
29. 3rd Generation Partnership Project. (2010). 3GPP; TSG radio access network; further advancements for e-utra physical layer aspects (release 9). Tech. Rep. TR 36.814, 3GPP
30. 3rd Generation Partnership Project. (2011). Technical specification group radio access network; evolved universal terrestrial radio access (e-utra); radio resource control (rrc); protocol specification (release 10). Tech. Rep. TS 36.331, 3GPP
31. 3rd Generation Partnership Project. (2012). Technical report 3rd generation partnership project; technical specification group radio access network; evolved universal terrestrial radio access (e-utra); mobility enhancements in heterogeneous networks (release 11). Tech. rep., 3GPP
32. 3rd Generation Partnership Project. (2014). Self-organizing networks (SON); self-healing concepts and requirements (release 12). Tech. rep., 3GPP TS 32.541 V12.0.0
33. GSMA Intelligence (2014) Understanding 5g: Perspectives on future technological advancements in mobile. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html
34. Guillet, F., & Hamilton, H. J. (Eds.). (2007). *Quality measures in data mining, studies in computational intelligence* (Vol. 43). Berlin: Springer.
35. Haidar, M., & O'Shaughnessy, D. (2012). Topic n-gram count language model adaptation for speech recognition. In Spoken language technology workshop (SLT), 2012 IEEE, pp 165–169, doi:10.1109/SLT.2012.6424216.
36. Hämälainen, S., Sanneck, H., & Sartori, C. (2012). *LTE self-organising networks (SON): Network management automation for operational efficiency* (1st ed.). Hoboken: Wiley Publishing.
37. Han, J., & Kamber, M. (2006). *Data mining: Concepts and techniques* (2nd ed., Vol. 54). Burlington: Morgan Kaufmann.
38. Hand, D. J., Smyth, P., & Mannila, H. (2001). *Principles of data mining*. Cambridge, MA, USA: MIT Press.
39. Hapsari, W., Umesh, A., Iwamura, M., Tomala, M., Gyula, B., & Sebire, B. (2012a). Minimization of drive tests solution in 3GPP. *Communications Magazine, IEEE*, *50*(6), 28–36.
40. Hapsari, W., Umesh, A., Iwamura, M., Tomala, M., Gyula, B., & Sebire, B. (2012b). Minimization of drive tests solution in 3GPP. *Communications Magazine, IEEE*, *50*(6), 28–36.
41. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transaction on Knowledge and Data Engineering*, *21*(9), 1263–1284. doi:10.1109/TKDE.2008.239.
42. He, Z., Cichocki, A., & Xie, S. (2009). Efficient method for tucker3 model selection. *Electronics Letters*, *45*, 805.
43. He, Z., Cichocki, A., Xie, S., & Choi, K. (2010). Detecting the number of clusters in n-way probabilistic clustering. *IEEE Transaction Pattern Analysis Machine Intelligence*, *32*(11), 2006–2021.
44. Holma, H., & Toskala, A. (2011). *LTE for UMTS: Evolution to LTE-advanced* (2nd ed.). Hoboken: Wiley Publishing.
45. Islam, A., & Inkpen, D. (2009). Real-word spelling correction using Google Web 1t n-gram with backoff. In Natural language processing and knowledge engineering, 2009. NLP-KE 2009. International conference on, pp 1 –8, doi:10.1109/NLPKE.2009.5313823
46. Johansson, J., Hapsari, W., Kelley, S., & Bodog, G. (2012). Minimization of drive tests in 3GPP release 11. *Communications Magazine, IEEE*, *50*(11), 36–43.
47. Jolliffe, I. (2002). *Principal component analysis. Springer series in statistics*. Berlin: Springer.
48. Kac, M., Kiefer, J., & Wolfowitz, J. (1955). On tests of normality and other tests of goodness of fit based on distance methods. *The Annals of Mathematical Statistics*, *26*(2), 189–211.
49. Kassis, E. (2010). Anomaly-based error detection in base station data. Master's thesis, Tel-Aviv University, Israel
50. Kela, P. (2007). Downlink channel quality indication for evolved universal terrestrial radio access network. Master's thesis, University of Jyväskylä, Finland.
51. Khanafer, R., Solana, B., Triola, J., Barco, R., Moltsen, L., Altman, Z., et al. (2008). Automated diagnosis for umts networks using bayesian network approach. *Vehicular Technology, IEEE Transactions on*, *57*(4), 2451–2461. doi:10.1109/TVT.2007.912610.
52. Kolehmainen, N. (2007). Downlink packet scheduling performance in evolved universal terrestrial radio access network. Master's thesis, University of Jyväskylä, Finland.
53. Laiho, J., Raivio, K., Lehtimaki, P., Hatonen, K., & Simula, O. (2005). Advanced analysis methods for 3g cellular networks. *Wireless Communications, IEEE Transactions on*, *4*(3), 930–942. doi:10.1109/TWC.2005.847088.
54. Luo, F. L., Unbehauen, R., & Cichocki, A. (1997). A minor component analysis algorithm. *Neural Networks*, *10*(2), 291–297.
55. Mueller, C. M., Kaschub, M., Blankenhorn, C., & Wanke, S. (2008). A cell outage detection algorithm using neighbor cell list reports. In K. Hummel & J. Sterbenz (Eds.), *Self-organizing systems* (Vol. 5343, pp. 218–229)., Lecture notes in computer science Berlin Heidelberg: Springer.
56. Nagao, Makoto, Mori, Shinsuke (1994) A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese. In: Proceedings of the 15th conference on computational linguistics - volume 1, association for computational linguistics, Stroudsburg, PA, USA, COLING '94, pp 611–615
57. Next Generation Mobile Networks (2008a) Recommendation on SON and O&M Requirements. Tech. rep., NGMN, URL http://www.ngmn.org/

58. Next Generation Mobile Networks (2008b) Use Cases related to Self Organising Network, overall description. Tech. rep., NGMN, URL http://www.ngmn.org/
59. Next Generation Mobile Networks (2015) NGMN 5G Initiative White Paper. https://www.ngmn.org/uploads/media/NGMN_5G_White_Paper_V1_0.pdf
60. Novaczki, S. (2013). An improved anomaly detection and diagnosis framework for mobile network operators. In Design of reliable communication networks (DRCN), 2013 9th international conference on the, pp 234–241.
61. Novaczki, S., & Szilagyi, P. (2011). Radio channel degradation detection and diagnosis based on statistical analysis. In Vehicular technology conference (VTC Spring), 2011 IEEE 73rd, pp 1–2.
62. NTT DOCOMO Inc (2014) Docomo 5g white paper: 5g radio access: Requirements, concept and technologies. https://www.nttdocomo.co.jp/english/binary/pdf/corporate/technology/whitepaper_5g/DOCOMO_5G_White_Paper.pdf
63. Osseiran, A., Braun, V., Hidekazu, T., Marsch, P., Schotten, H., Tullberg, H., Uusitalo, M., & Schellman, M. (2013). The foundation of the mobile and wireless communications system for 2020 and beyond: Challenges, enablers and technology solutions. In Vehicular technology conference (VTC Spring), 2013 IEEE 77th, pp 1–5, doi:10.1109/VTCSpring.2013.6692781.
64. Rabin, N. (2010). Data mining dynamically evolving systems via diffusion methodologies. PhD thesis, Tel-Aviv University, Tel-Aviv, Israel.
65. Raivio, K., Simula, O., Laiho, J., & Lehtimaki, P. (2003). Analysis of mobile radio access network using the self-organizing map. In Integrated network management, 2003. IFIP/IEEE Eighth international symposium on, pp 439–451, doi:10.1109/INM.2003.1194197.
66. Ramaswamy, S., Rastogi, R., & Shim, K. (2000). Efficient algorithms for mining outliers from large data sets. *SIGMOD Record*, *29*(2), 427–438.
67. Ramiro, J., & Hamied, K. (2012). *Self-organizing networks (SON): Self-planning, self-optimization and self-healing for GSM, UMTS and LTE* (1st ed.). Hoboken: Wiley Publishing.
68. Scully, N., et al. (2008). D2.1: Use cases for self-organising networks. URL http://www.fp7-socrates.eu
69. Sesia, S., Baker, M., & Toufik, I. (2011). *LTE - The UMTS long term evolution: From theory to practice*. Hoboken: Wiley.
70. Szilagyi, P., & Novaczki, S. (2012). An automatic detection and diagnosis framework for mobile communication systems. *IEEE Transactions on Network and Service Management*, *9*(2), 184–197.
71. Turkka, J., Ristaniemi, T., David, G., & Averbuch, A. (2011). Anomaly detection framework for tracing problems in radio networks. In The 10th international conference on networks, ICN 2011.
72. Turkka, J., Chernogorov, F., Brigatti, K., Ristaniemi, T., & Lempiäinen, J. (2012). An approach for network outage detection from drive-testing databases. Journal of Computer Networks and Communications.
73. Yilmaz, O. N. C., Hämälainen, J., & Hämälainen, S. (2011). Self-optimization of random access channel in 3rd generation partnership project long term evolution. *Wireless Communications and Mobile Computing*, *11*(12), 1507–1517.

Fedor Chernogorov received his M.Sc. degree with honors in Telecommunications in 2009 from P.G. Demidov Yaroslavl State University, Yaroslavl, Russia. From the University of Jyvaskyla, Finland he received his M.Sc. in Mobile Technology in 2010, and his Ph.D. in Mathematical Information Technology in 2015. Since 2011 he is with Magister Solutions, Jyvaskyla, Finland, where he currently works at position of Senior Researcher. He has authored and co-authored 9 conference and 2 journal publications in the fields of cellular mobile communications and data mining, anomaly detection. Areas of his interest are self-organizing and cognitive mobile networks, advanced performance monitoring in cellular, knowledge mining, anomaly detection and data mining.

Sergey Chernov received his M.Sc. degree with honors in radio physics and electronics in 2008 in Yaroslavl State University, Russia. Since 2011 he is Ph.D. student at the faculty of informational technologies in the University of Jyvaskyla, Finland. His current research is focused on Self-Organizing Network concept of Long-Term Evolution mobile cellular networks and the application of data mining and machine learning algorithms to the various radio network data.

Kimmo Brigatti received his M.Sc. (Information Technology) in 2011 from University of Jyvaskyla, Jyvaskyla, Finland. He has been involved in different research groups on data mining, anomaly detection and wireless network technologies starting from 2009 in University of Jyvaskyla. Kimmo has co-authored two publications N-Gram Analysis For Sleeping Cell Detection in LTE Networks (ICASSP, 2013) and An Approach for Network Outage Detection from Drive-Testing Databases (Journal of Computer Networks and Communications, 2012).

Tapani Ristaniemi (SM'11) received his M.Sc. degree in mathematics in 1995, the Ph.Lic. degree in applied mathematics in 1997, and the Ph.D. degree in wireless communications in 2000, all from the University of Jyväskylä, Jyväskylä, Finland. In 2001, he was appointed as a Professor in the Department of Mathematical Information Technology, University of Jyväskylä. In 2004, he moved to the Department of Communications Engineering, Tampere University of Technology, Tampere, Finland, where he was appointed as a Professor of Wireless Communications. In 2006, he moved back to the University of Jyväskylä to take up his appointment as a Professor of Computer Science. He is an Adjunct Professor of Tampere University of Technology. In 2013, he was a Visiting Professor in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. He has authored or co-authored over 200 publications in journals, conference proceedings, and invited sessions. He served as a Guest Editor of IEEE WIRELESS COMMUNICATIONS in 2011 and currently he is an Editorial Board Member of Wireless Networks and the International Journal of Communication Systems. His research interests are in the areas of brain and communication signal processing and wireless communication systems. Besides academic activities, Professor Ristaniemi is also active in the industry. In 2005, he co-founded a start-up, Magister Solutions, Ltd., in Finland, specializing in wireless systems (R&D) for telecom and space industries in Europe. Currently, he serves as a consultant and a Chairman of the Board.

PVI

THE INFLUENCE OF DATASET SIZE ON THE PERFORMANCE OF CELL OUTAGE DETECTION APPROACH IN LTE-A NETWORKS

by

Sergey Chernov, Mykola Pechenizkiy, Tapani Ristaniemi 2015

10th IEEE International Conference on Information, Communications and Signal
Processing (ICICS), Singapore

Reproduced with kind permission of IEEE.