Jyväskylän yliopiston julkaisuarkisto
Jyväskylä University Digital Archive

UNIVERSITY OF JYVÄSKYLÄ

This is an electronic reprint of the original article.
This reprint *may differ* from the original in pagination and typographic detail.

Author(s): Saarela, Mirka; Kärkkäinen, Tommi

Title: Weighted Clustering of Sparse Educational Data

Year: 2015

Version:

Please cite the original version:

Saarela, M., & Kärkkäinen, T. (2015). Weighted Clustering of Sparse Educational Data. In ESANN 2015 : Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (pp. 337-342). ESANN. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2015-24.pdf

All material supplied via JYX is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

Weighted Clustering of Sparse Educational Data

Mirka Saarela and Tommi Kärkkäinen

University of Jyväskylä - Department of Mathematical Information Technology
40014, Jyväskylä - Finland

Abstract. Clustering as an unsupervised technique is predominantly used in unweighted settings. In this paper, we present an efficient version of a robust clustering algorithm for sparse educational data that takes the weights, aligning a sample with the corresponding population, into account. The algorithm is utilized to divide the Finnish student population of PISA 2012 (the latest data from the Programme for International Student Assessment) into groups, according to their attitudes and perceptions towards mathematics, for which one third of the data is missing. Furthermore, necessary modifications of three cluster indices to reveal an appropriate number of groups are proposed and demonstrated.

1 Introduction

The application of clustering in a weighted context is a relatively unresearched topic [1]. PISA (Programme for International Student Assessment) is a worldwide study that triannually assesses proficiency of 15-year-old students from different countries and economies in the three domains, reading, mathematics, and science. Besides the reporting of student performances, PISA is also one of the largest public databases[1] in which students' demographic and contextual data, such as their attitudes and behaviors towards education related topics, is collected and stored.

PISA data are an important example of a large data set that includes weights. In general, weighting is a technique in survey research to align the sample to more accurately represent the true population. Namely, only a fraction of students from each country take part in the PISA assessment but, when taking the weights into account, they should be representative for the whole population. For example, the Finnish sample data of the latest PISA assessment consists of 8829 students whose analysis results, when multiplied with the respective weights, represent the whole 60047 15-year-old student population of the country. As can be seen from Fig. 1, in which the studentwise weights are depicted, the minimal weight in the Finnish national subset of PISA is 1, i.e. each students represents at least him/herself, while the maximal weight is more than 54.

A further important characteristic of PISA data is the large number of missing values. Because PISA uses a rotated design [2] and some students are not administered certain questions, the majority of the missing data in PISA is missing by design, which can be seen as a special case of *missing completely at random* [3, 4]. Altogether, there are 634 raw variables in the PISA student questionnaire data set of the latest assessment. However, a subset of 15 derived

[1]PISA data can be downloaded from `http://www.oecd.org/pisa/pisaproducts/`.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

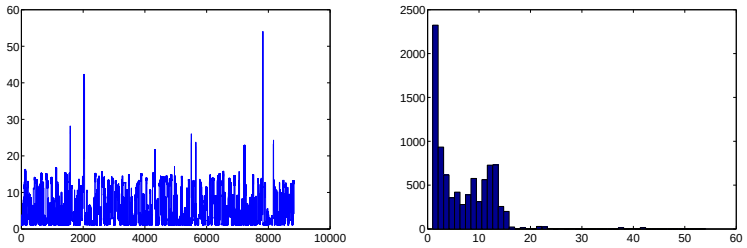

Fig. 1: Individual weights (left) and their discrete distribution (right) in Finnish 2012 PISA data.

variables, the so-called PISA scale indices[2], readily describe students' attitudes and perceptions, e.g., explaining the performance in mathematics [2, 5]. Each scale index is a compound variable and constructed using the students' answers to certain background questions. Nevertheless, mainly because of the rotated design, 33.24% of these scale indices are not available.

In [5] we utilized a robust clustering algorithm to the Finnish sample of PISA 2012 scale indices, which revealed very gender-specific contrasts in the different clusters. For the interpretation of the clustering result, we employed the weights to summarize the cluster prototypes on the population level. However, according to the PISA data analysis manual [6], one should always, particularly when over- or under-sampling has taken place, include weights at *each stage* of the analysis.

Therefore, the research questions of this paper are as follows: (i) how to efficiently cluster sparse student data on the population level, i.e., how the weights in the sample should be incorporated in the robust clustering algorithm and (ii) how much the two clustering results with and without weights (sample division vs. population division) differ from each other? Both questions are relevant for the Finnish subset of PISA data because immigrants as well as students from Swedish-speaking schools were deliberately over-sampled in the latest assessment.

2 Weighted robust clustering of sparse data

In general, partitioning-based clustering algorithms are composed of an initialization followed by the iterations of the two basic steps, where each observation is first assigned to its closest prototype and, then, each prototype is updated based on the assigned subset of data. As pointed out in [5], sparse data sets can be reliably clustered by utilizing the so-called *k-spatialmedians* [7] algorithm. Compared to k-means, the k-spatialmedians uses the spatial median to estimate the prototypes, which is statistically robust and can handle large amount of contamination (noise and missing values) in data.

However, because of the local search character of the partitioning-based clustering algorithms, their result depends on the initialization. For a sparse data set

[2]These scale indices are explicitly listed in [5].

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

with missing values, a proper initialization should posses, at least, two desired properties: it should reflect the subset of data with full observations, because inevitably missing values decrease reliability of the cluster allocations. Furthermore, the initial prototypes should be full, i.e., without missing values, because the cluster assignment and recomputation, e.g., as in [5], assumes this throughout the whole iterative procedure. Lately the k-means++ algorithm [8], where the random initialization is based on using a density function favoring distinct prototypes, has become popular.

Therefore, our general procedure to cluster the sparse data on the population-level is as follows. First of all, the subset of data that has no missing values is clustered using k-means++. Then, the robust clustering algorithm is applied for the whole sparse data by utilizing the obtained prototypes as initialization. Altogether, the final clustering result is statistically robust with respect to degradations in data, probably with full prototypes (especially when a small number of clusters is created from a large data set), and reflecting the spherical and possibly already separated shape of the full data subset.

The precise form of the general clustering criterion to be minimized (locally) by the iterative reallocation algorithm, with weights and missing values, reads as follows:

$$\mathcal{J}(\{c_k\}_{k=1}^K) = \sum_{k=1}^{K} \sum_{i \in I_k} w_i \|\boldsymbol{P}_i(\boldsymbol{c}_k - \boldsymbol{x}_i)\|_2^p, \tag{1}$$

where I_k denotes the indices of data assigned to the kth cluster and $\boldsymbol{P}_i$'s define the sparsity pattern (i.e., indicate available variables) observationwise:

$$(\boldsymbol{P}_i)_j = \begin{cases} 1, \text{if } (\boldsymbol{x}_i)_j \text{ exists,} \\ 0, \text{otherwise.} \end{cases}$$

In the k-spatialmedians algorithm for $p = 1$, the cluster prototypes are computed using a modifed SOR (Sequential Overrelaxation) algorithm [7], where weights are taken into account in the updates. Furthermore, in order the align the k-means-type initialization with $p = 2$ in (1) to the actual case $p = 1$, we propose to use $\{\sqrt{w_i}\}$'s as weights in k-means++ because, simply, $\alpha \, \|\boldsymbol{P}_i(\boldsymbol{c}_k - \boldsymbol{x}_i)\|_2^p = (\sqrt[p]{\alpha} \, \|\boldsymbol{P}_i(\boldsymbol{c}_k - \boldsymbol{x}_i)\|_2)^p$, for $\alpha > 0$.

To this end, to determine a single result of the partitioning-based weighted clustering procedure, one also needs to estimate the number of clusters K. For this purpose, we used three modified internal cluster validation indices, namely the Ray-Turi [9], the Davies-Bouldin [10], and the Davies-Bouldin* [10]. Essentially, we included the weights in the computations of the clusterwise scatter matrices, used the final value of (1) as the clustering error, and computed distances between the prototypes by using the Euclidean norm.

3 Experimental results

The tests concentrate on analyzing the use of weights in the initial partition utilizing k-means++, followed by the actual weighted k-spatialmedians. Namely,

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

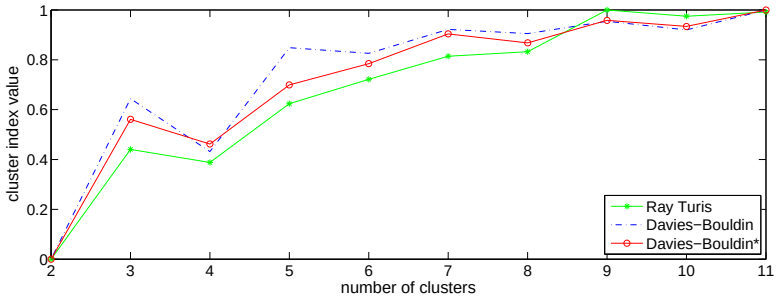

Fig. 2: Cluster indices for sparse data scaled into range $[0, 1]$.

one can use/omit the weights in i) the initialization of k-means++ and ii) the iterative reallocations of k-means++, which creates three possible algorithmic scenarios. First of all, all of these possibilities were applied to assess the number of clusters using the modified cluster indices. The result is given in Fig. 2 where the averages of 30 runs (ten for each variant for each k) is depicted. One concludes that all three cluster indices suggest that, for the Finnish 2012 population data, four clusters is an appropriate choice[3]. This is the same number that was obtained for the Finnish sample data without weighting (see [5]).

Next we fix $k = 4$, i.e., test the speed (number of iterations) and quality of the three algorithmic combinations for four clusters. The results with 10 repeated test runs are given in Table 1, together with the average of the ten repetitions in the last row. We report the number of iterations needed in the initialization (i.e. within k-means++), the number of iterations needed in the actual k-spatialmedians clustering with the whole sparse data, and also the final quality of the clustering result (i.e., the clustering error).

All three main columns of Table 1 show that including the weights in k-means++ for complete data before k-spatialmedians improves the performance of the latter as less iterations are needed. Similarly, to include square-rooted weights[4] in the initialization of k-means++ improves the performance of the whole initial procedure (see the last two main columns). Concerning the clustering error, we obtained similar error levels with all the approaches (see the last row of Table 1) but less variability when using the weights. Therefore, we conclude that appropriately scaled weights should be present in both places in the initialization in order to achieve an efficient and robust weighted clustering algorithm.

Using the fully weighted algorithm with the average of 10 runs, we obtain in practice the same four clusters as in the unweighted case (see [5] in which the clusters and their implications are discussed) with very similar characteristics

[3]Actually, all three indices have the best value at two but having only two clusters divides our data simply in high- and low-performing students which does not provide any interesting patterns additionally.

[4]Incorporating the weights into k-means++ simply as w instead of $\sqrt{w}$ was also tested. But since $\sqrt{w}$ gave, as we proposed in Sec. 2, better results, only these are reported here.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8.
Available from http://www.i6doc.com/en/.

(see Table 2). The prototypes that describe the four clusters are almost identical. In particular, also with weights the cluster *C2* of mostly girls, with very positive attitudes towards school and learning but no intentions to use mathematics later in life, appear. Also an opposite cluster *C3* with the majority of boys, that have the highest intentions to pursue a mathematics related career but otherwise very negative attitudes towards education, is present, together with the groups of advantaged high-performing students (*C1*) and their more disadvantaged lower performing peers (*C4*).

4 Conclusions

In this paper, we modified the k-spatialmedians algorithm [7], an algorithm that can handle large amounts of missing data, in such a way that it can be used also for weighted clustering. In order to have an as fast and deterministic approach as possible, we also introduced weights to the seeding as well as the actual main body of the k-means++ algorithm which we use in the initialization. Experiments showed that, indeed, the best, i.e. the fastest as well as most accurate, population-based clustering solution is obtained when weights are incorporated in all phases of the algorithm.

As pointed out in the introduction, though weighted clustering has been investigated in theory, it has not been examined much in an applied context. PISA data sets are prime examples of large data sets with many missing values as well as weights. We applied weighted clustering to the Finnish subset of the latest PISA data. Although over-sampling took place for some groups of the student population, no significant differences in the final results existed, i.e. the general

Without weights in k-means++			$\sqrt[p]{w_i}$ weights in iterative reallocation			$\sqrt[p]{w_i}$ weights in entire algorithm		
iter. in ini.	iter. in alg.	cluster error (quality)	iter. in ini.	iter. in alg.	cluster error (quality)	iter. in ini.	iter. in alg.	cluster error (quality)
23	34	5.9464	34	30	0.6458	21	28	0.6035
23	38	0.5176	34	30	0.6458	14	30	0.5424
19	33	0.5161	41	33	0.5176	23	30	0.5424
27	38	0.5176	42	30	0.5176	29	30	0.5424
23	34	0.4983	34	33	0.6458	18	29	0.5424
23	38	0.5176	34	30	0.6458	20	30	0.5424
21	44	6.0403	43	30	0.6458	22	30	0.5424
18	38	0.5176	39	33	0.5176	24	30	0.5424
25	38	0.5176	41	33	0.6458	26	28	0.6035
20	37	0.5176	34	30	0.6458	22	28	0.6035
20	38	1.6108	41	31	0.6073	22	29	0.5607

Table 1: Efficacy and quality of clustering result with and without weights in initialization. The base level 127450 has been subtracted from all cluster errors.

ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22-24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

	valid	sample	**population size**			**math score**		
cluster	indices	size	all	♀ (in %)	♂	∅	♀	♂
C1	65%	2009	13203	5311 (40%)	7893	574	581	569
C2	68%	2242	14418	8955 (62%)	5463	510	516	499
C3	67%	2450	16723	6495 (39%)	10229	532	539	528
C4	66%	2128	15703	8450 (54%)	7253	466	472	460
C1-C4	67%	8829	60047	29210 (49%)	30837	519	520	517

Table 2: Facts of population clusters

profiles of the clusters without weights (sample) and with weights (population) were almost identical. However, even though the algorithm is deterministic after the initialization, and the accuracy of clustering is improved when initialized with k-means++, still some randomness in the final clustering result remains due to the randomness in seeding. Hence, a complete comparison between clustering results persists challenging, not only for population- vs. sample-based clustering but also for clustering in general.

References

[1] Margareta Ackerman, Shai Ben-David, Simina Branzei, and David Loker. Weighted clustering. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[2] OECD. Pisa 2012 technical background. 2013.

[3] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[4] Donald B Rubin and Roderick JA Little. Statistical analysis with missing data. *Hoboken, NJ: J Wiley & Sons*, 2002.

[5] Mirka Saarela and Tommi Kärkkäinen. Discovering gender-specific knowledge from finnish basic education using pisa scale indices. In *Proceedings of the 7th International Conference on Educational Data Mining*, pages 60–68, 2014.

[6] OECD. *PISA Data Analysis Manual: SPSS and SAS, Second Edition*. OECD Publishing, 2009.

[7] Sami Äyrämö. *Knowledge Mining Using Robust Clustering*, volume 63 of *Jyväskylä Studies in Computing*. University of Jyväskylä, 2006.

[8] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[9] Siddheswar Ray and Rose H Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143, 1999.

[10] Minho Kim and RS Ramakrishna. New indices for cluster validity assessment. *Pattern Recognition Letters*, 26(15):2353–2363, 2005.