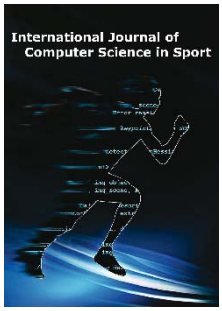
International Journal of Computer Science in Sport

Volume 18, Issue 3, 2019

Journal homepage: http://iacss.org/index.php?id=30

DOI: 10.2478/ijcss-2019-0021



Talent identification in soccer using a one-class support vector machine

Jauhiainen S.[1], Äyrämö S.[1], Forsman H.[2], Kauppi J-P.[1]

[1]Faculty of Information Technology, University of Jyväskylä, Finland

[2]Eerikkila Sports & Outdoor Resort, Training and Research Centre forFinnish Football Urheiluopistontie Eerikkila, Tammela

Abstract

Identifying potential future elite athletes is important in many sporting events. The successful identification of potential future elite athletes at an early age would help to provide high-quality coaching and training environments in which to optimize their development. However, a large variety of different skills and qualities are needed to succeed in elite sports, making talent identification generally a complex and multifaceted problem. Due to the rarity of elite athletes, datasets are inherently imbalanced, making classical statistical inference difficult. Therefore, we approach talent identification as an anomaly detection problem. We trained a nonlinear one-class support vector machine (one-class SVM) on a dataset (N=951) collected from 14-year-old junior soccer players to detect potential future elite players. The mean area under the receiver operating characteristic curve (AUC-ROC) over the tested hyperparameter combinations was 0.763 (std 0.007). The most accurate model was obtained when physical tests, measuring, for example, technical skills, speed, and agility, were used. According to our results, the proposed approach could be useful to support decision-makers in the process of talent identification.

KEYWORDS: TALENT IDENTIFICATION, ANOMALY DETECTION, ONE-CLASS SVM

Introduction

The amount of data in sports is rapidly increasing due to advances in data collection technologies (Brefeld & Zimmermann, 2017). This has opened many possibilities for data analysis and application development across all sports. Even though sports analytics is a relatively new field, a variety of different research questions, approaches and data sources are already documented in the literature (Brefeld & Zimmermann, 2017). For example, data analysis has been used for predicting outcome of a game (Aoki, Assuncao, & de Melo, 2017), decision support for passing in soccer (Power, Ruiz, Wei, & Lucey, 2017), and optimization of a training schedule (Knobbe, Orie, Hofman, van der Burgh, & Cachucho, 2017). Commonly used data analysis methods with examples from elite sports are introduced in (Ofoghi, Zeleznikow, MacMahon, & Raab, 2013). One example of data analysis utilized in talent identification is PECOTA (Player Empirical Comparison and Optimization Test Algorithm). It calculates different career paths for baseball players and forecasts player's performance utilizing similarity scores and projection (Silver, 2003). Another study, focusing on decision making in sport management, used the ordered weighted averaging (OWA) operator for selection of players (Merigó & Gil-Lafuente, 2011). In addition to talent identification, applications for suggesting the best sports in terms of athlete's individual capabilities have been developed (Papić, Rogulj, & Pleština, 2009).

The identification and selection of talented players at an early age is important in many sporting events. It will enable offering high-quality coaching and training environments for talented players and thereby accelerate their development (Williams & Reilly, 2000). However, the identification task, especially in team games, is a very complex process (Reilly, Williams, Nevill, & Franks, 2000). In soccer, for example, a great variety of physical features and technical skills are needed for success (Reilly et al., 2000). Moreover, psychological skills and characteristics also play an important role at elite level (Macnamara & Collins, 2011). Therefore, talent identification in sports should be based on a versatile set of variables.

Furthermore, the datasets in talent identification are inherently imbalanced due to the rarity of elite athletes in sports. In practice, this means that there are typically significant differences in the number of observations available from different classes (Chawla, Japkowicz, & Kotcz, 2004), which must be carefully taken into account when designing a machine learning method for the case at hand. Imbalanced datasets are common in many other real-life applications as well (He & Garcia, 2009). In this study only 14 observations were available for the minority class, whereas majority class consisted of almost thousand observations.

The two main approaches in data analysis are explanatory and predictive modeling (Breiman, 2001). Many previous studies on talent identification have concentrated on explanatory data analysis (Nieuwenhuis, Spamer, & Rossum, 2002; O'Connor, Larkin, & Mark Williams, 2016; Woods, Raynor, Bruce, McDonald, & Robertson, 2016). Although differences between talented and other players have been found, the predictive power of their models is unclear. Therefore, there is a need for models whose predictive power has been evaluated on independent test data. This need has also been noticed in other studies, such as the one by Smiths, Lipscomb, and Simkins (2007), where a predictive approach is applied on award prediction in baseball, a task that has been previously approached with explanatory methods.

In this research, we studied the potential of machine learning in talent identification, using data containing a diverse set of variables. Our goal was to analyze whether potential future elite players can be distinguished from majority of the players based on their test information already as juniors. Because of the limited number of observations in the minority class, the use of supervised machine learning methods would easily lead to overfitting. Therefore, we used one-class classification approach, where the training stage was completely unsupervised, and

information about the class labels was only used in model assessment (Goldstein & Uchida, 2016).

Methods

The target data of this study were collected by *The training and research centre for Finnish football* for monitoring the development of young soccer players. A total of 4991 junior soccer players (age mean±std 12.41±1.53 year, range 8-18 year) participated in the specific test events organized by the centre between the years 2011 and 2017. Out of the 47 participating teams, 41 were Finnish, but 293 players came from Sweden, Denmark, England, and the Netherlands. The participating Finnish teams are the best of their age group in Finland.

Each player participated in the test events twice a year together with their team. During the events, the players performed physical tests including, e.g., technical, speed, and agility tests. Moreover, they completed a self-assessment test including, e.g., perceived competence, tactical skills, and motivation. Description of the test protocol can be found in Forsman et al. (2016). The physical tests were measured in a continuous scale (such as length of a 5-jump or time of a speed test). The questionnaire scale was a discrete 5-point Likert scale concerning sport performance, anchored with 1 (almost never) and 5 (almost always).

Data selection

Our goal was to investigate how accurately we can detect potential future elite players among the large pool of players based on the collected test information. Some of the tested Finnish players in our dataset are currently pursuing an international soccer career and have already signed a contract with an international academy. In the absence of senior players who have reached the absolute elite performance level by playing, e.g., for a national team, these international academy players (from now on called ”academy players”) were labelled as talent category for the present study. The player categorization was defined by an educated person in charge of player development at the training and research centre for Finnish football. The players representing other than Finnish teams were excluded from the further analysis due to the insufficient information of their current career development.

All academy players were boys and all of them had performed the tests at the age of 14. For these reasons, 14-year-old Finnish boys were selected for our analyses. The age limit for signing a contract to an academy is 16 years. Therefore, we dropped out of the study those players born in 2003 or later as at the time of this study they could not have a signed contract even though they might be future academy players. In the whole dataset, the total number of academy players was 26. Twelve of these players were dropped out from the analysis due to the overly many missing test results. The final data set included 14 academy players.

Further pruning of the data set was performed due to a large number of variables with a significant proportion of non-random missing values, i.e., they followed *not missing at random* (NMAR) pattern (Little & Rubin, 2002). These missing values were caused, for example, by adjustments to the test protocols and questionnaires or inability of a player to participate all the test events.

Finally, the used data representation (called ”*phys large*”, N=951) consisted of 16 variables in which the test results were measured for at least half of the players (see Table 1). Since the questionnaire answers were missing from more than half of the players they were not included in “*phys large”*. In order to characterize univariate differences between the academy and nonacademy players two-sample t-tests were performed. Normality of the variables was tested

using Shapiro-Wilk test (if $n < 50$) or Kolmogorov-Smirnov test (otherwise). Homogeneity of the variances was tested with Levene test. When the assumption on normality failed Wilcoxon rank sum test was used. Significance level $\alpha = 0.05$ with Bonferroni correction was used and effect size Cohen's *d* reported (Cohen, 1988). All test were performed using MATLAB version R2016b.

Table 1: Mean/median and standard deviation of the variables in "*phys large*" separately for the 14 academy players and 937 non-academy players. A statistically significant difference between the groups was found with 5-jump, height, and weight ($*p < 0.05, d > 0.8$; $**p < 0.01, d > 0.8$).

Countinous variables	**Mean (std) of non-academy players**	**Mean (std) of academy players**
5 jump (m)**	10.90 (0.85)	11.69 (0.63)
Agility (sec)	6.98 (0.58)	6.87 (0.26)
Countermovement jump (cm)	29.87 (4.58)	32.54 (4.76)
Driving and shooting (sec)	15.03 (4.19)	13.36 (4.18)
Speed 10 meters (sec)	1.80 (0.09)	1.74 (0.06)
Speed 20 meters (sec)	3.19 (0.16)	3.05 (0.10)
Speed 30 meters (sec)	4.49 (0.25)	4.30 (0.14)
Speed 5 meters (sec)	1.03 (0.05)	0.99 (0.04)
Weight (kg)*	55.59 (9.48)	62.41 (5.65)
Height (cm)**	167.75 (8.93)	176.33 (6.06)
Juggling (sec)	24.63 (7.58)	22.34 (8.01)
Dribbling (sec)	25.74 (2.78)	25.20 (1.92)
Passing (sec)	37.15 (6.12)	34.94 (6.01)
Gymnastics (points)	12.22 (1.96)	12.05 (2.47)
Yo-Yo endurance (m)	2248.18 (319.00)	2480.00 (330.32)
Discrete variable	Median of non-academy Players	Median of academy Players
Mobility (1-3)	3	3

In order to investigate the predictive power of questionnaire data, another representation (called "*phys+quest*") of the data with fewer players (N = 468), but a greater number of variables including all the 16 "*phys large*" variables and additionally 18 variables from a specific questionnaire consisting of self-assessment of perceived competence, was defined (see Table 2). The questionnaire measures how the players rate their skills in offense, defense, and one versus one situations. Four out of the 14 academy players did not answer the questionnaire and they were dropped off. A detailed description of the questionnaire can be found in Forsman et al. (2016).

Table 2: Questionnaire variables used in this study.

M1. Mean of offensive skill questions
M2. Mean of 1-on-1 skill questions
M3. Mean of defensive skill questions

Q1. I can schedule my own movement correctly in offensive and defensive play

Q2. I have clear solution models about how to win 1-on-1 situations

Q3. I am usually the first player to reach the ball

Q4. I can easily lose my opponent in different game situations

Q5. I feel strong in match ups

Q6. In 1-on-1 situations, I am stronger/faster than my opponent

Q7. I can accomplish the typical play for my position in defensive play

Q8. I can, if necessary, help/support my teammates in defensive situations

Q9. I have a soft ”touch” on the ball

Q10. I dare to keep the ball to myself even in tight spaces

Q11. I have clear solution models about how I score in the different situations in the games

Q12. I can move to the empty spaces on the field, so that my teammates can pass me the ball

Q13. I can find my teammates with my sharp and accurate passes

Q14. I can accomplish the typical play for my position in offensive play

Q15. I know how my teammates are moving in attack situations and it is easy for me to pass them the ball

In order to analyze the predictive power of physical tests and perceived competence self-assessment independently of each other, the variables in ”*phys+quest*” were further split into two smaller representations consisting of only either physical (called ”*phys*”, N =468) or questionnaire (called ”*quest*”, N = 468) variables. The number of academy players was ten in both presentations. A comparison between ”*phys*” and ”*phys large*” representations was performed in order to evaluate the effect of sample size to the predictive accuracy. The summary of the four data representations is shown in Table 3.

Table 3: Different representations of the player data analyzed in this study.

Data representation	N	Variables	D
Phys large	951	Physical variables	16
Phys+quest	468	Physical variables + questionnaire	34
Phys	468	Physical variables	16
Quest	468	Questionnaire	18

Data preprocessing

Prior to model fitting all the physical variables were normalized and the Likert scale questionnaire variables were min-max scaled to range $[\frac{1}{10}, \frac{9}{10}]$. After removing observations and variables due to NMAR values, the remaining missing values were imputed using a self-implemented k-nearest neighbor (knn) imputation algorithm on MATLAB (Bishop, 2006). The estimate for each missing value was computed as the sample mean of the ten nearest neighbours based on Euclidean distance. Although the standard MATLAB knn classier uses $k = 1$ as default,

larger values of k have been found to control noise and perform better. Several studies have found the method relatively insensitive to the exact value of k between 10-20 (Beretta & Santaniello, 2016; Troyanskaya et al., 2001). In addition, principal component analysis (PCA) (Jolliffe, 1986) was used to eliminate correlations from the data. The minimal number of PCs which explained at least 90% of the total variance of the data was chosen, as suggested by Jolliffe (1986). The number of chosen PCs was ten for "*phys large*", "*phys*", and "*quest*". In the case of "*phys+quest*", PCs were calculated separately for physical and questionnaire variables yielding ten PCs for both subsets, and thereby altogether twenty PCs for "*phys+quest*".

One-class support vector machine

Because we only had 14 academy players available for our analysis, training a classifier with supervised methods can be highly sensitive to overfitting. For this reason, we trained one-class support vector machine (one-class SVM) (Chandola, Banerjee, & Kumar, 2009) to model the normal region of the data based only on the observations from the majority class, i.e., the non-academy players. The trained model can then be used to predict whether new observations belong to this normal region or not.

The primal problem of one-class SVM is (Chang & Lin, 2011):

$$\min_{\boldsymbol{w},\xi,p} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - p + \frac{1}{\upsilon N}\sum_{i=1}^{N}\xi_i \tag{1}$$

$$s.t.\begin{cases}\boldsymbol{w}^T\phi(\boldsymbol{x}_i) \geq p - \xi_i \\ \xi_i \geq 0,\end{cases} \tag{2}$$

where ϕ is a feature map that transforms data point $\boldsymbol{x}_\mathrm{i}$ into higher-dimensional space, $\boldsymbol{w}$ is a weight vector and p an offset parameterizing the region. ξ_is are slack variables, N is the number of observations, and υ is an upper bound on the fraction of training errors. More detailed description can be found in (Schölkopf, Platt, Shawe-Taylor, Smola, & Williamson, 2001). We evaluated the performance of one-class SVM (Python's scikit-learn library, version 0.20.0) using 16 different combinations of hyperparameters γ (0.1, 0.2, 0.3, and 0.4) and υ (0.05, 0.1, 0.2, and 0.4). Radial basis function (RBF) has been found to work best with one-class SVM (Bounsiar & Madden, 2014) and was chosen here as the kernel:

$$K\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = e^{-\gamma||\boldsymbol{x}_i - \boldsymbol{x}_j||^2}, \tag{3}$$

where γ is the kernel coefficient. To verify the need of non-linearity, a baseline classifier, where the RBF-kernel was replaced with the linear kernel, was trained.

Performance evaluation

The performance of the different one-class SVM models were assessed with 10-fold cross validation. The majority class was first divided into ten folds, one for testing and nine for training, and then the players of the minority class were added to each test fold. The learning process is unsupervised, because the information from the minority class is used only for performance evaluation and not for classifier training. Preprocessing, including normalization, knn-imputation and PCA, was performed separately in each fold, first for training data and then the obtained parameters were applied to the test cases before predicting the classes.

As performance metrics, we used the mean area under the receiver operating characteristic curve

(AUC-ROC) and the mean area under the precision recall curve (AUC-PR). The mean value of the metrics were calculated across all the ten folds. Although AUC-ROC is a widely used performance measure in machine learning (Narasimhan & Agarwal, 2013), AUC-PR might be a better option for highly imbalanced datasets (Davis & Goadrich, 2006). For both AUC measures the ideal classifier would yield the maximum score of one. The AUC-ROC score of a completely random classifier is 0.5, whereas the baseline level of AUC-PR depends on the class ratio in the data. The per fold AUC-PR baseline values are 0.869 and 0.818 for ”*phys+large*” and the three other representations, respectively. Mean AUC-ROC values were compared with Kruskal Wallis test and in case of differences, Tukey’s post-hoc tests were performed. Limit of statistical significance was set to p=0.05 and Bonferroni corrected. Based on Kolmogorov-Smirnov test, the values were not normally distributed.

In addition, mean sensitivity and specificity values across the test folds were also calculated using the default decision threshold. Sensitivity measures the proportion of correctly detected academy players, and specificity measures the proportion of correctly detected non-academy players.

Results

One-class SVM results

In Table 4, one-class SVM results in the talent identification task are summarized. The highest mean AUC-ROC value over the tested hyperparameters was 0.763 for data representation *“phys large”*. For representations ”*phys*”, ”*phys+quest*”, and ”*quest*”, the mean AUC-ROC values over the tested hyperparameters were 0.665, 0.643, and 0.585, respectively.

Table 4: Talent identification results for the proposed one-class SVM classifier using the four different data representations. The mean values over hyperparameter combinations and the cross-validation folds are reported for each performance measure (AUC-ROC, AUC-PR, sensitivity, specificity).

	“phys large”	***“phys”***	***“phys+quest”***	***“quest”***
Mean AUC-ROC	0.763(±0.007)	0.665(±0.016)	0.643(±0.013)	0.585(±0.062)
Mean AUC-PR	0.960(±0.002)	0.913(±0.009)	0.880(±0.003)	0.313(±0.194)
Mean sensitivity	0.795(±0.184)	0.732(±0.226)	0.838(±0.120)	0.313(±0.194)
Mean specificity	0.614(±0.142)	0.520(±0.176)	0.355(±0.235)	0.789(±0.125)

Differences in AUC-ROC values were significant between all data representation ($p<0.001$), except between ”*phys*” and ”*phys+quest*” ($p=0.113$). It can also be observed from the estimated accuracies of ”*phys*” and ”*phys+quest*” models, that the questionnaire variables did not improve the performance of the models. The results obtained with ”*phys*” and ”*phys large*” demonstrate, in line with the expectations, that the estimated classification performance tends to improve along with the number of available observations.

All the AUC-PR values were in line with the above-mentioned results. Note that in the case of AUC-PR, the baseline depends on the class ratio and therefore the results for ”*phys*” and ”*phys large*” are not directly comparable. When the non-linear kernel was replaced with the linear kernel in the one-class SVM classification model, the performance decreased notably for all the data representations. The mean of the AUC-ROC values in this case were: 0.548, 0.496, 0.612, and 0.582, for ”*phys large”*, ”*phys*”, ”*phys+quest*”, and ”*quest*”, respectively.

Discussion

The aim of this study was to investigate whether potential future elite soccer players can be identified from a large group of players using machine learning and data collected by physical and psychological tests in their youth. Application of data-driven approaches to talent identification can be generally considered a cumbersome research problem due to the scarcity of childhood data from elite players. Previous research on talent identification has focused on explanatory methods, i.e., explaining relationships and dependencies between variables without assessing generalization abilities of the fitted models on independent observations (Nieuwenhuis et al., 2002; O'Connor et al., 2016; Woods et al., 2016). In this study we evaluated the predictive ability of the one-class SVM anomaly detection method when trained on four different representations of the soccer player test data set.

The best classification performance (mean AUC-ROC value 0.763) was obtained with the set of variables representing physical tests and the greatest number of players (see Table 4). According to classification proposed by Youngstrom (2013), this result can be considered as "fair". In addition, one might argue that this result is satisfactory considering that the classification model has been fitted in an unsupervised manner using only cases from the category of non-academy players and tested using independent data (using CV) involving players from both categories. Besides, since the number of academy players was limited, one-class SVM hyperparameters γ and υ were not optimized in this study, but the average results over multiple classification models (trained using several combinations of γ and υ) were reported. Once more data for classifier validation becomes available, model selection based on CV can be applied to improve the current results.

While the estimated sensitivity of the "*phys large*" representation in the identification task was nearly 0.80, the estimated specificity of 0.614 shows that yet a large proportion of the players without an academy contract will likely be misclassified into the class of potential academy players. The results prove that there is still a long way to go before talent identification can be made by data-driven machine learning tools independently of human expertise. Realistically, the goal should not be full automation of the selection process, but rather modeling of the talent detection expertise possessed by the best professionals in coaching and player management. These data-driven decision support tools may be able to transfer knowledge and enhance decision making in local and regional development organizations.

Several studies have reported relatively high classification performance measures for various models, but the results can be optimistic from the predictive ability point of view, as their performances have not been tested on independent test observations. A multi-dimensional approach for talent identification among young soccer players with AUC-ROC value of 0.954 was presented by Woods et al. (2016). In O'Connor et al. (2016), 93.7% of young soccer players were correctly classified based on selection or nonselection for a full-time elite player scholarship. Nieuwenhuis et al. (2002) reported accuracy of 90.5% when young female field hockey players were classified as successful or less successful. A web-oriented expert system for talent identification in soccer was developed by Louzada, Maiorano, and Ara (2016). It applies principal component and factor analysis to compute general scores for the players in real time. However, without estimation of the generalization ability on independent test observations, the results are not directly comparable to ones presented in this study.

We also studied the relevance of two different types of tests for measuring players physical and psychological abilities by constructing four different representations of the data. The largest data representation, "*phys large*", produced the greatest overall classification performance. The highest sensitivity was achieved with the most versatile set of variables "*phys+quest*". The

representation ”*quest*” achieved the highest specifity, but the sensitivity was low. While some promising players may not become detected due to the low sensitivity, the higher specificity will lead to a lower number of false positives. This enables detection of a smaller group of players with potentially higher chances to succeed. In small countries, such as Finland, elite soccer players are a rarity and higher sensitivity should probably be preferred in order to prevent loss of unrecognized talents. Downsizing of the training group can be completed by coaches when necessary, and thereby ensure that all the talents receive special attention from their training organizations.

It should be noted that without a doubt some of the players that were assigned in the minority class by the SVM model can be potential future elite players, but they have been still too young for signing a contract at the time of this study. In addition, even if a player shows exceptional potential at the age of 14, numerous factors, such as maturity, injuries, coaching/scouting, or decision about whether to stay in Finland to finish school, affects her/his future as an elite soccer player. This is a limitation of this study, and must be taken into account when interpreting the results.

Another limitation of this study is the high number of missing values. Our results indicate that the performance of the model can be improved when more players will be available in both minority and majority classes. Thus, in the future, we can expect improvements due to the continuous accumulation of the data. In addition, some of the included players may sign a contract with a soccer academy after this study, which will enable further improvement to the current models. Moreover, increase in the size of the minority class would enable more thorough validation of the one-class model, or even use of supervised machine learning methods. Also, with larger data it can be possible to utilize different classifications as well, for example look at whether the player made it into the national team.

Furthermore, many of the relevant observations were incomplete. For instance, the self-assessment questionnaire measuring the player’s motivation could improve the classification performance, but in the present study these variables had to be discarded due to missing values. These missing values were caused by refinement and changes in the questionnaires and tests over the years as well as the fact not all questions were compulsory to answer. These issues have been considered and in the future, more complete data will be attained. Also, with more research, the most relevant features can be detected to improve the model. In the long-term perspective, it might become possible to include even more complex data types, such as player tracking or video data, to the machine learning process.

In this study, the parameters of the missing value imputation and dimension reduction methods were fixed based on the existing literature. However, data-based optimization of the parameters might improve the performance of the models.

Conclusion

Identifying talented athletes at young age is an interesting but difficult problem to be successfully solved by machine learning. Accurate identification may, however, enable better career development and level of performance for talented players. In this study, an unsupervised anomaly detection method, one-class SVM, was used to detect potential elite soccer players based on their test data in youth. The best results (mean AUC-ROC 0.763) were achieved when the largest dataset including physical test measurements was used. Considering the size and quality of the available data the present results are promising, but not yet able to provide practical tools to the field. The results also suggest that non-linear methods might be more efficient in the talent identification task than linear ones. Follow-up studies should focus on repeating the study with larger number of players and a more versatile set of variables.

Acknowledgment

This work has been carried out in two projects "Value from health data with cognitive computing" and "Watson Health Cloud", funded by Business Finland. Jukka-Pekka Kauppi was funded by the Academy of Finland Postdoctoral Researcher program (Research Council for Natural Sciences and Engineering; grant number 286019).

References

Aoki, R., Assuncao, R. M., & de Melo, P. O. S. (2017). Luck is hard to beat: The difficulty of sports prediction. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1367–1376).

Beretta, L., & Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC Medical Informatics and Decision Making*, *16*(3), 197–208.

Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). Springer.

Bounsiar, A., & Madden, M. G. (2014). Kernels for one-class support vector machines. In *2014 International Conference on Information Science & Applications (ICISA)* (pp. 1–4).

Brefeld, U., & Zimmermann, A. (2017). Guest editorial: Special issue on sports analytics. *Data Mining and Knowledge Discovery*, *31*(6), 1577–1579.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, *41*(3), 15.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27.

Chawla, N. V, Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, *6*(1), 1–6.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Routledge.

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning* (pp. 233–240).

Forsman, H., Gråstén, A., Blomqvist, M., Davids, K., Liukkonen, J., & Konttinen, N. (2016). Development and validation of the perceived game-specific soccer competence scale. *Journal of Sports Sciences*, *34*(14), 1319–1327.

Goldstein, M., & Uchida, S. (2016). A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS One*, *11*(4), 1–31.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

Jolliffe, I. T. (1986). *Principal component analysis* (1st ed.). Springer.

Knobbe, A., Orie, J., Hofman, N., van der Burgh, B., & Cachucho, R. (2017). Sports analytics for professional speed skating. *Data Mining and Knowledge Discovery*, *31*(6), 1872–1902.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). John Wiley & Sons.

Louzada, F., Maiorano, A. C., & Ara, A. (2016). iSports: A web-oriented expert system for talent identification in soccer. *Expert Systems with Applications*, *44*, 400–412.

Macnamara, Á., & Collins, D. (2011). Development and initial validation of the psychological characteristics of developing excellence questionnaire. *Journal of Sports Sciences*, *29*(12), 1273–1286.

Merigó, J. M., & Gil-Lafuente, A. M. (2011). Decision-making in sport management based on the OWA operator. *Expert Systems with Applications*, *38*(8), 10408–10413.

Narasimhan, H., & Agarwal, S. (2013). A structural SVM based approach for optimizing partial AUC. In *Proceedings of the 30th International Conference on Machine Learning* (pp. 516–524).

Nieuwenhuis, C. F., Spamer, E. J., & Rossum, J. H. A. van. (2002). Prediction function for identifying talent in 14-to 15-year-old female field hockey players. *High Ability Studies*, *13*(1), 21–33.

O'Connor, D., Larkin, P., & Mark Williams, A. (2016). Talent identification and selection in elite youth football: An Australian context. *European Journal of Sport Science*, *16*(7), 837–844.

Ofoghi, B., Zeleznikow, J., MacMahon, C., & Raab, M. (2013). Data mining in elite sports: a review and a framework. *Measurement in Physical Education and Exercise Science*, *17*(3), 171–186.

Papić, V., Rogulj, N., & Pleština, V. (2009). Identification of sport talents using a web-oriented expert system with a fuzzy module. *Expert Systems with Applications*, *36*(5), 8830–8838.

Power, P., Ruiz, H., Wei, X., & Lucey, P. (2017). Not All Passes Are Created Equal: Objectively Measuring the Risk and Reward of Passes in Soccer from Tracking Data. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1605–1613).

Reilly, T., Williams, A. M., Nevill, A., & Franks, A. (2000). A multidisciplinary approach to talent identification in soccer. *Journal of Sports Sciences*, *18*(9), 695–702.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural Computation*, *13*(7), 1443–1471.

Silver, N. (2003). Introducing PECOTA. *Baseball Prospectus*, *2003*, 507–514.

Smith, L., Lipscomb, B., & Simkins, A. (2007). Data mining in sports: Predicting cy young

award winners. *Journal of Computing Sciences in Colleges*, *22*(4), 115–121.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., … Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics*, *17*(6), 520–525.

Williams, A. M., & Reilly, T. (2000). Talent identification and development in soccer. *Journal of Sport Science*, *18*(9), 657–667.

Woods, C. T., Raynor, A. J., Bruce, L., McDonald, Z., & Robertson, S. (2016). The application of a multi-dimensional assessment approach to talent identification in Australian football. *Journal of Sports Sciences*, *34*(14), 1340–1345.

Youngstrom, E. A. (2013). A primer on receiver operating characteristic analysis and diagnostic efficiency statistics for pediatric psychology: we are ready to ROC. *Journal of Pediatric Psychology*, *39*(2), 204–221.