Prediction of Euroleague Games based on Supervised Classification Algorithm *k*-Nearest Neighbours

Tomislav Horvat[1], Josip Job[2] and Vladimir Medved[3]
[1]*Department of Electrical Engineering, University North, 104. brigade 3, Varaždin, Croatia*
[2]*Chair of Visual Computing, Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek, Osijek, Croatia*
[3]*Department of General and Applied Kinesiology, Faculty of Kinesiology, University of Zagreb, Zagreb, Croatia*

Keywords: Basketball, Classification, Database Analysis, Euroleague, Information System, *k*-Nearest Neighbours, Machine Learning, Prediction.

Abstract: Machine learning becomes one of the top fields in world of ICT (Information and communications technology). In the last few decades machine learning has taken a big boost in sports. Sport fans are using machine learning in sport betting and sport managers are using it for player selection, performance evaluation and even in outcome prediction. This paper gives basketball outcome prediction algorithms using *k*-nearest neighbours (*k*-nn). Few methods of preparing data calculated with different *k* and time period for *k*-nn algorithm will be presented and closely explained. Feature selection will also be discussed and results based on feature selection will be presented.

1 INTRODUCTION

The main goal of a machine learning algorithm is to predict an outcome of the observed process. Machine learning is a field of computer science closely related to computational statistics with the background in probability theory, linear algebra, information theory and cognitive sciences. The definition of machine learning comprises that goal of machine learning is to build a model that can serve as a good and useful approximation of data (Alpydin, 2010). The area of machine learning can be divided into supervised, unsupervised and reinforcement learning (Jacobs, 2017). The main difference between supervised and unsupervised learning lies in the fact that in supervised learning target variable should be specified. Based on the type of target variable, supervised learning can be divided into classification and regression. In classification, the target variable is discrete value while in regression target variable can take infinitive number of numeric values (Harrington, 2012). On the other hand, in unsupervised learning, there's no label or target value given for the data. The main goal of unsupervised learning is to find regularity in given data. Sports prediction is usually treated as a classification problem by which one class is predicted (victory or defeat), and rare cases are predicted by numerical values. In this paper data with a known outcome will be used, so problem will be classified as a supervised classification problem.

In this paper, prediction model based on *k*-nearest neighbours (later *k*-nn), one of supervised classification method, will be presented. *K*-nn method is easy to grasp and thus very effective, usually gives high prediction accuracy, it is insensitive to outliers and with no assumptions on data, but also computationally very expensive and requires a lot of memory (Harrington, 2012). Because of its simplicity, *k*-nn method gives a researcher many options and opportunities. Thus, two algorithms for data preparation, with and without feature selection, will be presented in this paper, as well as various possibilities for defining coefficient *k* and different time periods. Feature selection methods are used to achieve dimensionality reduction and exploration of model simplification and acceleration execution algorithms.

For the purposes of research, an information system called Basketball Coach Assistant (later BCA) was built. BCA is a web application based on database, built with PHP and MySQL technologies. The application is supported by JavaScript and

Horvat, T., Job, J. and Medved, V.
Prediction of Euroleague Games based on Supervised Classification Algorithm *k*-Nearest Neighbours.
DOI: 10.5220/0006893502030207
In *Proceedings of the 6th International Congress on Sport Sciences Research and Technology Support (icSPORTS 2018)*, pages 203-207
ISBN: 978-989-758-325-4

jQuery on the client side. The first version of the BCA information system, called *AssistantCoach*, was presented at the international conference icSports 2015 in Lisbon (Horvat et al., 2015).

The second chapter will give an insight into previous research primarily intended for basketball, not necessarily based only on predicting game outcome. The data preparation and methods used for predicting the outcomes will be discussed in the third chapter. In the chapter Results, the results of the research will be presented and explained more closely. In the last chapter the conclusions are given as well as the guidance for future research.

2 RELATED RESERCH

This chapter will provide some basic information on current research and use of machine learning in basketball. Some interesting conclusions and actual results of prediction outcome from prior research will be presented. Sports prediction has gained in popularity because of availability of various datasets, statistics, tools and applications. Different researchers use various machine learning methods to predict the outcomes. Most research in basketball is based on the NBA league, while this research focuses on the strongest European basketball league called Euroleague.

Methods for basketball games prediction started to develop in the second half of the 20th century. According to research from 1979, if team scores are 0.7 points per possession, the one will win the game (Shanahan, 1979). In 1997, the application named *Advanced Scout* was introduced with the goal of revealing interesting patterns in NBA game data (Bhandari et al., 1997). At the symposium in Budapest, a study was presented where authors were using neural networks for analysing five seasons of the second Serbian league and concluded that the most influence on the final outcome have defensive rebounds, then percentage for two point shots, percentage for three point shots, stolen balls, lost balls, offensive rebounds, percentage of free throws, blocks and the slightest influence have assists (Ivanković et al., 2010).

One of the best results in an outcome prediction is shown in (Loeffelholz et al., 2009) where a prediction model of the basketball games outcome is presented in the NBA league using various types of neural networks. The research also included methods for feature selection. Compared to the experts' predictions whose accuracy was 68.67%, the best modelled neural network gave an accuracy of 74.33%. In the paper (Cheng et al., 2016), the authors presented a classification model with accuracy of 74.4% for prediction of NBA league games by using Maximum Entropy.

From other results it is worth mentioning a paper (Amorim Torres, 2013.) in which various machine learning methods were used. The best prediction rate was achieved using the Multilayer Perceptron Method that achieved 68.44% prediction rate. The Linear Regression has achieved a performance of 67.89% which was better than the likelihood method that achieved a performance of 66.81%. In paper (Miljković et al., 2010) Naive Bayes method is used. Besides actual outcome prediction, for each game, system calculates the spread, by using multivariate linear regression. The system was evaluated on the dataset of 778 games and it correctly predicted the winners of about 67% of matches. As for the prediction of spread, the system has successfully provided predictions for 78 matches (10%).

3 DATA AND METHODS

Sufficient amount of relevant data is a basic condition for building a good prediction model. It is very important to well define the methods that will be used. Two different data preparation methods and two ways of using the *k*-nn algorithm with different *k* coefficient will be explained in detail in this chapter.

3.1 Data Acquisition

For the purposes of research, public statistics of Euroleague seasons 2012/2013 to 2016/2017 were used. 80% of games are used for model training and 20% for result evaluation, regardless of the time period. Using web scraping process, structured data from Euroleague official Web were extracted, transformed and loaded into relational database suitable for further analysis. Process of **e**xtracting, **t**ransforming and **l**oading is called ETL process. Due to specificity of data retrieval, Web scraper in scripting programming language PHP has been programmed. The web scraper passes through the domain *www.euroleague.net*, extracts data from page, transforms them if necessary and stores them into relational database. As was previously stated, sufficient amount of relevant data is a basic condition for building a good prediction model and that is exactly the task of the web scraper.

3.2 The *k*-nn Method

As mentioned above, the *k*-nn method is selected because it is easy to grasp and thus very effective, but also computationally very expensive. This is a non-parametric method used for classification and regression. The method fundamentally relies on a metric distance value. The most common metric, used in this research also, is Euclidian distance, but there are other metrics that can be used. Euclidian distance is shown in equation (1).

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i}^{n} (x_i - y_i)^2} \quad (1)$$

In this paper, two data preparation algorithms, both Euclidian distance based, will be compared and tested for different *k* values as well as for different dataset lengths, i.e. time periods. The main goal is to find the optimum *k* and the time period for predicting games outcome.

3.3 Data Preparation Variants

As mentioned above, two variants of data preparation for the *k*-nn method were used. All variants of data preparing are based on basic basketball statistics which are shown in Table 1. Due to the specificity of each team, the outcome predication is based only on the observed team games.

Table 1: Basic basketball statistics elements.

Name/Abbrev.	Full name / Explanation
2fgm, 2fga	2 field goals made / attempts
3fgm, 3fga	3 field goals made / attempts
ftm, fta	free throws made / attempts
dr, or	defensive / offensive rebounds
as, st, to	assists, steals (stolen balls), turnovers
bl, bl_against	blocks made by player / against player
f, f_drawn	fouls committed / drawn
madeFG_FT	made field goals and free throws
missFG_FT	missed field goals and free throws

First variant is called *DefenseOfense*. The goal of that variant is to group defensive and attacking elements, thus separating the defensive from the attacking part. Two components, I_{ofense} and $I_{defense}$, were calculated and later used in the outcome prediction. Equations (2) and (3) show how the components are calculated. Index *r* (rival) refers to opponent / rival team.

$$I_{ofense} = madeFG_{FT} - missFG_{FT} + or + as + f_{drawn} - to - bl_{against} \quad (2)$$

$$I_{defense} = -madeFG_{FT_r} + missFG_{FT_r} + dr + st + bl - or_r - f \quad (3)$$

Euclidian distance based on every game in test period and game in evaluation period is calculated and then, based on user defined parameter *k* predicted outcome is calculated. Parameter *k* is always defined as odd number to avoid the inability of defining game outcome. The same principle of calculation is used in another data preparation variant, but unlike in *DefenseOfense* more elements are used.

The purpose of the second variant, called *Components*. is to group the logical elements of the game (defensive and offensive rebounds, stolen balls and turnovers…) and thus obtain a multidimensional problem. List of logical groups and their calculation is shown in Table 2.

Table 2: Logical components groups.

Name	Explanation
points	*madeFG_FT – missFG_FT*
rebounds	*dr + or*
steals_turnovers	*st - to*
assists	*as*
blocks	*bl – bl_against*
fouls	*f – f_drawn*

Both algorithms (*DefenseOfense* and *Components*) were also compared after applying the feature selection process. Dimensionality reduction (van der Maaten et al., 2009) or in computing feature selection (Jović et al., 2015) is a very important step in estimating output value because significantly determines the algorithm predictability. The selection process can be applied in several ways depending on the goal, the available resources, and the desired level of optimization. In this research, the feature selection is performed by measuring the information gain for each feature (Hoque et al., 2014). Information gain is a measure of an amount of information whose specific feature contributes to

the class. In this case specifically, it is a measure that shows how much each element of basketball statistics (Table 1) contributes to the final game outcome. Feature selection will be defined based on the average information gain and on *zero based* information gain. If the information gain of statistical component is greater or equal to the average value of the information gain, component will be used in the outcome prediction process. *Zero based* information gain means that features will be included into prediction model only if their information gain is greater than zero. Final feature subset which will be used in the testing phase of the algorithm will be generated by a union of feature subsets defined for the observed and rival team.

4 RESULTS

The research results, based on two different data preparation variants, each of them defined with different coefficients *k*, different number of analysed seasons and different subset of features are presented in Table 3.

Table 3: Prediction algorithm results (DefenseOfense).

Feature selection	Seasons	Accuracy (%)			
		k=3	*k=5*	*k=7*	*k=9*
	1	80,19	83,96	83,96	83,02
	3	83,39	79,55	80,51	80,83
	5	80,66	79,88	78,53	79,11
inf. gain > 0	1	80,19	83,96	83,96	83,02
	3	83,39	79,55	80,51	80,83
	5	80,66	79,88	78,53	79,11
inf. gain > average value	1	75,47	72,64	71,7	74,53
	3	72,2	70,29	72,2	70,61
	5	74,27	73,89	75,63	75,63

Table 4: Prediction algorithm results (Components).

Feature selection	Seasons	Accuracy (%)			
		k=3	*k=5*	*k=7*	*k=9*
	1	70,75	71,7	73,58	68,87
	3	76,36	77,32	75,4	74,44
	5	71,95	73,11	73,11	76,6
inf. gain > 0	1	70,75	71,7	73,58	68,87
	3	76,68	77,32	75,4	74,44
	5	71,57	72,92	73,11	76,6
inf. gain > average value	1	61,32	61,32	61,32	57,55
	3	57,51	61,02	59,11	61,34
	5	60,93	62,67	60,93	62,09

As can be seen from result in Table 3 and Table 4 better results are achieved with using *DefenseOfense* data preparation variant. It is noticeable that coefficient *k* does not have significant impact on model accuracy while feature selection process and grouping of input features has. Much more impact on model accuracy has the feature selection and defining of logical subgroups of input features. Both algorithm variants use a set of 15 features. The first variant, *DefenseOfense*, divides features into two logical subgroups, while the second variant, called *Components*, divides input features into six logical subgroups. From all this, it is easy to notice that dividing no excessive large input set of features on even smaller subgroups does not yield better results (speaking about *k*-nn classification method). This is an example where all the features are significant and each of them contributes to the model accuracy. Excluding features with information gain equal to 0 does not produce better results, but it reduces the algorithm execution time. Feature selection variant that excludes all features with information gain less than the average information gain for observed team gave worse results. The conclusion is that there are still too few features that would lead to a good description of the process (in this case process is a basketball game). The problem of lower model accuracy, in case where features are selected based on the average information gain, occurs in both cases, but still gives good results. Better results of the prediction would certainly be provided with a greater number of input features. In that case, feature selection would get greater significance and would also accelerate algorithm execution speed.

Compared with prior research, the proposed data preparation variant *DefenseOfense* gives improved results. Most of the prior research is focused on predicting NBA league outcomes, and typical dataset is consisted of 5 seasons, but additional difference between two leagues is the fact that the NBA league has as twice as many games per season than in Euroleague. Despite the fact that in this calculation dataset with less records is used, this approach achieved better results, but the method should be tested on the same dataset before drawing conclusions.

5 CONCLUSIONS

In this paper, traditional *k*-nn algorithm on game prediction is presented. Based on the analysis, two data preparation methods have been proposed.

Sport predictions present challenging and interesting field because there is no universal algorithm that solves prediction problem for every sport. In addition to expert knowledge related to ICT technology, there is also a need for expert knowledge related to the analyzed sport.

According to the presented results, it can be concluded that the *k*-nn is a very convenient method for solving outcome prediction problems in sport. The combination of *k*-nn classification method and feature selection is also an interesting field of research. Various approaches have been tried up to this research. An approach that might give better results is to define the feature weight based on the information gain when calculating Euclidean distance (Sun and Huan, 2010), with or without feature selection. A possible formula for such research to calculate Euclidian distance based on feature information gain is:

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i}^{n} w_i (x_i - y_i)^2} \qquad (4)$$

Although there is no universal algorithm for predicting game outcomes, there is no reason not to try to find the sport specific one, hence in addition to knowledge of ICT technologies and machine learning research, advanced knowledge of the observed process, in this case basketball, is also required.

REFERENCES

Alpydin, E., 2010. *Introduction to Machine Learning: Second Edition*, Cambridge, Massachusetts, London, England.

Amorim Torres, R., 2013. *Prediction of NBA games based on Machine Learning Methods*, University of Wisconsin Madison.

Bhandari, I., Colet E., Parker, J., Pines, Z., Pratap, R., Ramanujam, K., 1997. Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. *Data Mining and Knowledge Discovery, 1, 121-125*.

Cheng, G., Zhang, Z., Kyebambe, M.N., Kimbugwe, N., 2016. Predicting the Outcome of NBA Playoffs Based on the Maximum Entropy Principle. *Entropy, 18 (12), 1-15.*

Harrington, P., 2012. *Machine Learning in Action*, Manning Publications, Shelter Island.

Hoque, N., Bhattacharyya, D.K., and Kalita, J.K., 2014. MIFS-ND: A mutual information-based feature selection method, *Expert Systems with Applications, vol. 41, issue 14, pp. 6371–6385.*

Horvat, T., Havaš, L., Medved, V., 2015. Web Application for Support in Basketball Game Analysis. *icSports 2015, Lisboa, 225-231.*

Ivanković, Z., Racković, M., Markoski, B., Radosav, D., Ivković, M., 2010. Analysis of basketball games using neural networks. *Proceedings of the 11th International Symposium on Computational Intelligence and Informatics, Budapest, 251-256.*

Jacobs, S., 2017. *The Ultimate Guide to Machine Learning (Neural Networks, Random Forests and Decision Trees)*, Amazon Distribution.

Jović, A., Brkić, K., Bogunović, N., 2015. A review of feature selection methods with applications. *Information and Communication Technology, Electronics and Microelectronics (MIPRO).*

Loeffelholz, B., Bednar, E., Bauer, K.W., 2009. Predicting NBA Games Using Neural Networks, *Journal of Quantitative Analysis in Sport*, 5(1), article 7.

Miljković D., Gajić Lj., Kovačević A., Konjović Z., 2010. The use of data mining for basketball matches outcomes prediction. *IEEE 8th International Symposium on Intelligent Systems and Informatics, Subotica, Serbia*.

Shanahan, K.J., 1979. *An analysis of game statistics in basketball*, Masters thesis. Faculty George Williams.

Sun, S., Huan, R., 2010. An Adaptive k-Nearest Neighbor Algorithm. *7th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010).*

Van der Maaten, L., Postma, E., van den Herik, J., 2009. *Dimensionality Reduction: A Comparative Review*, Tilburg Centre for Creative Technology, Tilburg University.