MINISTERUL EDUCAȚIEI NAȚIONALE
ROMÂNIA
UNIVERSITATEA DE MEDICINĂ, FARMACIE, ȘTIINȚE ȘI TEHNOLOGIE „GEORGE EMIL PALADE" DIN TÂRGU MUREȘ

Acta Marisiensis. SeriaTechnologica
Vol. 18 (XXXV) no. 2, 2021
ISSN 2668-4217, ISSN-L 2668-4217

sciendo

10.2478/amset-2021-0015

CYCLING ACTIVITY DATASET CREATION AND APPLICATION FOR FEEDBACK GIVING

Dorina K. FERENCSIK[1] and **Erika B. VARGA**[2]

[1,2]*Institute of Information Science, University of Miskolc*
3515 Miskolc-Egyetemváros, HUNGARY

[1]dorina.ferencsik@gmail.com
[2]vargaerika@iit.uni-miskolc.hu

Abstract

Our research aims at supporting personal cycling trainer applications in training planning and feedback giving to nonprofessional outdoor cyclists, based on a general reference. In this paper we present the created dataset. According to our present knowledge, this data collection is the first public dataset containing cycling activities recorded outdoor. Its usability for training planning and feedback giving is demonstrated through an example. The dataset is clustered according to age groups, considering distance and average speed as the two most influential features when predicting the time required for training. These clusters are then applied as references in feedback giving and goal setting.

Key words: Outdoor cycling dataset, Sports data mining, Cycling performance feedback, Personal cycling trainer

1. Introduction

Goal setting is a mental tool one can use to maintain a high level of motivation in any activity. In effective goal setting SMART goals are needed. Although historically the philosophy behind goal setting dates back to the Greeks, the idea of creating SMART goals was formalized later. Doran [8] was the first to introduce the Specific, Measurable, Assignable, Realistic and Time-related (SMART) method for writing effective management goals. Over time the SMART acronym experienced changes and a variety of other words were used as more people saw the benefits of the concept in business and education. SMART goals became SMARTER goals where Exciting and Recorded are the additional characteristics.

The application of SMART goals in sport activities was studied by Locke and Latham [15]. Their general hypothesis was that goal setting works as well in sports as in business and laboratory tasks, and they set ten specific hypotheses to be tested in competitive sports. They also state that *goal setting in the absence of informative feedback becomes meaningless.*

A similar statement can be derived from the behavioral sports psychology literature (see Horn [11]) for a review), that also emphasizes the importance of feedback. In book of Ward [25], three primary principles of goal setting are identified. 1) Goals should be specific and difficult, 2) goal statements should define the consequences of meeting or not achieving the goal, and 3) *goal setting is more effective when combined with performance feedback.* Performance feedback can be given by a person (self, or other, e.g. coach or teammate), or technology.

Technology have become increasingly useful in providing feedback in sports settings and in fitness and health, see Sykora et al. [22]. Functionally, feedback types can be divided into three main groups: 1) monitoring actual activity, 2) comparing actual performance with personal records, and 3) showing actual performance with respect to a general reference.

Whilst gadgets providing first-type feedback are simple electronic measuring devices, for second- and third-type feedback an information system is needed that collects, filters, processes, and displays the data coming from the measuring units, see Chi [6] and Baca et al. [3]. Tracking devices, like sports watches with

built-in GPS, or mobile devices running tracking apps, produce vast amount of training data. They usually track and monitor performance in real time. This means that for example mobile applications dedicated to running, cycling or hiking activities record information regarding the duration time, length and altitude of the course in parallel with the athlete's moves. Examples here are Strava and Garmin Connect, which are web-based training services, offering their users sports activity data visualization and analysis. Strava stores training data in a file in GPS exchange format (GPX), which is a light-weight XML data format for the interchange of GPS data (waypoints, routes, and tracks) between applications and Web services on the Internet (see: `http://www.topografix.com/gpx.asp`). Garmin introduced the TCX (Training Center XML) dataset format, which is similar to the GPX format (see: `http://www8.garmin.com/xmlschemas/TrainingCenterDatabasev2.xsd`). It exchanges also GPS tracks, but treats a track as an activity rather than simply a series of GPS points. TCX includes additional data with each track point: heart rate, running cadence/bicycle cadence, calories, as well as summary data in the form of laps. The movement of the athlete is visualized using Mapbox in Strava, and on Google Maps in Garmin Connect, together with some statistical measures like average heart rate during sports session, velocity, or length of the performed course. This analysis, however, represents only the actual sports session.

Combination of technology with data mining methods opens new perspectives. First attempts include activity identification Lee and Hoff [13], analysis of the outcomes of sport events by De Marchi [7], and sport result prediction by Min et.al. [18], Baulch [5], McCabe and Trevathan [16] and Miljkovic et al [17]. Lately, relevant second-type feedback presenting systems have been created with the purpose of achieving better performance and avoiding excessive load in professional sport settings by Baca [1] or Jaitner and Trapp [12]. The Mobile Motion Advisor introduced by Baca et al. [4], by applying the remote coaching concept, supports athletes and coaches in different sports in continuously supervising individual performance level and giving real-time feedback by Baca [2]. Recent scientific interest leans towards recommending training plans. Fister et al. [10] or Rauter et al. [20] started to build an automated personal trainer that is able to aggregate data from GPX files and on the basis of an analysis using data mining methods, gives second-type feedback and training recommendations for triathlon athletes based on personal records. Their intelligent sports trainer is aimed to be capable of dynamically planning sports sessions according to personal characteristics and external circumstances (e.g. weather) and applies a modified bat algorithm (MBA) Fister et al. [9].

The popular web-based training services, like Strava and Garmin Connect, also apply data mining methods. When recommending route or training plans for runners and cyclists the data come from a global dataset recorded by their users. Second-type feedback is provided when displaying personal progress or fitness level. Third-type feedback is implemented as showing personal results in view of the list of best records of other users. For these kinds of feedback, recorded data need to be stored under personal accounts and to keep privacy, they are managed and used privately.

Large amount of historical data is needed for the purpose of personal cycle trainer applications that produce feedback based on general reference. Since we have not found such a dataset that 1) contains outdoor cycling data in an amount that is sufficient for feedback giving, and 2) is publicly available, we have collected several thousand records of cycling activities from volunteer users of different sports tracking applications. These users exported their own activities and gave us permission to use their data for research, keeping their anonymity. On the basis of this data collection we have created a public dataset that can be used as a reference to compare with one's actual performance. The dataset is available at `kaggle.com` and can be downloaded from `http://www.kaggle.com/dorinaferencsik/outdoor-cycling-metrics` (uploaded on October 4, 2019).

In this paper we introduce the dataset. We present the method of its creation and demonstrate its usability for personal cycling trainer applications through an example.

2. Data acquisition and cleaning

For our research we have collected over ten thousand cycling activities. Among our volunteer data providers there are professional cyclists and commuters having more than 20 activities/month, and hobby cyclists with less than one activity/month. We have collected activity data from all groups so that we can offer a diverse reference base for a wide range of users.

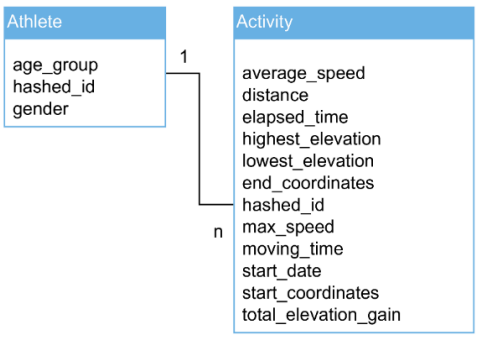


Fig. 1: Structure of the collected dataset

Originally, we have collected activities completed by female and male cyclists as well, because we planned to create separate reference bases for each gender in order to grant more precise and reliable feedback. Unfortunately, there were only about a few hundreds of activities completed by female cyclists –

that were made available for us – which is not enough for data mining. For this reason, the present paper covers only the reference base of male cyclists.

More than 10 thousand cycling activities of male athletes were uploaded to our website. As the collected files consist of only latitude-longitude geminates characterized by a timestamp and an elevation above sea level, first we had to extract useful information.

This procedure involved the exploration of the possibilities of transforming GPS positions into features, and also the selection process of these extracted features. We have extracted several traits from the files and by combining them with the cyclists' characteristics (such as age and gender) we have ended up with a database containing 14 features (see Figure 1).

In the next step, we had to clean the dataset. The discovery and correction of the faulty data involved investigating missing values, dimension reduction and detecting outliers. This process was greatly supported by the Pandas Python package, which can create summaries about a given dataset and can facilitate the whole course of data processing. The `describe()` function was used to create a feature-wise summary table and was the base for identifying the number of missing values, and finding the uninterpretable or impossible values by checking the range of a feature.

In our dataset, we found 2 activity features where over 80% of the values were missing. These are the start and end coordinates (latitude and longitude). The reason for this is that in the process of data acquisition, the users could decide whether they want to share the exact GPS positions of their cycling routes, or not. Missing value pairs mean that the user did not give us permission to include these data in the public dataset. As mentioned earlier, the gender feature came out to be indifferent as well, because of the small number of female users. Hence, the insufficient number of usable values motivated the deletion of these features on the one hand. On the other hand, if we want to give real time feedback on one's actual performance based on a general reference, dimension reduction is inevitable for reducing computation time. Consequently, the final dataset contains 11 cycling activity features: age group, average speed, distance, elapsed time, highest elevation, lowest elevation, hashed ID, maximum speed, moving time, total elevation gain, and local start date.

Table 1: Numerical features of the cleaned dataset

	Mean	STD	Min	0.25	0.50	0.75	Max
average_speed [m/s]	5.76	1.22	2.03	4.99	5.74	6.46	12.48
distance [m]	26120.85	29199.25	161.0	7451.3	13702.2	34599.1	436806.0
elapased_time [s]	5817.39	19632.69	61	1645.0	2998.0	7338.5	1806211.0
highest_elevation [m]	270.73	166.61	-71.2	162.3	229.4	290.5	956.0
lowest_elevation [m]	129.92	47.88	-134.4	103.2	130.2	148.3	809.0
max_speed [m/s]	12.22	3.24	2.3	9.7	11.6	14.4	22.0
moving_time [s]	4509.68	4948.21	61.0	1440.0	2415.0	6099.0	90983.0
elevation_gain [m]	274.47	410.03	0.0	29.8	87.0	359.45	3896.8

The last step was the investigation of outlier values. Outliers significantly differ from other observations. They can represent outstanding performance of professional cyclists or can originate from human error (e.g. one forgets to switch off the tracking device when the activity is stopped) or from measuring failures. Since our dataset is a mixture of two user groups (professional and hobby cyclists), both causes are probable. We have checked these values feature-wise and applied a mixture method to remedy the problem. Records containing outstanding, but correct values were retained, but records having impossible values, such as 441 km/h as maximum speed or 12100 meter as highest elevation above sea level, were deleted. As the result of eliminating outliers and missing values, our dataset was reduced to 9474 sample records.

The summary of the cleaned and prepared dataset's numerical features is shown in Table 1. This table shows the mean, the standard deviation (STD), the minimum and the maximum value of each numerical feature. Furthermore, the first, second and third quartile of the dataset are listed in the corresponding 0.25, 0.50 and 0.75 columns.

The 9474 cycling activity samples in the dataset are categorized by the users' age groups as follows.

- Age group 0: users under 30 years, 914 activities
- Age group 1: users between 30 and 45 years, 7219 activities
- Age group 2: users above 45 years, 1050 activities.

3. Method of dataset application

The collected dataset contains wide range of information about the completed cycling activities without any categorization. If these were classified by, for example the difficulty level, we would be able to use these classes as reference: cyclists could compare and rank their actual performance according to these

difficulty classes. Since the dataset includes several attributes, each contributing somehow to the difficulty level of an activity, it is hard to predefine difficulty classes. Therefore, we used unsupervised clustering methods where similar activity instances are grouped based on the values of their properties. Phyton's `scikit-learn` package offers a number of implemented clustering algorithms. Among them we used and compared the ones that work well with large dataset and small number of clusters in a low dimensional space.

Firstly, the K-means method by Lloyd [14] was applied. It is one of the most well-known clustering algorithms which is based on calculating the Euclidean distance of the samples. The idea behind this method is to create *k* clusters with equal variance, minimizing a criterion known as the inertia or within-cluster sum-of-squares. The K-means method scales well to large datasets [21] but it has some drawbacks as well. Among others, K-means clustering requires all variables to be continuous, it will always try to create convex and isotropic clusters, and it requires a priori specification of the number of clusters. Though this can be done empirically from the data, improper choices can lead to erroneous clusters.

Secondly, we used the Mini-batch K-means clustering algorithm introduced by Ward [23], which is a variant of the original K-means method. This method uses mini batches to decrease the computation time while trying to optimize the same objective function. The mini batches are randomly selected subsets of the dataset and are assigned to the nearest cluster centroid as a unit. The Mini-batch variation usually results in only a slightly worse outcome than the original K-means.

Finally, we have tried the Spectral clustering method, see Ng and Jordan [19], which is not based on distance like K-means, but on connectivity: even if two samples are close to each other, if they are not connected, they will not be in the same cluster. Spectral clustering requires the number of clusters to be specified and examines the pair-wise connectivity of the samples. Two samples are connected if their distance is less than a specified ε constant.

4. Results

The applied clustering methods allow for computing personal, as well as general reference for providing feedback. However, when composing the dataset, we have not limited the data collection to a single user. Our dataset includes diverse activities from different users thereby guaranteeing a sufficient base for general reference. On the other hand, since the ages of the users were available and we could create age groups within the sample, it is worth calculating age-group based references as well. Foreseeably, these clusters will represent one's actual performance better than the general reference.

Before creating the clusters in the dataset for later use in a feedback giving and training planning system, we need to answer the following questions:

1. What are the most significant features affecting the prediction variable?
2. Which clustering algorithm yields the best results and how many clusters are needed to describe the dataset?

In predictive models, the features that contribute most to the prediction variable are chosen with feature selection. We have run this process using Phyton's `feature_selection.SelectKBest` method in the `scikit-learn` package, where the prediction variable was the time required for a training (`moving_time`). As a result, we got 107938.28 for distance and 15655.19 for `total_elevation_gain`, while a few hundred scores for all other features. Beside these two features, we also considered `average_speed`, as moving time can be calculated as distance divided by speed. We found that distance and average speed have strong effect on moving time and the resulting clusters can be more easily explained, than for distance and total elevation gain, so – in order to present a demonstrative example – we applied the clustering models on distance and average speed.

We have experimented with the clustering methods and the number of groups. We found that the Spectral method is not the best to describe our dataset. For the whole dataset it could create only 4 clusters, and when applied to age group 1, it resulted an unconnected cluster which is hard to explain. The Mini-Batch algorithm yielded very similar results to the K-means algorithm, and since the latter method is better known and has several existing implementations, we decided to apply this one to demonstrate the usability of the created dataset. The K-means algorithm requires the specification of the number of clusters in advance. We wanted to characterize a training with high/average/low distance and average speed, so we have tried to run the algorithm with 3, 4, 5 and 6 clusters and found that our dataset is best represented by grouping the data into 5 clusters.

5. Discussion

The clusters created by the K-means algorithm are shown in Figure 2.

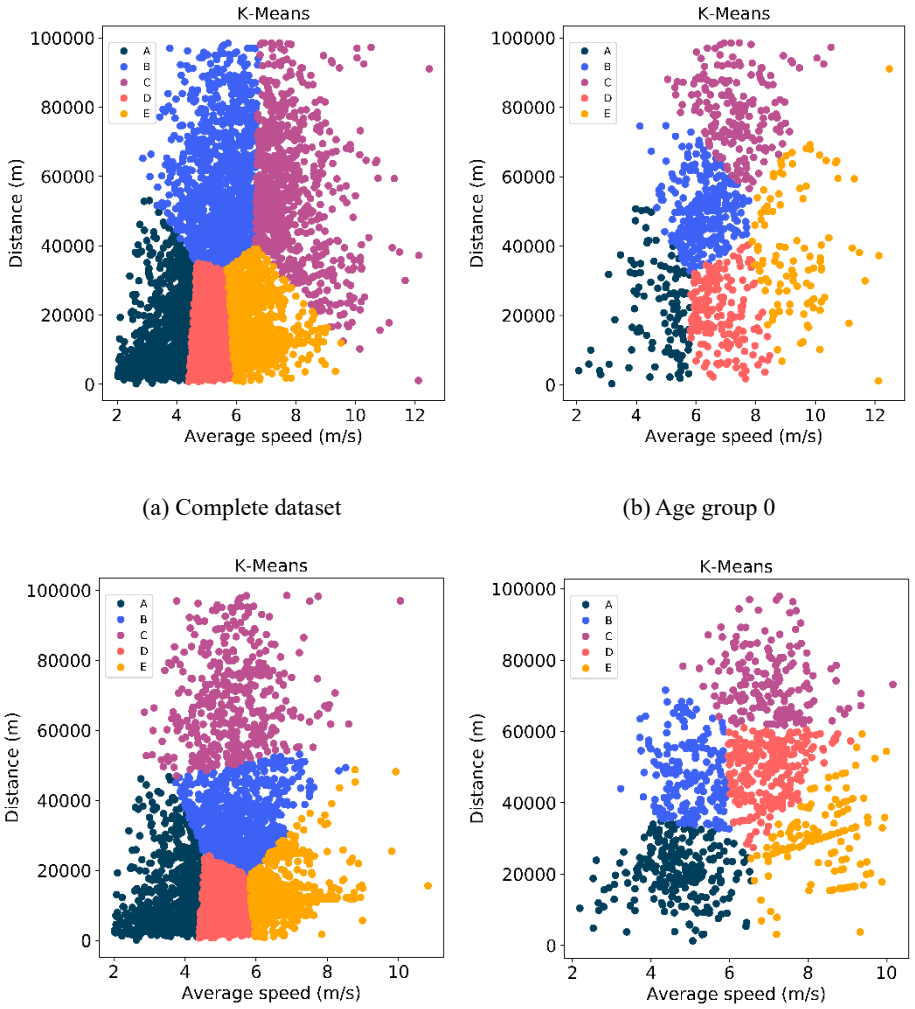
For age group 0, having the smallest sample size (cyclists between 18 and 30 years, 914 samples), the formed clusters can be characterized by:

A. average and short distance – low speed
B. average distance – average speed
C. long distance – average and high speed
D. short distance – average speed
E. high speed

For age group 1 (cyclists between 31 and 45 years, 7219 samples) the created clusters can be interpreted as:

A. average and short distance – low speed

B. average distance – average speed
C. long distance
D. short distance – average speed
E. average and short distance – high speed



(a) Complete dataset

(b) Age group 0

(c) Age group 1

(d) Age group 2

Fig. 2: Clusters created by the K-means algorithm

For age group 2 (cyclists above 46 years, 1050 samples) the formed clusters can be explained as:

A. average and short distance – low speed
B. average distance – low speed
C. long distance
D. average distance – average speed
E. high speed

Analyzing the results, we can conclude that among the clusters computed from the entire dataset the average cycling performance can hardly be identified. This is because the individual activities scatter widely. Dividing the dataset by age groups, the resulting clusters can be more easily explained, and the average performance can be clearly seen. It is also worth mentioning that while most clusters are characterized by the distance and speed attributes, there is at least one cluster for each age group where only one of the attributes determine cluster membership. This phenomenon can be explained by the small number of considered features, but at the same time can be realistic as well regarding the given age group.

The clustering example shows that, with categorizing the cycling activities into different clusters, a useable reference base can be achieved for a feedback giving system. After finishing an activity, cyclists can check their performance's membership in one of the clusters. Thereby the system evaluates their performance and can show the activity's ranking on a percentile diagram.

The same dataset and clustering methods can also be used in a personal cycling trainer application for training planning. After selecting an age group, the user can choose from 3 training samples within each cluster. For example, if a 40-year-old cyclist in age group 1, wants to go for a short, but intensive ride that belongs to cluster E, he can get easy (Q1), average (Q2) or hard (Q3) training recommendations

concerning the selected features, which are distance and average speed in our example. These are calculated as the quartile points of the data in the given cluster. In Table 2 one training example is given for clusters C, D and E for each age group. The estimated time required for the training is calculated from distance and average speed.

The parameters of the program that calculates the training plans are age group, the clustering method to be applied, the label of the target cluster (A-E), and a number between 0 and 1 to determine the desired difficulty level of the training (easy – 0.25, average – 0.5 and hard – 0.75 in the example above).

Table 2: Example training plans based on K-means clustering results

	Distance [m]	**Speed** [m/s]	**Time** [h:m]
	Age group 0		
C – average	77410.3	7.3 (26.3 [km/h])	2:57
D – easy	14982.8	6.3 (22.7 [km/h])	0:40
E – hard	51853.1	9.8 (35.3 [km/h])	1:28
	Age group 1		
C – average	66389.3	5.3 (19.1 [km/h])	3:26
D – easy	7350.9	4.9 (17.6 [km/h])	0:25
E – hard	12489.7	6.8 (24.5 [km/h])	0:31
	Age group 2		
C – average	72773.5	7.0 (25.2 [km/h])	2:53
D – easy	42353.3	6.4 (23.0 [km/h])	1:50
E – hard	33997.5	8.9 (32.0 [km/h])	1:04
	General plan		
C – average	54080.9	7.5 (27.0 [km/h])	1:59
D – easy	7362.7	4.9 (17.6 [km/h])	0:25
E – hard	18097.4	6.8 (24.5 [km/h])	0:45

The design of the proposed system can be seen in Figure 3. The functions of each system component are as follows:

- Data collecting application: providing user interface for collecting reference data.
- Server: performing and storing the clustering of reference data.
- Intelligent software module:
  - checking actual performance's membership in one of the clusters and giving feedback,
  - calculating training plan.
- Training tracking and planning application:
  - collecting actual cycling data,
  - displaying performance ranking and feedback,
  - providing user interface for training planning,
  - displaying training plan.

6. Conclusion

From a sport psychological point of view, the application of both goal setting and feedback are the most powerful incentives. There are many applications that can be used free of charge for tracking a range of data from a ride and visualizing one's weekly and monthly mileage and time spent for cycle training. This kind of feedback, however, is based only on personal records.

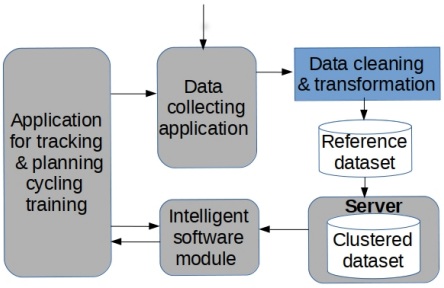


Fig. 3: Data flow between cycling training app, intelligent software module and database server

On the other hand, if someone would like to set specific goals that are difficult − that is one should increase his effort to achieve it − but realistic at the same time, and want to get feedback on his performance, he needs to pay for a professional application. These applications may take the athlete's current fitness level into account and can customize the intensity and volume of the training plan. Some of them are even more flexible: they provide real-time feedback and can adapt the training plan to the current external circumstances. They work with great datasets collected by thousands of registered users.

Our aim was to create a publicly available dataset that can be used in personal cycling trainer applications for goal setting and feedback giving. We

have collected and cleaned 9474 cycling activity records from male users of third-party cycling applications and made it available at `kaggle.com`. We have created the structure of the dataset that can be the basis of applying data mining methods on the acquired data. We have made some experiments with clustering and linear regression algorithms, and in this paper, we presented the results and showed that the dataset is applicable for training planning and feedback giving.

Acknowledgement

The described article was carried out as part of the EFOP- 3.6.1-16-00011 "Younger and Renewing University Innovative Knowledge City institutional development of the University of Miskolc aiming at intelligent specialisation" project implemented in the framework of the Szechenyi 2020 program. The realization of this project is supported by the European Union, co-financed by the European Social Fund.

References

[1] Baca, A. Feedback systems. *WIT Transactions on State of the Art in Science and Engineering,* 32:43-67, 2008. DOI: 10.2495/978-1-84564-064-4/02,

[2] Baca, A. *Social Networks and the Economics of Sports*. chapter Adaptive Systems in Sports. Springer, Cham., 2014. DOI: https://doi.org/10.1007/978-3-319-08440-4_7

[3] Baca, A., Dabnichki, P., Heller, M., and Kornfeind, P. Ubiquitous computing in sports: A review and analysis. *Journal of Sports Sciences*, 27(12):1335-1346, 2009. DOI: 10.1080/02640410903277427

[4] Baca, A., Kornfeind, P., Preuschl, E., Bichler, S., Tampier, M., and Novatchkov, H. A server-based mobile coaching system. *Sensors*, 10(12):10640-10662, 2010. DOI: 10.3390/s101210640

[5] Baulch, M. *Using Machine Learning to Predict the Results of Sporting Matches*. PhD thesis, University of Queensland, 2001. PhD Thesis

[6] Chi, E. Sensors and ubiquitous computing technologies in sports. *WIT Transactions on State of the Art in Science and Engineering, Computers in Sport*, 32:249-268, 2008. DOI: 10.2495/978-1-84564-064-4/09

[7] De Marchi, L. *Data mining of sports performance data.* PhD thesis, University of Leeds, School of Computing Studies, 2011. Erasmus computing 2010/2011.

[8] Doran, GT. There's a s.m.a.r.t. way to write management's goals and objectives. *Management Review*, 70:249-268, 1981

[9] Fister, I, Rauter, S, Yang, Xin-She, Ljubiˇc, K, and Fister, I. Planning the sports training sessions with the bat algorithm. *Neurocomputing*, 149:993-1002, 2015. DOI: https://doi.org/10.1016/j.neucom.2014.07.034

[10] Fister, Jr. I, Fister, I, Fister, D, and Fong, S. Data mining in sporting activities created by sports trackers. In *International Symposium on Computational and Business Intelligence*, pages 88-91, 2013. DOI: 10.1109/ISCBI.2013.25

[11] Horn, TS. *Advances in sport psychology*. Human Kinetics, Champaign, IL., 2009

[12] Jaitner, T. and Trapp, M. Application of service oriented software architectures in sports: Team training optimization in cycling. *International Journal of Computer Science in Sport*, 7(2):34-45, 2008

[13] Lee, JY. and Hoff, W. Activity identification utilizing data mining techniques. In *IEEE Workshop on Motion and Video Computing*, pages 12-12, 2007. DOI: 10.1109/WMVC.2007.4

[14] Lloyd, S. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129-137, 1982. DOI: 10.1109/TIT.1982.1056489

[15] Locke, EA. and Latham, GP. The application of goal setting to sports. *Journal of Sport Psychology*, 7:205-222, 1985

[16] McCabe, A. and Trevathan, J. Artificial intelligence in sports prediction. In *Fifth International Conference on Information Technology: New Generation*, pages 1194-1197, 2008. DOI: 10.1109/ITNG.2008.203

[17] Miljkovic, D., Gajic, L., Kovacevic, A., and Konjovic, Z. The use of data mining for basketball matches outcomes prediction. pages 309-312. In *8th IEEE International Symposium on Intelligent Systems and Informatics*, 2010

[18] Min, B., Kim, J., Choe, C., Eon, H., R., Ian, and McKay, A. Compound framework for sports prediction: The case study of football. *Knowledge-Based Systems*, 21(7):551-562, 2008

[19] Ng, AY., Jordan, MI., and Y., Weiss. *On spectral clustering: Analysis and an algorithm.* Advances In Neural Information Processing Systems, 2001.

[20] Rauter, S., Fister, S., I., and I., Fister Jr. How to deal with sports activity data sets for data mining and analysis: some tips and future challenges. *International Journal of Advanced Pervasive and Ubiquitous Computing*, 7(2):1-11, 2015

[21] Sculley, D. Web Scale K-Means clustering. *Proceedings of the 19th International Conference on World Wide Web*, 2010

[22] Sykora, M., Chung, PWH., Folland, JP., Halkonand, BJ., and Edirisinghe, EA. *Advances in Sports Informatics Research*. 2015. Phon-Amnuaisuk S., Au T. (eds)

[23] Ward, P. Goal Setting and Performance Feedback. Ch. 6 in J.K. Luiselli, Springer Science+Business Media, LLC, 2011. DOI: 10.1007/978-1-4614-0070-7_6