A minimalistic toolbox for extracting features from sport activity files

Iztok Fister Jr.*, Luka Lukač*, Alen Rajšp*, Iztok Fister*, Luka Pečnik*, Dušan Fister†
*Faculty of Electrical Engineering and Computer Science
University of Maribor
Koroška cesta 46, 2000 Maribor, Slovenia
Email: iztok.fister1@um.si
†Faculty of Economics and Business
University of Maribor
Razlagova 14, 2000 Maribor, Slovenia

***Abstract*—Nowadays, professional, as well as, amateur athletes are monitoring their sport activities/training using modern sport trackers. These devices allow athletes to capture many indicators of sport training, e.g. location of training, duration of training, distance of training, consumption of calories. Until recently, not enough devotion was given to those indicators that are not visible directly, but can be obtained as the result of extensive data analysis, e.g. information extracted from topographic maps, weather conditions, and interval data. In line with this, the present paper is dedicated to describing the new toolbox for extracting features hidden in sports activity files. The results of the extraction serve as entry points for deep data analysis, that allows us to build intelligent systems for training support.**

DOI: 10.1109/INES52918.2021.9512927

I. Introduction

Nowadays, Machine Learning (ML) is causing a revolution in many research areas, and sports training is no exception. Not long ago, a common conviction was presumed that one cannot train without a personal coach. Today, the picture is a totally different. More and more efficient computational methods in ML have enabled rising the researches in automatic planning of sport training sessions. This domain has been gaining in popularity since its inception, and is nowadays a desire of not only individual athletes, but also of complete sports teams [2].

Since their inception, affordable sports trackers have proven costly for monitoring of sport activities. These allow the athletes to track their sports activities from start-to-end, and to upload the sport datasets into the cloud after realization of the sports training sessions. Sports trackers are a kind of mobile and pervasive technologies and can monitor different training load indicators, such as:

- the GPS position of a training, which is expressed as longitude, latitude and altitude; these are tracked by the GPS sensor integrated in the sports tracker.
- the ambient temperature,
- the power, the heart rate and the pedalling cadence on the bike.

Monitoring the data has become a very natural process of not only professional, but also semi-professional (amateurs) and curious recreational athletes. Websites, that collect such data, are able to contain plenty of database samples, but still have two common bottlenecks preventing a breakthrough in this area: On the one hand, the majority of research data in the sports domain are not publicly available [11], while on the other, data extraction methods from cyclist's databases demand a lot of preprocessing skills. The paper focuses on data preprocessing.

Typically, data preprocessing is one of the most intensive and demanding tasks in ML. It is one of the natural criteria that the data, which enter into the ML pipelines, must be prepared and aligned properly. A sample of these tasks include: filtering, data normalization, data discretization, feature selection, and feature extraction. The aim of the feature extraction is to generate new features from the observed datasets. Fister et al. [3] exposed the opportunities of data that are monitored by the sport trackers for data mining applications.

Most of the research referred to analysis of load indicators that were extracted easily from the set of sports activity datasets, e.g., the total distance, the total duration, and the average power. Unfortunately, no attention was paid practically to the indicators that are actually hidden in the data and thus must be extracted first. Some of these include basic topographic data, such as the number of hills and their categories, distances between hills, peak heights and distances of descents. On the other hand, many important data referring to the interval training sessions are hidden within the sports activity datasets. Altogether, the motivation of this paper is divided into four issues, as follows:

- to make an overview of indicators that can be extracted from sport activity datasets as features,
- to propose a tool for detection of topographic features in sport activity datasets,
- to propose a tool for extracting historical weather data,
- to develop a tool for detection of interval training sessions from datasets.

Finally, the developed tools are collected into a toolbox and applied to an archive of sports activity datasets. The proposed toolbox is then included into AutoML, providing complete

pipelines of methods for intelligent data sports training analysis.

The whole source code of the proposed toolbox is available in GitHub repository at:

https://github.com/firefly-cpp/sport-activities-features

The structure of the remainder of the paper is as follows: The proposed toolbox is described in Section II. The analysis of the activity datasets from a topology point of view is the subject of Section III. Parsing the historical whether data is presented in Section IV, while the algorithms for detecting the interval sports training sessions into activity datasets are described in Section V. Section VI shows how the preprocessed data can be incorporated into intelligent data sports training analysis. The paper concludes with Section VII, where summarizing of the performed work is done, and directions for the future work are outlined.

II. Overview of indicators in sports activities

The problem of overall (integral) load indicators found in sports activity datasets, such as total duration, total distance, average heart rate, etc., is commonly associated with the following biases:

- details are not expressed sufficiently,
- only a general/integral outlook of the race/training is captured,
- the intensity indicator of the realized race/training may be fallacious and
- different stages/phases during the sport race/training, i.e. warming-up, endurance, intervals, etc., are not recognized directly.

In cycling, for example, relying on integral indicators can be especially fallacious in cases of a climb (mountain) training, which is of very high intensity by itself, but the average speed and total distance are low compared to a relaxation ride, where average speed and total distance are similar to those of climb training, but the intensity of a ride is practically none. Also, these indicators can lack transparency in cases of specialized training, for example, when the majority of a ride is of descent intensity, but the climb is of maximum intensity. In such cases, a more detailed analysis of sport training data is necessary to extract the main point of the training (by focusing on the climb data) and disregarding the rest.

An interesting point of view is also weather conditions having a big influence on the realization of the sports training session. It does really matter if the sports training session is performed on a sunny or rainy day. In summary, almost three sources for extracting features hidden in cycling sports activity datasets can be found that identify (Fig. 1):

- hills,
- interval training sessions,
- weather conditions.

Identification of the mentioned sources for extracting features are described in detail in the remainder of the paper.

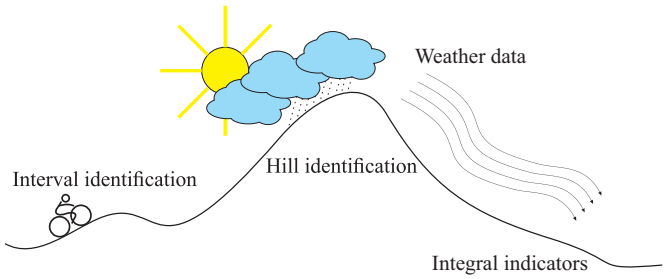

Fig. 1. Sources for extracting features hidden in cycling sports activity datasets.

III. Topographic maps' extraction

Topographic maps are dedicated to representation of relief on the Earth's surface typically, using contour lines that connect points of equal elevation. Elevation maps are a type of topographic maps that refer to representation of graphic location height above or below a fixed reference point. In cycling, the reference point is a start of the course, while heights above this denote climbs, and below this descents, that need to be overcome by the cyclists towards the end of the race. Obviously, the geolocation of the cyclist during the race in the sense of altitude, longitude and latitude, as well as time information is measures nowadays using the sports trackers equipped with GPS receivers, based on information obtained from four or more GPS satellites.

An example of the elevation map drawn according to analysis of the cycling sports activity is illustrated in Fig. 2, where

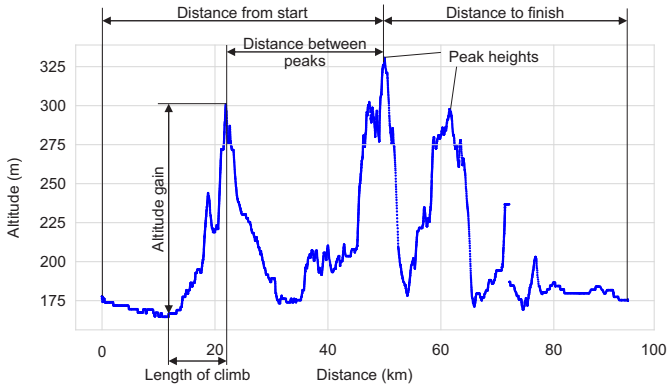

Fig. 2. An example of topographic map and its basic labelled features.

the total distance in kilometers is plotted on the x-axis, and the altitude in meters on the y-axis. Interestingly, these maps are commonly analyzed by cyclists and sports trainers before the cycling race in order to determine the length, altitude, intensity and number of hills within a route. Additionally, several other indicators can be outlined from topographic maps, such as: peak height, climb gradient, distance from the start when the climb begins, distance to finish, etc. On the basis of analyzing topographic maps, the daily team tactics are determined (keep in mind that in the case of a multi-stage cycling race, such as Giro d'Italia, Tour de France and Vuelta a Espana topographic

data are analyzed from several days further). Thus, cyclists who are good climbers (fr. grimpeur) are fully engaged during the short, but intensive mountainous stages, while, on the other hand, sprinters come into effect in more distant and flat stages. Commonly, the overall topographic profiles (considering all stages in a multiple stage cycling race) determine who will attend a specific race, which makes the elevation maps a very usable tool.

Hills in cycling are often treated as the most demanding of all terrains. We can realize that, in the case of a final climb, numerous bunches, consisting of 5-6 cyclists at most, arrive at the top of the climb separately. The primary focus of analyzing the elevation maps is, thus, to recognize the specifics of each climb. and identify any "hidden traps", such as short brutal slopes that may deteriorate team tactics.

According to the figure, we define some important features that are presented in Table II.

IV. Weather data extraction

Weather has a significant impact on the performance and behavior of cyclists. It can be checked before-hand and unfavorable weather conditions can be a reason to abandon a planned training session. It is well-known [12] that air temperature, relative and specific humidity, wind speed, solar shortwave radiation, thermal long-wave radiation, and precipitation can have significant influences on the performance of athletes in sports training. Wind can work in the cyclists' favor or against them, depending on its direction, while a strong crosswind may also influence the cyclist's stability. Temperature in combination with humidity impacts the effort needed by the cyclist determinedly, and high temperatures and humidity may pose significant stress on the cyclist performing training. The solar radiation resulting from sun can increase the perceived temperature and influence the cyclist in the same way as high temperatures. Precipitation can decrease the grip of the bike, which means that on a wet terrain the cyclist must adjust his speeds accordingly.

Travelling the same route can lead to very different outcomes, depending on the weather conditions which need to be measured as accurately as possible. In this sense, our toolbox incorporates the tracking of historical weather data from an external source [13]. The recorded exercise data (for each hour of exercise) are then sampled from the nearest weather station. The following parameters, which can be analyzed later, are collected: temperature, maximum and minimum temperature (during the time period), wind chill, precipitation (in mm/h), snow depth (if there is any), wind speed (in km/h), wind gust, wind direction, visibility (in km), cloud cover (in %) relative humidity (in %), conditions (e.g. clear, partially cloudy, rain, snowy), date of the measurement, and location of the weather station.

This allows us to evaluate each training better, and also shows that no two training sessions are the same, even if the cyclist follows the same route and maintains roughly the same speed.

V. Detection of interval in sports training sessions

Intervals in cycling training sessions are shorter periods of time, during which an athlete puts a significantly greater amount of energy into the exercise. In order to enhance maximal oxygen uptake ($VO_{2\,\max}$), which is one of the most important factors for determining success in an aerobic endurance sport, interval training sessions with longer intervals are recommended [7]. Thus, the detection of intervals from sport activity data is crucial in order to analyze all aspects of an activity thoroughly.

There are several ways to detect intervals within the given activity dataset, for example, according to heart rates or input power given at a certain time. Despite the fact that extracting intervals according to heart rate is very efficient, it is not possible if those data are insufficient or missing. On the other hand, calculating power from the other data is possible in almost all cases, but this calculation may return worse results, because the power can swing heavily during an interval.

According to these facts, both power based and heart rate based detection of intervals have been implemented in this toolbox. If detecting intervals according to power, power indicators of all segments have to be calculated. An interval is identified if the power of a certain segment exceeds the average power of the whole activity. The next step is to merge near intervals, which can be very problematic, since there is no good way of knowing whether the time between them is small enough to consider more intervals as one. After this is done, the last step is to remove intervals which are too short to be considered as intervals.

Algorithm 1: Detection of intervals by heart rate

```
1  function DETECT-INTERVALS(segments, minimumTime)
2      intervals ← {};
3      avg ← getAverageHeartRate(segments);
4      foreach segment ∈ segments do
5          if segment.heartrate > avg then
6              intervals ← intervals + {segment};
7          end
8      end
9      for i ← 1 to intervals.length do
10         if getAverageHeartrateBetween(intervals[i - 1],
             intervals[i]) < 10 then
11             merge(intervals[i - 1], intervals[i]);
12         end
13     end
14     foreach interval ∈ intervals do
15         if interval.time < minimumTime then
16             intervals ← intervals - {interval};
17         end
18     end
19     return intervals;
```

TABLE I
LIST OF EXTRACTED FEATURES BY TOPOGRAPHIC MAP EXTRACTION.

ID	*Feature*	*Num./Cat.*	*Description*
1	Number of hills	Numerical	How many hills exist in particular topographic map.
2	Average ascent of hills	Numerical	Average ascent of all hills in sport activity.
3	Average altitude of hills	Numerical	Average altitude of all hills in particular topographic map.
4	Distance of hills	Numerical	Total distance of all identified hills in sport activity.
5	The percentage of hills	Numerical	How many parts of activity is crowded with hills.
6	Total ascent of all hills	Numerical	The sum of ascent of all hills in map.
7	Total descent of all hills	Numerical	The sum of descent of all hills in map.

The algorithm based on heart rate, illustrated in Algorithm 1, is very similar to the one based on power, except for the detecting factor that enables combining more intervals. As the heart rate cannot change substantially in a glimpse of time, it is significantly easier to determine whether two or more intervals can be merged into one.

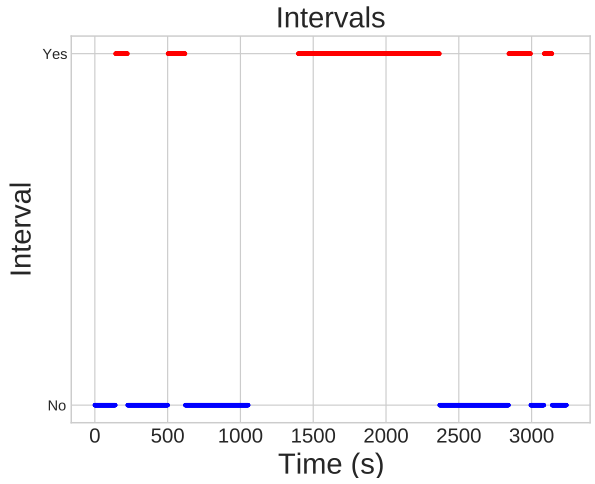

Fig. 3. An example of interval detection. Value "Yes" means that at that specific timepoint an interval has been detected. Value "No" means there was no interval at that timepoint.

VI. INTELLIGENT DATA ANALYSIS

The features described in the previous chapters determine detailed information about realized sports activities. In many cases, this information is hidden, as a very different set of feature values can belong to the same activity. For this reason, the construction of an ML model, i.e., classifier, that would be able to determine the activity from the values of the features, is a very demanding and time-consuming task [6], [4]. It also often makes sense to process existing features in such a way that they carry even more information, and in certain cases even eliminate certain features.

In these cases, we also need feature selection algorithms and feature transform algorithms [8]. Unfortunately, these algorithms require a lot of domain-specific knowledge in order to achieve the desired performance. Therefore, it is easier to use Automated Machine Learning methods (AutoML) [1], [5] to find the best combination of classifiers. These methods are able to propose the more successful set of algorithms that can solve a given problem based on the input data automatically. The set of algorithms is also called the Machine Learning pipeline [9].

There are many approaches and AutoML, one of which is NiaAML [4], [10]. The NiaAML framework searches for classification pipelines using nature-inspired algorithms. Obviously, all the proposed preprocessing methods within the sport-activities-features toolbox are included into this framework. Presumably, the best possible classification pipeline can be found on the basis of a given training set for determining sports activities from the features. Later, this pipeline can be exported and reused on the unseen data.

VII. CONCLUSION

The results of an athlete involved in sports training are not a product of coincidence, but the hard work of athletes, as well as sports trainers. Therefore, the domain of sports training has been connecting more and more with the computer science that offers tools for analysis data obtained by mobile devices worn during the training session with modern ML methods. Moreover, the results of this analysis can be useful to the sports trainers as well as amateur athletes training alone.

Unfortunately, a lot of data obtained from mobile devices are hidden, and can be explained with sophisticated algorithms before entering into pure ML methods. The paper is focused on the development of extracting hidden features within sports activities. In line with this, the following preprocessing methods were proposed, constituting the minimalistic toolbox: detailed analysis of topological maps, weather conditions, and identification of interval training sessions in cycling. All three preprocessing methods are included into the AutoML framework, that enables an intelligent data analysis of sports training data.

The minimalistic toolbox reveals a lot of potential directions for the future work, as follows: The methods could be included into Artificial Sport Trainer (AST) [2]. Primarily, the analysis was performed in cycling. Obviously, the same methods could also be applied in other sports. Last but not least, these could also be widened to team sports (e.g., football, basketball, volleyball).

REFERENCES

[1] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated

TABLE II
LIST OF EXTRACTED INTERVAL FEATURES

ID	*Feature*	*Num./Cat.*	*Description*
1	Number of intervals	Numerical	Number of intervals in an exercise
2	Minimum duration of an interval	Numerical	Time of the shortest interval in an exercise
3	Maximum duration of an interval	Numerical	Time of the longest interval in an exercise
4	Average duration of an interval	Numerical	Time of the average interval in an exercise
5	Minimum distance of an interval	Numerical	Minimum distance of an interval in an exercise
6	Maximum distance of an interval	Numerical	Maximum distance of an interval in an exercise
7	Average distance of an interval	Numerical	Average distance of an interval in an exercise
8	Minimum heart rate of an interval	Numerical	Minimum heart rate during an interval in an exercise
9	Maximum heart rate of an interval	Numerical	Maximum heart rate during an interval in an exercise
10	Average heart rate of an interval	Numerical	Average heart rate during an interval in an exercise

machine learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28, pages 2962–2970. Curran Associates, Inc., 2015.

[2] Iztok Fister, Iztok Fister Jr, and Dušan Fister. *Computational intelligence in sports*. Springer, 2019.

[3] Iztok Fister Jr., Iztok Fister, Dušan Fister, and Simon Fong. Data mining in sporting activities created by sports trackers. In *Computational and Business Intelligence (ISCBI), 2013 International Symposium on*, pages 88–91. IEEE, 2013.

[4] Iztok Fister Jr., Milan Zorman, Dušan Fister, and Iztok Fister. Continuous optimizers for automatic design and evaluation of classification pipelines. In *Frontier Applications of Nature Inspired Computation*, pages 281–301. 2020.

[5] I. Guyon, K. Bennett, G. Cawley, H. J. Escalante, S. Escalera, Tin Kam Ho, N. Macià, B. Ray, M. Saeed, A. Statnikov, and E. Viegas. Design of the 2015 chalearn automl challenge. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2015.

[6] Xin He, Kaiyong Zhao, and Xiaowen Chu. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, page 106622, 2020.

[7] Jan Helgerud et al. Aerobic high-intensity intervals improve vo2max more than moderate training. *Official Journal of the American College of Sports Medicine*, 2007.

[8] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. *Automated machine learning: methods, systems, challenges*. Springer Nature, 2019.

[9] Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, and Jason H. Moore. Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, GECCO '16, pages 485–492, New York, NY, USA, 2016. ACM.

[10] Luka Pečnik and Iztok Fister Jr. Niaaml: Automl framework based on stochastic population-based nature-inspired algorithms. *Journal of Open Source Software*, 6(61):2949, 2021.

[11] Alen Rajšp and Iztok Fister Jr. A systematic literature review of intelligent data analysis methods for smart sport training. *Applied Sciences*, 10(9):3013, 2020.

[12] Timo Vihma. Effects of weather on the performance of marathon runners. *International Journal of Biometeorology*, 54(3):297–306, may 2010.

[13] Visual Crossing Corporation. Weather Data & Mapping — Visual Crossing, 2020.