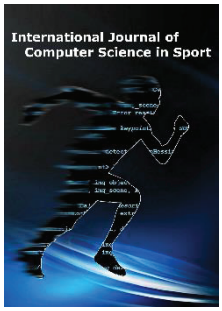
International Journal of Computer Science in Sport

Volume 21, Issue 2, 2022

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

sciendo

DOI: 10.2478/ijcss-2022-0008



Time Series Data Mining for Sport Data: a Review

Rumena Komitova, Dominik Raabe, Robert Rein, Daniel Memmert

Institute of Exercise Training and Sport Informatics, German Sport University Cologne, Am Sportpark Müngersdorf 6, 50933 Cologne, Germany

Abstract

Time series data mining deals with extracting useful and meaningful information from time series data. Recently, the increasing use of temporal data, in particular time series data, has received much attention in the literature. Since most of sports data contain time information, it is natural to consider the temporal dimension in form of time series. However, in sports, the effective use of time series data mining techniques is still under development. The main goal of this paper is therefore to serve as an introduction to time series data mining and a glossary for interested researchers from the sports community. The paper gives an overview about current data mining tasks and tries to identify their potential research direction for further investigation. Furthermore, we want to draw more attention with respect to the importance of mining approaches with sport data and their particular challenges beyond usual time series data mining tasks.

KEYWORDS: DATA MINING, TIME SERIES, REPRESENTATION, TIME SERIES SIMILARITY MEASURES, CLASSIFICATION, CLUSTERING, SPORT SCIENCE

Introduction

One of the essential features many data have is a time dependency. Tracking the behavior of specific data in time can produce important information. Various real world applications depend on the analysis of time-dependent data. This includes for example data from the medical domain, human activity recognition, motion data, and sports data (e.g. electrocardiogram (ECG), and electroencephalogram (EEG), data generated from wearable sensors, location of moving objects, etc.). In all these cases, observations over time are collected which constitute so-called *time series*. For instance, in the sports domain one major source of temporal data generated stems from sensor data (Memmert & Raabe, 2018; Memmert et al., 2017; Rein et al., 2017). Most of these data are of time-series or similar character, where data are recorded continuously. Application of the techniques and principles of data mining for the analysis of time series resulted in the concept called *Time Series Data Mining* (TSDM) (Esling & Agon, 2012; Fu, 2011; Mitsa, 2010).

Traditional time series analysis (Box, Jenkins, & Reinsel, 2016) contains methods to analyze time series data in order to extract temporal rules from the structure of time series, such as trends, changes in value, seasonality, or other characteristics of the data to generate an accurate forecast. Time series analysis in sports can be used to predict the result of sports competition, or sports popularity (Miller, Schwarz, & Talke, 2017). For example, Yong, Lingyun, & Jia, (2020) collected the Boston marathon results over several years. They analyzed the athletes' performance through statistical analysis and used the autoregressive moving average model (ARMA) to predict the athletes' next average competition performance. Li, Wang, & Li (2021) used multivariate logistic regression analysis to predict the relationship between the winning probability and game results. Subsequently the player's and the team's historical data was used to predict game outcomes, which can be used for the team's future training. TSDM, in contrary, deals with much larger amounts of time series data and much greater numbers of time series. However, the focus of TSDM is less on the analysis of the statistical properties of time series data, but instead focuses on the discovery of hidden relations between time series data and extract potentially useful and meaningful information from them, where the terms "useful" and "meaningful" depend on the application. In other words, the data miner is seeking for patterns (or motifs), novelties and anomalies in a time series dataset that are frequently repeating or atypical, whereas the statistician is usually seeking for comprehensive structure that explains the data. Furthermore, the application of TSDM benefits from feature extraction and transformation techniques designed in other scientific communities such as traditional time series analysis, signal processing, and machine learning.

With the amount of information and data in the sport domain, opportunities for data mining and machine learning applications can be extremely widespread, and benefits from the results can be enormous (Bonidia, Rodriges, Avila-Santos, A.P., Sanches, & Brancher, 2018; Horvat & Josip, 2020; Li & Zhang, 2012; Ofoghi, Zeleznikow, MacMahon, & Raab, 2013; Schumaker, Soleiman, & Chen, 2010). The evolution of tracking systems has become another opportunity for researchers in the sport domain to extract new knowledge relate to player performance, decision-making, and movement patterns, among others (Biermann et al., 2021; Jamil et al., 2021; Raabe et al., 2022; Rein & Memmert, 2016; Stein et al., 2017). Recurring and surprising events extracted from stored sport data by TSDM methods can be important in assisting trainers to make informed decisions, thereby improving training techniques. For instance, time series motif discovery, one of the tasks in TSDM to discover frequent patterns which are unknown previously, can be helpful for sport activity recognition (Yeh, Kavantzas, & Keogh, 2017) and patterns are helpful when forecasting future events, such as prediction of performance, and in athletes training (Schmidt, 2012). The automatically activity classification could thus help document player's statistics such as type of movement each player has performed and the number of movement repetitions allows coaches to quantitatively supervise player's physical exertion, and further help prevent potential injuries. A review of existing data mining techniques and their possible application to sports data can be also found in Stein et al. (2017).

We believe that TSDM - a topic that has been largely glossed over in the sport domain literature - has a great potential for exploratory data analysis in sports and can supply the sport community with a set of methodologies that can be used to model and study sports-related processes and phenomena. TSDM techniques could help in obtaining a better understanding and exploitation of sport data and thus provide a basis for more objective evaluation of performance and potentially novel methodologies for the sport sciences. Furthermore, we believe that TSDM has the potential to become an integral part of future sport analytics approaches. From a scientific point of view, the availability of time series datasets presents an opportunity for the TSDM and machine learning communities to develop, validate, and apply new techniques in the sport

domain. Our objective in this paper is not to be exhaustive on TSDM tasks but rather to give some instances of such tasks in the sport domain.

Key Concepts in Time Series Data Mining

TSDM tasks in the literature are time series classification, clustering, motif discovery, and anomaly detection. While these different tasks used to mine time series data, they all require some high level representation of the time series, rather than the original raw time series.

Time Series Representations

TSDM can be exploited from research areas dealing with signals, such as signal processing. For example, sport data can be transformed into time series. However, although analysis of the whole time series dataset might be essential for statistical purposes and forecasting, this is not required for most data mining tasks. Time series are high-dimensional data (Keogh & Kasetty, 2003). When dealing with large volumes of data, the representation of time series data is especially important and is a key to efficient and effective solutions since direct manipulation of the original data may be very difficult. A *time series representation* transforms a time series into another time series to design low-dimensional representations which preserve the local and global shape characteristics of a time series.

Several techniques have been proposed to represent time series by certain types of transformation, such as the Discrete Wavelet Transform (DWT) (Chan & Fu, 1988), and the Discrete Fourier Transform (DFT) (Faloutsos, Ranganthan, & Manolopoulos, 1994), which transform the original time series into the frequency domain. For example, in order to preserve the useful information from time series data from wearable sensors, Ahmadi et al. (2014) used the coefficients from DWT to classify time series of sport activities, including agility cuts, jogging and jumping. Mitchell, Monaghan, & O'Connor (2013) used the DWT coefficients as the features to classify different activities during soccer and field hockey.

Symbolic Aggregate Approximation (SAX) (Lin, Keogh, Lonardi, & Chiu, 2003; Lin, Keogh, Wei, & Lonardi, 2007) were developed to transform time series data into symbolic form and to reduce dimensionality of time series data by discretization of the original data into a collection of symbolic string alphabet. SAX has traditionally been used when the goal of time-series analysis is to model and compare. SAX was not originally meant for classification even however, Junejo & Al Aghbari (2012) used a SAX time series representation to model human action recognition. The majority of works use SAX for analysis of ECG, EEG, or accelerometer/inertia sensor data. Many papers referencing one of the SAX papers often refer to SAX as a possible tool for time series classification without applying it direct in the sport science domain.

Similarity Measures for Time Series Data

After representing the time series data in a suitable form, it is necessary to carefully define a similarity measure in order to determine if the different times series match. Given atime series database, most of the TSDM tasks rely on a definition of *similarity* measure function between two time series, which calculates the distance between the two time series. The term similarity and distance measures are usually used interchangeably. Which distance measure fits best depends thereby on the context and the choice of a proper similarity measures, the characteristic of time series, the length of time series, and the representation method (Aghabozorgi et al., 2015; Fu, 2011; Wang et al., 2012).

One way to estimate the similarity between two time series is to use the L_p norm as distance measure. For p=2, we obtain the Euclidean distance, which is one of the simplest and most

widely used *shape-based* similarity measure for time series (Keogh & Kasetty, 2003). Shape-based similarity measures compare the global shape of time series. Other variants of distance measures based on the L_p-norms are the Manhattan Distance (p=1), and Maximum Distance (p=∞) (Ding et al. (2008); Mitsa (2010)).

Euclidean distance-based time series similarity plays a big role in time series data mining. However, although the Euclidean distance is simple to understand and easy to compute, it can be used only when two time series have the same length. This may be not appropriate when the time series data is generated at different sampling rates, or when the data sampling may change over a long time. Another limitation is that the Euclidean distance is very sensitive with respect to signal transformations, such as shifting, uniform amplitude scaling, or uniform time scaling (Keogh & Ratanamahatana, 2002).

To obtain more robust similarity metrics, other distance measure techniques were proposed. For example, the well-known *Dynamic Time Warping* (DTW) (Keogh & Ratanamahatana, 2002) makes distance comparisons between time series less sensitive to the signal transformations mentioned above and can also detect similar shapes at different time steps and amplitude. The advantage of DTW is that it can compare the similarity between time series data by accommodating possible length differences. DTW has been advocated to be the most accurate similarity measure (see (Ding et al., 2008) for an extensive comparison of distance measures of time series). It has been widely used in applications such as speech recognition (Berndt & Clifford, 1994), motion data clustering, or activity template classification (Seto, Zhang, & Zhou, 2015). For example, Sempena, Maulidevi, & Aryan (2011) used DTW to recognize various human activities such as punching, boxing or jogging. Srivastava et al. (2015) classified tennis shots into forehand, backhand or serve by applying DTW on accelerometer and gyroscope data. Similarly, Hu et al. (2020) demonstrate a basketball activity classification model based on a DTW algorithm applied to body kinematic measures. However, DTW has not been fully implemented in sport applications.

Amplitude differences and scaling problems can produce a large distance between time series even if they have a similar shape. To avoid the problem of amplitude shifting and in order to make meaningful comparison between two (one-dimensional) time series under Euclidean distance, DTW or any other distance measure, data are typically normalized, i.e. both time series are transformed to have a mean of zero and unit standard deviation (Keogh & Kasetty, 2003). An extensive comparison of representation and distance measures of time series can be found in (Ding et al., 2008).

Time Series Data Mining Tasks

After representing the time series data, one can proceed with the typical data mining tasks such as classification, clustering, anomaly detection, and motif discovery. In this section, a brief overview of major tasks considered by the TSDM community will be provided. We refer interested readers to Fu (2011) and Esling & Agon (2012) for a more in-depth review.

Time Series Classification

Classification is one common type of supervised learning widely used to develop predictive models. The term "supervised" indicates that the output is known, meaning we assume we have some domain knowledge about the problem we try to solve. Given an unlabeled time series, the goal of time series classification is to assign it to one class of a given number of predefined classes (Keogh & Kasetty, 2003).

Time series classification is an important and challenging problem in time series data mining and

is applied in various domains. For example, in recent years, the use of time series data mining methods in medical domains has gained increasing interest. The signals generated by an Electrocardiography (ECG), and by an Electroencephalography (EEG) are time series, whose analysis has brought major advances in the medical domain (Anguera, Barreiro, Lara, & Lizcano, 2016). EEG time series generated after monitoring the brain activities contain a series of waves characterized by their frequency and amplitude. It is possible to classify certain types of special waves such as anomalies, or motifs (special time series subsequences that are interesting for domain experts) that are characteristic of some neurological pathologies, like epilepsy (Anguera, Barreiro, Lara, & Lizcano, 2016). In case of ECG time series data from some patient, a medical doctor can make use of supervised algorithms to search for some similar patterns in a time series ECG dataset and classify for example heart rates into certain diagnostic groups and require training sets in advance. The retrieved information can help to provide the correct diagnoses. In other various scenarios, where temporal data is involved, class labels can be associated with time stamps or time sequences corresponding to frequently recurring motifs and abnormal behaviors.

Recently, variousmethods to classify time series data have been proposed. One of the most popular performing classifier in the field is the k-Nearest-Neighbor classifier (k-NN). k-NN can be easily applied for a particular problem by choosing a representation and a similarity measure. As indicated by the name, the k-NN classifier takes the *k* nearest neighbors into account. Examples are the Nearest-Neighbor classifier using the Euclidean distance metric (Mueen, 2014) or Dynamic Time Warping (DTW) (Senin, 2008). Recent contributions have focused on developing ensembling methods that significantly outperforms the NN coupled with DTW (NN-DTW) (Bagnall, Lines, Bostom, Large, & Keogh, 2017). Plenty of research proposes DTW as the best distance-based measure to use along k-NN (Seto, Zhang, & Zhou, 2015). Moreover, Xi, Keogh, Shelton, Wei, & Ratanamahatana (2006) claim that the 1-NN-DTW classifier which combines DTW as similarity measure and a 1-NN classifier is the best time series classification approach. Another popular method in time series classification are decision trees. Decision tree infer a set of rules from the training data which supports simple interpretation of the results. The set of rules is subsequently applied to any new time series data to be classified (Ratanamahatana et al., 2010). This method does not require feature selection. Nevertheless, the most of the classifiers need prelabeled training data and hence expert knowledge and supervision.

In the sport domain, time series classification has been applied mostly for activity recognition in some specific sports such as table tennis (Maeda, Fujii, Hayashi, & Tasaka, 2014; Blank, Hoßbach, Schuldhaus, & Eskofier, 2015), and soccer (Hossain, Khan, & Roy, 2017). Hossain et al. (2017) studied the use of wrist-worn sensors to classify motion performed by soccer field players such as passes, kicks, sprints and runs. One approach to classification is to use the training data to build time series sequences representing each activity, and subsequently classify the test data according to their similarity to these sequences (Seto, Zhang, & Zhou, 2015). Also, wearable sensors have been used for detection and classification of training exercises for goalkeepers (Haladjian, Schlabbers, Taheri, Tharr, & Bruegge, 2020), to study the impact of the ball on the head (Worsey et al., 2020), to segment swimming stroke phases using sensors (Wang et al., 2020), to automatically detect collisions in Rugby (Kelly, Coughlan, Green, & Caulfield, 2012), and other activities in sports (Wang et al., 2018). Haladjian et al. (2020) present an algorithm to automatically detect and classify goalkeeper training exercises using a wearable sensor attached to a goalkeeper glove. Their approach detects the exercises using DTW to detect and eliminate irrelevant motion instances and extracts a set of statistical and heuristic features to describe the different exercises and train a machine learning classifier. Another approach represents pitch-by-pitch sequences as time series data using baseball's linear runs and build a model using the k-Nearest Neighbors classification method (Soto-Valero, González-Castellanos, & Pérez-Morales, 2017). In order to compare time series of pitcher's performance, DTW was

used as similarity measure. Hu, Mo, & Qu (2020) proposed a basketball activity classification model based on body kinematic measures where DTW was used to classify four different basketball activity patterns such as shooting, passing, lay-ups and dribbling. Finding interesting patterns within the time series data can thus provide meaningful information not only about single player but also about team performance.

Often, the activity classification algorithms are used to recognize various human activities based on the represented features. Features can be used by algorithms, such as machine learning to perform classification of the data. A large set of features can be extracted from time series to describe and represent the information contained within the data. There are many different strategies of features extraction. For instance, Tanaka, Iwamoto, & Uehara (2005) transformed multiple time series into one dimension using PCA. However, this approach will not able to capture activities from sensors with different data lengths. Another challenge often is the high dimensionality of the feature space for the time series data and thus the cost associated with computations (Xing, Pei, & Keogh, 2010). In general, much emphasis has been placed on deep learning-based methods, as they are capable of automatic and problem-specific feature extraction. Recently, deep learning is increasingly embraced for solving time series related tasks, including time series classification (Ismail Fawaz et al., 2019). However, sensor-based activity recognition, feature extraction is a difficult task because of inter-activity similarity (Bulling, Blanke, & Schiele, 2014). Different activities may have similar characteristics. Therefore, it is difficult to produce features to represent activities uniquely.

Time Series Clustering

In case of huge time series datasets, using supervised approaches such as classification is often impossible, while clustering can solve this problem using unsupervised solutions (Aghabozorgi et al., 2015). While time series classification assigns unlabeled time series data to predefined (labeled) classes after a supervised learning process, time series clustering refers to the division of data into natural groups, so called *clusters*, such that the time series belonging to a cluster are similar to one another, and time series belonging to different clusters differ significantly from one another (Liao, 2005).

Recently, clustering of time series has attracted the interest of researchers and has become an important topic (Aghabozorgi et al., 2015; Esling & Agon, 2012; Lin et al., 2002; Mitsa, 2010). Time series classification and clustering can be a stage for several other time series mining tasks, such as motif discovery, and anomaly detection algorithms (Keogh, Lin, Lee, & Van Herle, 2006). A main motivation for representing a time series in the form of clusters is to better represent the main characteristics of the data. Furthermore, the aim of time series clustering is to find clusters that are meaningful, in some sense, based on the information in the data. For example, in (Siirtola, Laurinen, Haa- palainen, Roning, & Kinnunen, 2009), the authors presented a clustering-based method for classifying activities. The activity recognition was performed in two stages. In first stage, the sample is clustered into one of six clusters (one activity can be included in more clusters). In the second step, a decision tree model is used to recognize the sports activity. These clusters are rarely known a priori provided by a domain expert, thus they are best learned through data. However, without a priori knowledge, is often difficult to interpret what each of the clusters refers to.

Time Series Anomaly Detection

Detecting anomalous subsequences in time series data is a broad field which has attracted much attention and has been studied in many research areas (Agarwal et al., 2015; Braei & Wagner, 2020; Fu, 2011; Mitsa, 2010; Pimentel et al., 2014). Time series *anomalies* can be defined as unexpected or unusual patterns in time series data that do not conform to a well-defined notion

of the expected (normal) behavior (Zolhavarieh et al., 2014). In other words, anomalies appear when the underlying process deviates from its normal behavior. The problem of finding them in a given time series is referred to as *time series anomaly detection*. Figure 1 depicts an example of anomaly detection. Different terms and inter- changeable synonyms are used in literature that have similar meaning to anomaly detection, such as *event detection*, *novelty detection*, *change point detection*, *fault detection*, and *outlier detection* (Gupta, Gao, Aggarwal, & Han, 2013; Pimentel, Clifton, Clifton, & Tarassenko, 2014). For example, change point detection is mainly used in the statistics literature and seek data points in a time series at which characteristics of the time series change (Agarwal et al., 2015). Fault detection methods are often applied for monitoring technical systems in industrial applications and describe the change of a system behavior to a faulty state. What is common for them is that anomalies are defined as observations that are significantly different from other observations (Chandola, Banerjee, & Kumar, 2009).

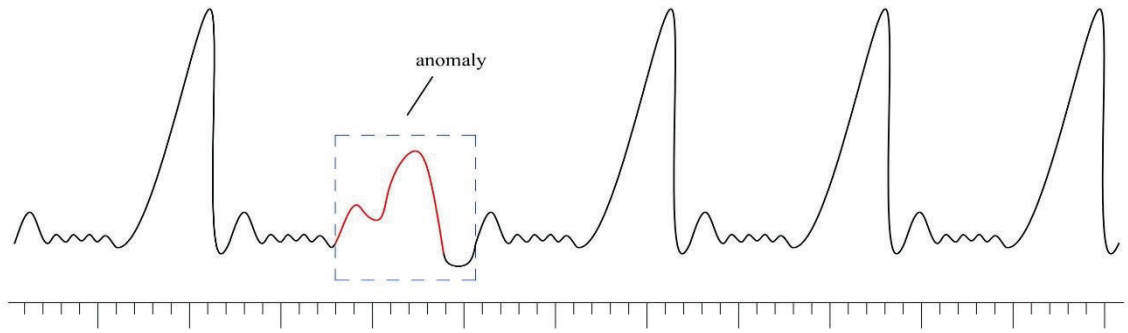


Figure 1. An example of the anomaly detection task

Anomalous data can be broadly classified into three general categories (Chandola et al., 2009). A *point anomaly* is the simplest type of anomaly and refers to a data point being anomalous, i.e. a point that deviates significantly from all the points in the dataset (Braei & Wagner, 2020). Point anomaly occurs in almost all datasets. *Contextual anomalies* are data points whose values are anomalous with respect to a specific context, but not otherwise. That is, a given behavior might be "normal" in concrete context but abnormal on another. In various cases, defining a context is not easy, making it difficult to apply anomaly detection techniques. Contextual anomalies are generally studied in time series data or spatial data (Gupta et al., 2013). Finally, a *collective anomaly* refers to a collection of related data instances that individually may not be anomalies, but their collective appearance is anomalous. Knowing a priori which type of anomaly the time series data might contain, meaning what an anomaly actually looks like in the given dataset, helps the data analyst to choose the appropriate detection method or to create a model. A survey that provides a structured overview of the research on anomaly detection can be found in Chandola et al. (2009), Gupta et al. (2013) and Schmidl et al. (2022).

Anomaly detection requires a combination of domain expertise as well as in machine learning and time series data mining algorithms. Since many anomaly detection algorithms have been developed independently and by different research communities, choosing the best detection technique is a difficult task (see Schmidl, Wenig, & Papenbrock (2022) for a comprehensive survey of existing state-of-the-art anomaly detection algorithms). Furthermore, the performance of the anomaly detection technique depends on the type of anomalies to be detected, their duration, as well as the anomalous time series and evaluating their performance on real-world anomaly detection tasks is far from trivial, event when labeled time series datasets are available

(Wu & Keogh, 2021).

Time Series Motif Discovery

A common problem in the TSDM and machine learning community is the identification of previously unknown and approximately repeated subsequences of single (or multiple) time series which are very similar to each other, also called *motifs* (Lin et al., 2002). *Time series motif discovery* is the technique to find them. Figure 2 depicts an example of motif discovery. The exact definition of the most relevant motifs varies slightly among authors and areas of application. Examples for motifs can be peaks (e.g. local minima or maxima), changes of noise characteristics of time series, or a variations of time or spectral components, which repeatedly occur in a time series.

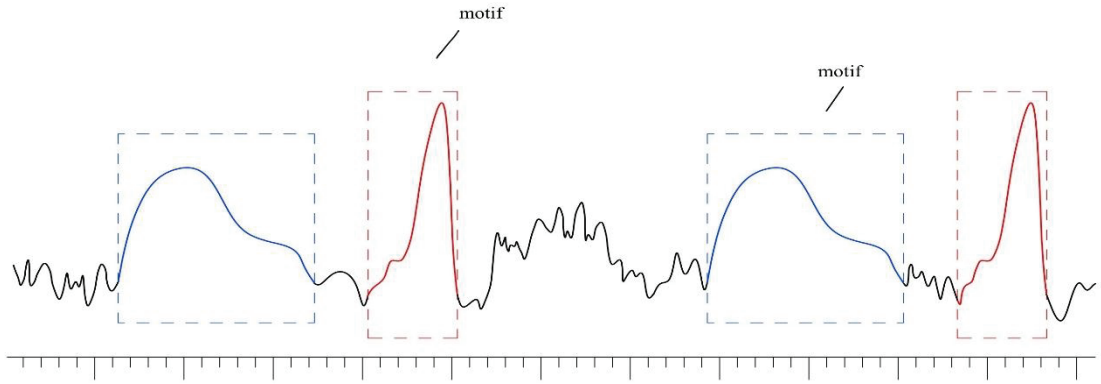


Figure 2. An example of the motif discovery task. The task of motif discovery consists in finding frequency appearing subsequences in longer time series.

In recent years, motif discovery has received a great amount of attention. Therefore, the available approaches and literature are manifold. It has been used as a subroutine in many TSDM tasks and applications such as clustering, and classification (Esling & Agon, 2012; Fu, 2011; Mueen, 2014; Torkamani & Lohweg, 2017). What makes motif discovery in time series data interesting is its applicability to a diverse range of domain applications, such as medicine (Lin & Li, 2010; Liu, Li, Chen, Tan, Chen, & Zhou, 2015; Sivaraks & Ratanamahatana, 2015), sensors (Agarwal et al., 2015), motion capturing (Tanaka et al., 2005; Minnen et al., 2006), and trajectory mining (Oates et al., 2013), among others. For example, in motion capture a particular gesture (e.g. throwing a ball). In case of ECG time series, motif discovery can help to locate the highly similar and rapid beats in the ECG.

Many techniques have been proposed for time series motifs extraction. The main objective is often to split a long time series into a sequence of smaller time series (or motifs), which are semantically meaningful to the domain experts. For instance, long series of sport activities can be classified into a frequently repeating patterns, representing a sequence of shooting, dribbling, kicking, and other activities. To that aim, each sequence is labeled as one specific activity class. Thus, an efficient time series motif discovery algorithm can be useful as a tool for visualizing big time series databases. However, finding motifs is a difficult task, even when they have the same or very similar general characteristics, because in most cases the number of occurrences of motifs, their shape, and duration of occurrences may be unknown (Mitsa, 2010). Note that, depending on the application, distortions such as time and amplitude shifts may be accepted (Torkamani & Lohweg, 2017). Common challenges in this motif discovery problem include scalability, the detection of motifs with various lengths (Linardi, Zhu, Palpanas, & Keogh, 2018),

multivariate time series (Tanaka et al., 2005; Yeh et al., 2017), how to formalize the usefulness of such motifs, and how the most useful ones can be found.

One key issue in motif discovery is the (low dimensional) representation of the time series or the time series subsequences. Another key issue is an appropriate similarity measure for time series used for comparing possible pairs of time series subsequences. Different aspects of motif discovery including domain dependent preprocessing, and algorithmic techniques based on their similarity measure and representation are described byMueen (2014). Torkamani & Lohweg (2017) analysed motif discovery algorithms with respect to several data characteristics like for example time complexity. As there can be many repeated patterns in a long time series appearing in different locations, lengths, similarities and frequencies, most of the motif discovery algorithms involve definition of the motifs. Since motifs of different lengths occur in many other real-world applications (Keogh et al., 2003; Lin et al., 2002; Tanaka et al., 2005), discovering variable-length motifs in time series has received a great amount of attention (Gao & Lin, 2018). When the lengths are unknown, trying different lengths can be very time consuming. Since the motifs have different lengths, existing (fixed length) motif discovery methods can only detect some, or just one of these motifs if the proper length is defined. However, most papers tend to focus on the author's own data sets and scientific problem. Hence, various motif discovery methods are described with different vocabulary and different focus, making it difficult to spot similarities between methods.

Conclusion

Although, traditionally, analysis of sport data is based on expert knowledge and statistical analysis, data mining techniques have become increasingly interesting to the field. Besides classical time series analysis methods for handling seasonality and analyzing the temporal structure of time series to forecast future values, time series data mining (TSDM) community produced techniques for large data sets. A major challenge in the sport domain today is how to exploit the huge amount of data that this field generates. Nevertheless, time series data types are common in the sport domain and require often particular analysis techniques. Especially since increasingly novel sensor technologies are used in sports, it is necessary to process and evaluate the multitude of physiological as well as biomechanical data. However, little use has been made of data mining techniques in sport for systematic knowledge discovery to date. Therefore, before trying to discover knowledge from any sport time series dataset, it is fundamental to have a clear understanding of these data. Extracting information about the behavior of time series data is potentially very important and useful for the analysis of sport data, yet inherently difficult problem. Apart from understanding the data of the sport domain, the data miner must be familiar with the difficulties poses by the particular characteristics of time series sport data, especially their high dimensionality and their noise characterisitics.

There exist many works dealing with TSDM. In this paper, we have presented an overview to the field of TSDM as well as to identify future research aspects to contribute on. The TSDM tasks presented in this paper are typical of the machine learning field, however time series are domain dependent and often common approaches developed in the general machine learning literature cannot be applied. For instance, original time series data can not always be considered suitable to be fed to machine learning algorithms. However, training predictors from time series often requires a similarity measure between time series, either based on the raw time series or their representations. Although TSDM techniques in order to discover unknown knowledge are a good candidate for the application on sports data, they have not attracted much attention in the sport domain. Approaches are required that are capable of discovering knowledge supporting decision making processes in the sport domain. One of the aspects that we may be looking for is

the existence of patterns, recognition of movement patterns (motifs), the recognition of movement errors (anomalies), or also the anticipation of movement results in different sports. Considering the large volume of research and development of motif discovery tools, it is surprising that these tools have rarely applied to time series in the sport domain.

Sport analysis is interesting research field where experts for various computer science subjects (data mining, data science, computer vision, machine learning, etc.) can bring in their expert knowledge. For meaningful analysis, domain requirements coming from different experts must be taken into account. We see that both data-informed approaches from TSDM and concepts developed in sport science can go hand in hand to design meaningful analysis. Future work needs to better recognize the role and influence of both approaches and how to combine them.

Acknowledgements

This research was supported by a grant from the German Research Council (DFG, Deutsche Forschungsgemeinschaft) as well as Federal Ministry of Education and Research (BMBF; MM4SPA) to the fourth autor DM (grant ME 2678/26.1; grant 01IS20021A).

References

Agarwal, P., Shroff, G., Saikia, S., & Khan, Z. (2015). Efficiency discovering frequent motifs in large-scale sensor data. *Proceedings of the second ACM IKDD conference on data sciences*, (pp. 98–103).

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering–a decade review. *Informion Systems*, *53*, 16–38.

Ahmadi, A., Mitchell, E., Richter, C., Destelle, F., Gowing, M., O'Connor, N., & Moran,

K. (2014). Toward automatic activity classification and movement assessment during a sports training session. *IEE Internet of Things Journal*, *2*(1), 23–32.

Anguera, A., Barreiro, J., Lara, J., & Lizcano, D. (2016). Applying data mining tech- niques to medical time series: an empirical case study in electroencephalography and stabilometry. *Computational and structural biotechnology journal*, *14*, 185–199.

Bagnall, A., Lines, J., Bostom, A., Large, J., & Keogh, E. (2017). The great time se- ries classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *31*(3), 606–660.

Biermann, H., Theiner, J., Bassek, M., Raabe, D., Memmert, D., & Ewerth, R. (2021). A Unified Taxonomy and Multimodal Dataset for Events in Invasion Games. *In Proceedings of the 4th International Workshop on Multimedia Content Analysis in Sports.*

Berndt, D., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. *KDD workshop*, *10*(16), 359–370.

Blank, P., Hoßbach, J., Schuldhaus, D., & Eskofier, B. (2015). Sensor-based stroke de- tection and stroke type classification in table tennis. *Proceedings of the 2015 ACM International Symposium on Wearable Computers*, *2018*, 93–100.

Bonidia, R., Rodriges, L., Avila-Santos, A.P., Sanches, D., & Brancher, J. (2018). Computational intelligence in sports: a systematic literature review. *Advances Human-Computer Interaction*, *2018*, 1–13.

Box, G., Jenkins, G., & Reinsel, G. (2016). *Time series analysis: forecasting and control*. New Jersey: John Wiley & Sons, Inc., Hoboken, fifth ed.

Braei, M., & Wagner, S. (2020). Anomaly detection in univariate time-series: a survey on the state-of-the-art. *arXiv preprint arXiv:2004.00433*.

Bulling, A., Blanke, U., & Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, *46*(3), 1–33.

Chan, K., & Fu, A. (1988). Efficient time series matching by wavelets. In *Proceeding of the 15th International Conference on Data Engineering, (Cat. No. 99CB36337), IEEE*, (pp. 126–133).

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: a survey. *ACM Computing Surveys*, *41*(3), 1–58.

Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., & Keogh, E. (2008). Querying and mining of time series data: experimental comparison of representations and distance measures. *PVLDB Endowment*, *1*(2), 1542–1552.

Esling, P., & Agon, C. (2012). Time series data mining. *ACM Computing Surveys (CSUR)*, *45*(1), 1–34.

Faloutsos, C., Ranganthan, M., & Manolopoulos, Y. (1994). Fast subsequence matching in time-series databases. *ACM SIGMOD International Conference on Management of Data*, *23*(2), 419–429.

Fu, T.-C. (2011). A review on time series data mining. *Engineering Applications of Artificial Intelligence*, *24*(1), 164–181.

Gao, Y., & Lin, J. (2018). Efficient discovery of variable-length time series motifs with large length range in million scale time series. *arXiv preprint arXiv:1802.04883*.

Gupta, M., Gao, J., Aggarwal, C., & Han, J. (2013). Outlier detection for temporal data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *26*(9), 250–2267.

Haladjian, J., Schlabbers, D., Taheri, S., Tharr, M., & Bruegge, B. (2020). Sensor-based detection and classification of soccer goalkeeper training exercises. *ACM transactions on Internet of things*, *1*(2), 1–20.

Horvat, T., & Josip, J. (2020). The use of machine learning in sport outcome prediction: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(5), e1380.

Hossain, H., Khan, M., & Roy, N. (2017). Soccermate: A personal soccer attribute profiler using wearables. In *2017 IEEE International Conference on Parvasive Computing and Communications Workshops (PerCom Workshops)*, (pp. 164–169).

Hu, X., Mo, S., & Qu, X. (2020). Basketball activity classification based on upper body kinematics and dynamic time wraping. *International journal of sport medicine*, *41*(4), 255–263.

Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. (2019). Deep learning for time series classification: a review. *Data mining and knowledge discovery*, *33*(4), 917–963.

Jamil, M., Phatak, A., Mehta, S., Beato, M., Memmert, D., & Connor, M. (2021). Using multiple machine learning algorithms to classify elite and sub-elite goalkeepers in professional men's football. *Scientific reports, 11*(1), 1-7.

Junejo, I., & Al Aghbari, Z. (2012). Using sax representation for human action recogni- tion. *Journal of Visual Communication and Image Representation*, *23*(6), 853–861.

Kelly, D., Coughlan, G., Green, B., & Caulfield, B. (2012). Automatic detection of collisions in elite level rugby union using a wearable sensing device. *Sport Engineering*, *15*(2), 81–92.

Keogh, E., & Kasetty, S. (2003). On the need for time series data mining benchmarks: a survey and empirical demonstration. *Data Mining and Knowledge Discovery*, *7*(4), 349–371.

Keogh, E., Lin, J., Lee, S., & Van Herle, H. (2006). Finding the most unusual time series subsequence: algorithms and applications. *Knowledge and Information Systems*, *11*(1), 1–27.

Keogh, E., Lin, J., & Truppel, W. (2003). Clustering of time series subsequences is meaningless: Implications for previous and future research. In *Proceedings of the third IEEE international conference on data mining, Wahington, DC: IEEE Computer Society*, (pp. 115–122).

Keogh, E., & Ratanamahatana, C. (2002). Exact indexing of dynamic time warping. *Proceedings of the 26th International Conference on Very Large Data Bases*, *7*(3), 406–417.

Li, Y., Wang, L., & Li, F. (2021). A data-driven prediction approach for sports team performance and its application to national basketball association. *Omega*, *98*(102123).

Li, Y., & Zhang, Y. (2012). Application of data mining techniques in sports training. In *5th International Conference on BioMedical Engineering and Informatics*, (pp. 954– 958).

Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition,*, *38*(11), 1857–1874.

Lin, J., Keogh, E., Lonardi, E., & Patel, S. (2002). Finding motifs in time series. In *Proceedings of the Eighth ACM SIGKDD Iternational Conference on Knowledge Discovery and Data Mining 2nd Workshop on Temporal Data Mining*, (pp. 53–68).

Lin, J., Keogh, E., Lonardi, W., & Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, ACM Press*, (pp. 2–11).

Lin, J., Keogh, E., Wei, L., & Lonardi, S. (2007). Experiencing sax: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery*, *15*(2).

Lin, J., & Li, Y. (2010). Finding approximate frequent patterns in streaming medical data. In *IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS), IEEE*, (pp. 13–18).

Linardi, M., Zhu, Y., Palpanas, T., & Keogh, W. (2018). Matrix profile x: Valmod - scalable discovery of variable-length motifs in data series. In *Proceedings of the 2018 International COnference on Management Data*, (pp. 1053–1066).

Liu, B., Li, J., Chen, C., Tan, W., Chen, Q., & Zhou, M. (2015). Efficient motif discovery for large-scale time series in healthcare. *IEEE Transactions on Industrial Informatics*, *11*(3), 583–590.

Maeda, T., Fujii, M., Hayashi, I., & Tasaka, T. (2014). Sport skill classification using time series motion picture data. In *Industrial Electronics Society, IECON 2014-40th Annual Conference of the IEEE*, (pp. 5272–5277).

Memmert, D., & Raabe, D. (2018). *Data Analytics in Football. Positional Data Collection, Modelling and Analysis.* Abingdon: Routledge.

Memmert, D., Lemmink, K. A. P. M., & Sampaio, J. (2017). Current Approaches to Tactical Performance Analyses in Soccer using Position Data. *Sports Medicine, 47*(1), 1-10.

Miller, R., Schwarz, H., & Talke, I. (2017). Forecasting sports popularity: application of time series analysis. *Academic Journal of Interdisciplinary Studies*, *6*(2), 75.

Minnen, D., Starner, T., Essa, I., & Isbell, C. (2006). Discovering characteristic actions from on-body sensor data. In *Wearable computers, 2006 10th IEEE international sym- posium on wearable computers. IEEE*, (pp. 11–18).

Mitchell, E., Monaghan, D., & O'Connor, N. (2013). Classification of sporting activities using smartphone. *Sensors*, *13*(4), 5317–5337.

Mitsa, T. (2010). *Temporal data mining*. Chapman and Hall/CRC.

Mueen, A. (2014). Time series motif discovery: dimensions and applications. *Wiley Interdiscilinary Reviews: Data Mining and Knowledge Discovery*, *4*(2), 152–159.

Oates, T., Boedihardjo, A., Lin, J., Chen, C., Frankenstein, S., & Gandhi, S. (2013). Motif discovery in spatial trajectories using grammar inference. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, (pp. 1465–1468).

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.

Pimentel, M., Clifton, D., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *In Signal Processing*, *99*, 215–249.

Ratanamahatana, C., Lin, J., Gunopulos, D., Keogh, E., Vlanchos, M., & Das, G. (2010). Mining time series data. *Data mining and knowledge discovery handbook*, (pp. 1069–1103).

Raabe, D., Nabben, R., & Memmert, D. (2022). Graph Representations for the Analysis of Multi-Agent Spatiotemporal Sports Data. *Applied Intelligence, 1-21.*

Rein, R., & Memmert, D. (2016). Big data and tactical analysis in elite soccer: future challenges and opportunities for sports science. *SpringerPlus*, *5*(1), 1–13.

Rein, R., Raabe, D., & Memmert, D. (2017). "Which pass is better?" Novel approaches to assess passing effectiveness in elite soccer. *Human Movement Science, 55*, 172–181. https://doi.org/10.1016/j.humov.2017.07.010

Schmidl, S., Wenig, P., & Papenbrock, T. (2022). Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, *15*(9), 1779–1797.

Schmidt, A. (2012). Movement pattern recognition in basketball. *Human movement science*, *31*(2), 360–382.

Schumaker, R., Soleiman, O., & Chen, H. (2010). Sports knowledge managemet and data mining. *Annual Review of Information Science and Technology*, *44*(1), 115–157.

Sempena, S., Maulidevi, N., & Aryan, P. (2011). Human action recognition using dy- namic time warping. *Proceedings on the 2011 Interantional Conference on Electrical Engineering and Informatics, ICEEI*, (pp. 1–5).

Senin, P. (2008). Dynamic time warping algorithm: review. *Information and Computer Science Department University of Hawaii at Menoa Honolulu*, *855*, 1–23.

Seto, S., Zhang, W., & Zhou, Y. (2015). Multivatiate time series classification using dynamic time warping template selection for human activity recognition. *IEEE sympo- sium series on computational intelligence*, (pp. 1399–1409).

Siirtola, P., Laurinen, P., Haapalainen, E., Roning, J., & Kinnunen, H. (2009). Clustering-based activity classification with a wrist-worn accelerometer using basic features. *2009 IEEE Symposium on Computational Intelligence and Data Mining*, (pp. 95–100).

Sivaraks, H., & Ratanamahatana, C. (2015). Robust and accurate anomaly detection in ecg artifacts using time series motif discovery. *Computational and mathematical methods in medicine*, *2015*.

Soto-Valero, C., González-Castellanos, M., & Pérez-Morales, I. (2017). A predictive model for analysing the starting pitchers' performance using time series classification methods. *International Journal of Performance Analysis in Sport*, *17*(4), 492–509.

Srivastava, R., Patwari, A., Kumar, S., Mishra, G., Kaligounder, L., & Sinha, P. (2015). Efficient characterization of tennis shots and game analysis using wearable sensor data. *2015 IEEE sensors*, (pp. 1–4).

Stein, M., Jenezko, D., H.and Seebacher, Jäger, A., Negel, J., M.and Hölsch, Kosub, S., Schreck, T., Kleim, D., & Grossniklaus, M. (2017). How to make sense of team sport data: from acquisition to data modeling and research aspects. *Data*, *2*(1).

Tanaka, Y., Iwamoto, K., & Uehara, K. (2005). Discovery of time series motif from multidimensional data based on mdl principle. *Machine Learning*, *58*(2), 269–300.

Torkamani, S., & Lohweg, V. (2017). Survey on time series motif discovery. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *7*(2).

Wang, J., Wang, Z., Gao, F., Zhao, H., Qui, S., & Li, J. (2020). Swimming stroke phase segmentation based on wearable motion capture technique. *IEEE Transactions on Instrumentation and Measurement*, *69*(10), 8526–8538.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., & Keogh, E. (2012). Experimental comparison of representation methods and distance measures for time series data. *Data Mining and Knowledge Discovery*, *26*(2), 275–309.

Wang, Z., Li, J., Wang, J., Zhao, H., Qiu, S., Yang, N., & Shi, X. (2018). Inertial sensor-based analysis of equestrian sports between beginner and professional riders under different horse gaits. *IEEE Transactions on Instrumentation and Measurement*, *67*(11), 2692–2704.

Worsey, M., Jones, B., Cervantes, A., Chauvet, S., Thiel, D., & Espinosa, H. (2020). Assessment of head impacts and muscle activity in soccer using a t3 inertial sensor and a porable electromyography (emg) system: A preliminary study. *Electronics*, *9*(5), 834.

Wu, H., & Keogh, E. (2021). Current time series anomaly detection benchmarks are flawd and are creating the illusion of progress. *IEEE Transaction on Knowledge and Data Engineering*.

Xi, X., Keogh, E., Shelton, C., Wei, L., & Ratanamahatana, C. (2006). Fast time series classification using numerosity reduction. In *Proceedings of the 23rd international conference on Machine learning*, (pp. 1033–1040).

Xing, Z., Pei, J., & Keogh, E. (2010). A brief survey on sequence classification. *ACM SIGKDD Explorations Newsletter*, *12*(1), 40–48.

Yeh, C., Kavantzas, N., & Keogh, E. (2017). Matrix profile vi: meaningful multidimen- sional motif discover. In *IEEE international conference on data mining (ICDM). IEEE*, (pp. 565–574).

Yong, W., Lingyun, P., & Jia, W. (2020). Statistical analysis and arma modeling for the big data of marathon score. *Science & Sports*, *35*(6), 375–385.

Zolhavarieh, S., Aghabozorgi, S., & Teh, Y. (2014). A review of subsequence time series clustering. *The Scientific World Journal,*, *2014*.