Journal of Intelligent & Fuzzy Systems 29 (2015) 1981–1994
DOI:10.3233/IFS-151676
IOS Press

Interestingness measure for mining sequential patterns in sports

Goran Hrovat[a,*], Iztok Fister Jr.[a], Katsiaryna Yermak[b], Gregor Stiglic[c] and Iztok Fister[a]
[a]*Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia*
[b]*Faculty of Medicine, University of Maribor, Slovenia*
[c]*Faculty of Health Sciences, University of Maribor, Slovenia*

Abstract. The increasing availabilities of tracking devices, including mobile devices and sports trackers with heart-rate monitors, accelerometers and GPS receivers, have increased the interest in developing fitness applications. The aims of these applications are to improve the motivations of athletes during training, as to track the histories of their sports activities, to advise the type of training for the future, and even to share this information with friends on social networks. This study proposes a novel method for analyzing the time series data gathered from a single athlete over an extensive time period of training. Using this method, the transformed time series data are exploited by a sequential pattern mining algorithm, then the novel trend of interestingness measures are calculated for discovering sequential patterns and finally these patterns are visualized. Essentially, the main novelty of the proposed method is significance testing for trends that serve as interestingness measures for mined sequential patterns. As a result, two types of trend plots together with glyph-based sequence charts are provided to trainers for determining the progresses of their athletes based on time periods of several months. Beside the trainers, this algorithm is also useful for amateur athletes usually preparing without trainers.

Keywords: sequence analysis, interestingness measure, data mining visualization, sport tracker, TCX

1. Introduction

Analyzing performance during sports training has always been a very complex process. Over past centuries the trainers have been the persons preparing athletes for competitions and tournaments. These trainers have been very experienced persons with broad theoretical backgrounds about sports training but always getting into difficulties when the progresses of the athletes' forms need to be monitored because they did not have any possibilities for manually dealing with these data [1]. In line with the progression of mobile and ubiquitous technologies, new possibilities have emerged for smarter sports training [2]. The primary advantage of these technologies is that they enable an athlete to store information about his training activities in so-named Training Center XML (TCX) files suitable for further analyses. These data are represented as time series of data points measured at non-uniform time intervals and contain information about the movement of an athlete during his/her sports activities. This tracked information consists of the current position, heart-rate, speed, and time stamp. The current position measured by Global Positioning System (GPS) receivers [3] are expressed according to the altitude, longitude, and latitude. These data may also be used for predicting half-marathon race time [4], where running speed is required during training.

Data gained from the tracker devices can serve for extensive analyses by trainers. For instance, the heart-rate is one of the main indicators for an athlete's endurance [5]. When used in conjunction with the

*Corresponding author. Goran Hrovat, Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia. Tel.: +386 41 947 746; E-mail: goran.hrovat@um.si.

duration of the training, it can be used as an indicator of the training intensity. Higher heart-rate over a specific period of time means higher intensity of training. This measure is suitable during interval training that represents highly intensive training with rest periods in between. On the other hand, decline in the heart-rate during a specific time period is one of the indicators for successfully implementing programs of endurance training by trainers. However, the heart-rate is related to the performances of an athlete. When this measure is increased after maximum effort, it can be an indicator of tiredness. In contrast, gradually decreasing the heart-rate at rest by an athlete can indicate that the endurance training program has been implemented well by trainer.

This type of training has a great impact on heart-rate deviation. Heart-rate deviation is the deviation of the heart rate from the average heart rate over a specific time interval. This can be calculated from the time series of heart-rates to be used as attributes for sequential pattern mining. When an athlete's heart-rate is constant, although higher during the whole training period, the heart-rate deviation is close to zero. This means, the athlete is in good form and therefore ready for competitions. On the other hand, any fluctuation of the heart-rate which relates to a high-level of heart-rate standard deviation, can indicate either a lesser trained athlete or a well-trained athlete performing interval training.

Unfortunately, the problem arises when one wants to visualize the gathered data in TCX files so that they become appropriate for trainers either to identify the current form of an athlete or to serve as an indicator of an athlete's progress. Most of the related studies based on visualizing the heart-rate from TCX files, e.g., data obtained either from running races as proposed by Oliveira et al. [6] or analyzing real-time sports performances, as proposed by Legg et al. [7].

The purpose of the proposed interdisciplinary study was to discover interesting sequential patterns from the time series data induced by the training of a professional cyclist and visualizing them in the form of glyph-based sequence charts together with daily and duration trend plots, in order to make them user-friendly. Thus, the daily trend represents a normalized number of days, where an interesting sequential pattern occurs at least once, whilst the duration trend represents the normalized time coverage of the interesting sequential pattern aggregated by months. Normalized number of days means the proportion of days when the pattern appears with regard to all training days in each month. Normalized time coverage relates to the proportion of time when a pattern appears with regard to all training time in a month. Interesting sequential patterns are measured by significance testing of trends, the results of which represents an interestingness measure. In line with this, a permutation test was used to test the difference in sequential pattern occurrence between the first half and the second half of the observed time period. The duration and daily trend plots based on several months add additional information to the discovered sequential patterns. The professional trainers can use the proposed method to monitor the athlete's progress . In addition to sports trainers, amateur athletes may also use the proposed method in order to analyze their sport activities because this can increase their motivation [8, 9].

The paper is organized as follows. Section 2 surveys the field of sequential pattern mining. Section 3 begins with a description of sports trackers and related work . Section 4 proposes the method for analyzing the sports activities using the novel interestingness measures. Section 5 presents the experiments and results showing the interesting sequential patterns for the more significant trends. Section 6 discusses meanings for some of the sequential patterns. Section 7 concludes the study and provides some directions for future research.

2. Sequential pattern mining

This section introduces a short survey of sequential pattern mining methods. Market-basket sequential pattern mining arose in the year 1995 [10] when Agrawal and Srikant extended their research of frequent association rule mining [11, 12]. Market-basket analysis means to mine interesting patterns within customers' purchases, where the purchases of transactions consist of products called items. Sequential pattern mining is to discover all sequential patterns with predefined minimum support within a database of transactions where transaction times and customer identities are present regarding the transactions. Suppose we have a database D that is a set of transactions $\{t_1, t_2, ..., t_j\}$ and set of items $I = \{i_1, i_2, ..., i_k\}$. Each transaction has customer id, transaction time and item-set $X \subset I$. The quantity of items within the item-set is unknown, only information about the item's presence. The size of item-set $|X|$ is the number of items within the item-set. Database D has as many sequences as the number of unique customer ids. A sequence $s = < s_1, s_2, ..., s_m >$ is constructed from transactions with the same customer id. Elements of sequence s_1 to s_m are item-sets of transactions ordered by the transaction time. The length of a sequence is defined as

$l = \sum_{i=1}^{m} |s_i|$ and such a sequence is called l-sequence. The size of a sequence m is the number of elements within a sequence. Sequence $a = < a_1, a_2, ..., a_n >$ is a subsequence of sequence $b = < b_1, b_2, ..., b_m >$ and b is a super-sequence of a if there exist integers $i_1 < i_2 < ... < i_n$ such that $a_1 \subset b_{i_1}, a_2 \subset b_{i_2}, ..., a_n \subset b_{i_n}$ and $n \leq m$. A support of a sequence s, denoted as $sup(s)$, is the number of sequences within the database, where the sequence s is a subsequence. A frequent sequential pattern fs is a sequence that satisfies user predefined minimum support $sup(fs) \geq minSupp$. A maximal sequential pattern is a sequence with no super-sequences within the database. A closed sequential pattern is a sequence with no super-sequences within the database having the same support.

In 1995, Agrawal and Srikant [10] proposed three algorithms for the problem of mining sequential patterns: AprioriAll, AprioriSome and DynamicSome. In [13], a new algorithm "Generalized Sequential Patterns (GSP)" is proposed which performs much faster and is generalized to solve a wider range of problems. Time constraints have been added that specify minimum and maximum time periods between adjacent elements within a pattern, a sliding-time window has been introduced that allows for items within an element of a sequential pattern to be present within two different transactions, and taxonomies have been considered. The proposed method is based on the generation of candidates that is followed by pruning, where frequent sequences of length k are used to generate $k + 1$ length sequences. Masseglia et al. [14] continued their work and proposed a new algorithm PSP based on GSP but by efficiently storing candidates and sequences. Han et al. [15] introduced FreeSpan which significantly reduces candidate subsequence generation with the use of projected sequence databases. Pei et al. [16] improved FreeSpan and further reduced candidate generation with the proposed algorithm PrefixSpan, which examined only the prefix subsequences and projects only their corresponding postfix subsequences into projected databases. Zaki [17] proposed SPADE that improves the sequence mining task by reducing the database scans needed to find all the frequent sequential patterns to only three. Ayres et al. [18] presented Sequential PAttern Mining (SPAM) and further improved mining sequential patterns, especially when sequential patterns within that database are very long. SPAM uses a depth-first search, with a vertical bitmap representation of the database. Such representation is extremely efficient for support counting but is quite space inefficient in comparison to SPADE. SPAM was further improved by Yang et al. [19] with LAst Position INduction Sequential PAttern Mining (LAPIN-SPAM) which exploits the fact that if an item's last position is smaller than the current prefix position, the item can not appear behind the current prefix within the same sequence. Many other sequential pattern mining algorithms have emerged which are optimised only for mining closed sequential patterns [20–22].

The presence of several algorithms for mining sequential patterns bears witness to its importance in many applications. Pitman et al. [23] extended the Bi-Directional Extension (BIDE) algorithm [21] for landing page personalization and product recommendation within the e-commerce domain. Baralis et al. [24] proposed an approach based on the extraction of frequent closed sequences for analyzing the exam log data of diabetic patients in order to identify those sets of exams that are frequently done together and the sequences of exam sets frequently performed by patients. Yung-Piao et al. [25] presented a model for predicting the stock-trend based on a combination of sequential chart pattern, K-Means [26], and the AprioriAll [10] algorithm. They used only attribute price for mining sequential patterns. The correlation between sequential patterns was used to predict the movement of shares (up or down). In our case, we mine sequential patterns, using multiple attributes. The proposed new interestingness measure is then used to rank mined patterns, which can be used for monitoring the progress of the athlete. Yap et al. [27] introduced a personalized sequential pattern mining-based recommendation framework and Competence Score measure for improving next-item recommendation for the target user. Concaro et al. [28] exploited the SPAM [18] algorithm over the Nationwide Inpatient Sample (NIS) [29] database, in order to find frequent sequential patterns of diagnoses shared by US hospitals. Hu et al. [30] introduced three constraints - compactness, repetition and recency - and considered them jointly with frequency in selecting sequential patterns. Their method is useful in extracting sequential patterns within a B2B environment.

3. Sports trackers and sport activities

Sports trackers are mobile devices that enable users to track data during their sports activities. Actually, they are a combination of hardware and software. The hardware of these mobile devices consists of miscellaneous sensors, whilst the software is comprised of

Fig. 1. Garmin Forerunner watch.

⟨Trackpoint⟩
⟨Time⟩**2011-07-23T09:30:54.000Z**⟨/Time⟩
⟨Position⟩
⟨LatitudeDegrees⟩**45.05720363929868**⟨/LatitudeDegrees⟩
⟨LongitudeDegrees⟩**6.034371312707663**⟨/LongitudeDegrees⟩
⟨/Position⟩
⟨AltitudeMeters⟩**727.4000244140625**⟨/AltitudeMeters⟩
⟨DistanceMeters⟩**0.6700000166893005**⟨/DistanceMeters⟩
⟨HeartRateBpm⟩
⟨Value⟩**80**⟨/Value⟩
⟨/HeartRateBpm⟩
⟨Extensions⟩
⟨TPX⟩
⟨Speed⟩**2.2209999561309814**⟨/Speed⟩
⟨/TPX⟩
⟨/Extensions⟩
⟨/Trackpoint⟩
⟨Trackpoint⟩
...
⟨/Trackpoint⟩

Fig. 2. Example of Trackpoint element in TCX file.

an operating system and applications. Several applications have already been developed in order to track the sports activities e.g., Endomondo, Runtastic. However, sport tracking is not just limited to mobile devices only. There are also more domain-specific devices dedicated to professional as well as amateur athletes. For example, Garmin developed a special sport devices named Forerunner, shown in Fig. 1, that are tailored as watches appropriate for wearing by athletes during their sports activities. Such sports watches have many useful functions, e.g., measuring the duration of a training session, average heart-rates, wasted calories, and the movements of an athlete's temperature during the sports activities. Speed is tracked by the GPS receivers. Trackers usually save data in TCX files, which are also becoming popular within the research community [6]. An example of the TCX file is shown in Fig. 2.

It can be seen from Fig. 2 that the TCX file contains a set of tracked points embraced by tags ⟨Trackpoint⟩...⟨/Trackpoint⟩ which consists of data, like: the time-stamp, the latitude (in degrees), the longitude (in degrees), the altitude (in meters), the distance from the last measured position (in meters), the heart-rate (in beats per minute) and the speed (in meters per second). All the mentioned data representing features in this study are presented as bold in the figure.

With the mass production of wearable mobile devices, ubiquitous computing is also becoming important within the field of sports analytics. A related study was proposed by Oliveira et al. [6], where heart-rate temporal information in TCX data obtained from a single race was visualized for comparisons between multiple runners. Legg et al. [7] analyzed the real-time sports performances of matches and visualized them within an interactive glyph-based chart. Li et al. [31], thus proposing the idea of using ubiquitous computing technology for creating a comprehensive system that monitors people's exercise habits and health and provides information for motivating and helping them tailor their activities. A MObile PErsonal Trainer (MOPET) was proposed by Buttussi et al. [8]. The MOPET is a wearable system based on real-time data coming from sensors, providing expert knowledge and a periodically updated user model. It motivates users and provides safety and health advice whilst alternating jogging and fitness exercises during outdoor environments. Tapia et al. [32] presented an algorithm for real-time recognition of physical activities and their intensities using five triaxial wireless accelerometers and a wireless heart-rate monitor. Wylie et al. [33] presented a general health monitoring software for mobile phones called Heart Angle with the purpose of helping users monitor, record, and improve their fitness levels through built-in cardio-respiratory tests, a location tracking application for analyzing heart-rate exertion over time and location. Pernek et al. [34] predicted heart-rate during a sport activity using different combinations of acceleration features on a smart phone device. Reddy et al. [35] presented Biketastic, a platform for cyclists where participants can share routes, ride statistics, sense information to infer route roughness and noise, and media that documents the cyclist's experience when cycling. Speed and location information is obtained using a GPS sensor and the physical dynamics of the route are documented using an accelerometer and microphone so road roughness and noise level along a route can be inferred.

4. Analyzing the sequential patterns' trends for cycling

This section introduces a novel method for analyzing the sequential patterns in TCX data obtained by professional cyclists over longer periods of time. However, this method can usefully be also applied for other sports' disciplines, where TCX data can be produced. Sequential patterns' temporal information need to be analyzed in order to explore sequential patterns' trends, which can reflect a cyclist's form progression over a longer period of time. Some preprocessing is necessary before preparing the data for sequential pattern mining. Our proposed method is focused on discovering the sequential patterns with increasing and decreasing trends. This means, the sequential pattern has an increasing trend when it occurs more frequently in the second half of the training period and has decreasing trend when it occurs more frequently in the first half of the training period. The proposed method consists of the following four phases:

1. transforming the data (preprocessing),
2. mining the sequential patterns on transformed data,
3. calculating the trends' interestingness measures for frequent sequential patterns and
4. visualizing interesting sequential patterns for increasing and decreasing trends.

The remainder of this paper illustrates these phases in detail.

4.1. Transforming the data

This subsection proposes a transformation process for preparing the data in TXC format to be appropriate for sequential pattern mining.

Initially, cycling data within TCX format are grouped according to the days of creation. For each day, activity data are divided into certain predetermined intervals (e.g., 5 min), during which features are identified. The following features are taken into account due to their availabilities, in this study: a cycling speed (Speed), a heart-rate (HR), and an altitude (Alt).

However, different values of the mentioned features can be extracted from the TCX files. In line with this, characteristics of these features are determined using the standard statistical measurements, like average (avg), minimum (min) and maximum values (max), and standard deviation (stdev). For instance, the feature Speed is fully determined by its minimum (min.Speed), maximum (max.Speed), and average (avg.Speed) values together with its standard deviation (stdev.Speed), during each interval.

Table 1
Discretized features with levels and intervals

Feature	Low level	Medium level	High level
Average speed	[0.00, 5.06)	[5.06, 7.48)	[7.48, 33.45]
Heart-rate standard deviation	[0.0, 5.8)	[5.8, 13.5)	[13.5, 50.0]
Altitude standard deviation	[0.00, 8.28)	[8.28, 21.09)	[21.09, 92.11]

Selection of the discretization method is essential for sequential pattern mining. In this study, the features are discretized into three levels of equal-frequencies [36], i.e., each level contains the same number of values. The discretization by equal-frequency prevent little or too much instances in one of the category, because the distribution of values for each feature (e.g., heart-rate) is not uniform. In such a case the prevalent category (e.g., min.HR=[37,93]) would therefore almost always occurs in mined sequential patterns. An example of discretized features with levels and intervals are illustrated in Table 1.

Fig. 3 shows an example of transformed data, where each of the first two five minute intervals and some of the features' values can be seen for the N-training days. Data in this form are ready for the next phase of the proposed method.

4.2. Mining the sequential patterns in the transformed data

In this phase, the sequential pattern mining algorithm is applied to the transformed data. The sequential pattern mining algorithm requires the following parameters:

- minimal support,
- maximal size of the elements within a sequence,
- minimal and maximal size of the sequence, and
- maximal gap between elements in the sequence.

The characteristics of the parameters are as follows. Too small a support can lead to sequential patterns with lower p-value due to the lack of its occurrence in one half of trend vector. Here, the p-value obtained as a result of the permutation test and, is explained later in the section, represents a trend interestingness measure for sequential patterns. The minimal size of the sequence ensures the mining of sequential patterns,

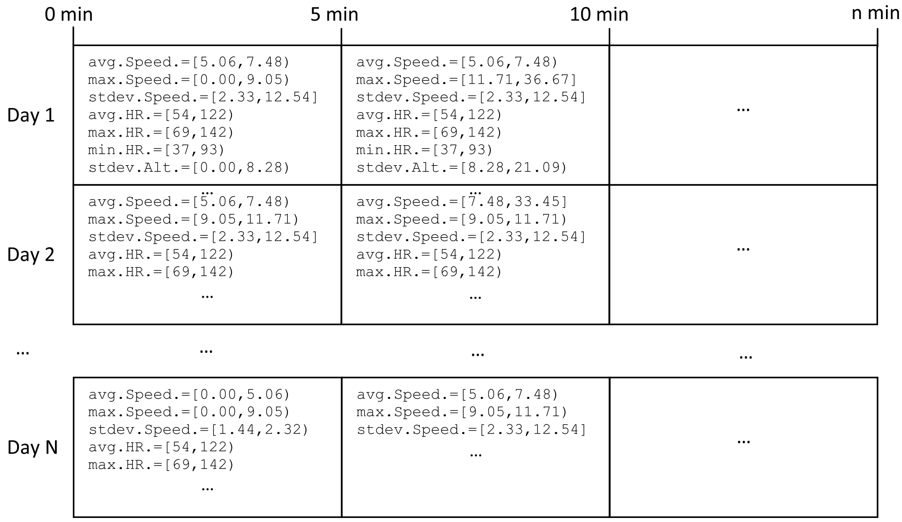

Fig. 3. Example of transformed data used for mining sequential patterns.

```
  items                                support
1 <{stdev.HR.=[5.8, 13.5)},
   {avg.Speed.=[0.00, 5.06)},
   {stdev.Alt.=[21.09, 92.11]}>         0.6835
```

Fig. 4. Example of sequential pattern in text format and its support.

whilst sequential patterns of length 1 are actually those item-sets which can be obtained by other mining algorithms (e.g. Apriori [10]). The maximal sequential pattern length represents the upper boundary limiting the execution time of sequential pattern mining. The maximal gap between events within the sequence is set to 1, which means that the rest periods in the sequence are excluded.

Indeed, the days are considered as time sequences and five minute intervals as events. Rest periods of more than 30 seconds are noted as time periods of length 2 between events. In line with this, events in sequential patterns are set to be adjacent (maximal gap between two adjacent events set to 1), i.e., mined sequential patterns do not include rest periods of more than 30 seconds. As a result, the discovered sequential patterns are candidates for further analysis. An example of discovered sequential pattern in raw text format and its support is shown in Fig. 4.

As can be seen from Fig. 4, the cyclist had the first five minutes heart-rate standard deviation at interval [5.8, 13.5) (Medium level), the next five minutes average speed at interval [0.00, 5.06) (Low level), and the last five minutes altitude standard deviation at interval [21.09, 92.11] (High level). This sequential pattern occurred in 68.35% of training days as determined by the support, calculated during sequential pattern mining.

4.3. Calculating the trend interestingness measures for frequent sequential patterns

This subsection introduces the essential part of sequential patterns analysis, the novel interestingness measures, which are the important contribution in ranking sequential patterns.

During this phase, trend interestingness measures are calculated for all the sequential patterns. Here, training data are grouped by months. As a result, each monthly report contains the number of days where the specific sequential pattern occurred and the total number of its occurrences during the whole analysis. Indeed, two different trends are identified: duration and daily. The former is expressed as a normalized total time of the sequential pattern occurrences, whilst the latter as the normalized number of days, where this pattern occurs.

Let sp denote a sequential pattern, k the number of training months, t_{sp} a sequential pattern duration (e.g., 15 min), t_i total training time in month i, d_i the number of days in a month i, $f_i(sp)$ frequency of sp in month i, $f_{i_j}(sp)$ frequency of sp in j-th day of i-th month and $supp_i(sp)$ a support of sequential pattern sp in month i, where a day is considered as a sequence, in this study. The duration trend vector $\overrightarrow{dut}$ is based on the fact that a sequential pattern can occur more than once within a sequence, in other words

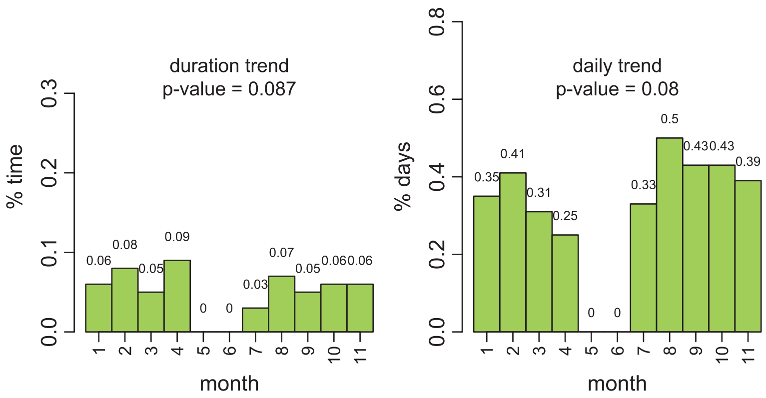

Fig. 5. Example of duration and daily trend.

$$f_i(sp) = \sum_{j=1}^{d_i} f_{i_j}(sp).$$

Normalized duration of sp in i-th month is defined as

$$dur_i(sp) = \frac{t_{sp} \cdot f_i(sp)}{t_i},$$

whilst the duration trend vector as

$$\overrightarrow{dut} = \langle\, dur_1(sp),\ dur_2(sp),\ \ldots,\ dur_k(sp)\,\rangle.$$

On the other hand, the daily trend vector is defined as

$$\overrightarrow{dat} = \langle\, supp_1(sp),\ supp_2(sp),\ \ldots,\ supp_k(sp)\,\rangle.$$

For both trend vectors, $\overrightarrow{dut}$ and $\overrightarrow{dat}$, a one-tailed permutation test is performed for each sequential pattern. Months with a lack of training days are excluded from the permutation test in order to prevent undesirable influence on the calculation. The permutation test is performed using 10,000 permutations, where the difference between the sample means of the first half and the second half of the trend vector was tested. The null hypothesis states that the mean of one half of the trend vector is not bigger than the other half. A calculated p-value obtained as the result of the permutation test represents a significance value between the sequential pattern occurrence in the first half and the second half of a training period (expressed in months). Thus, p-value trend interestingness measure is defined obtained as the result of a permutation test.

Computational complexity for calculating the novel interestingness measure for one sequential pattern depends on the number of training days d and is $O(d)$. It is then calculated for all n sequential patterns returned by an algorithm of sequential pattern mining. Overall computational complexity for calculating the interestingness measures for all n sequential patterns is therefore $O(n \cdot d)$.

An example of temporal trends regarding the sequential pattern occurrences can be seen on the diagrams in Fig. 5, where the decreasing duration trend and increasing daily trend are shown, together with the p-values. In this example, the fifth and sixth months of training are excluded from the calculation, as the cyclist in each of them trained only one day. The permutation test is performed between the first four months and the last five months.

4.4. Visualizing the interesting sequential patterns for increasing and decreasing trends

Finally, the interesting sequential patterns with their temporal trends are visualized. In this study, we considered lower p-value as the more interesting because it clearly reflects the change of sequential pattern occurrence over time. For sequential pattern visualization, we propose a sequence chart with bars reflecting low, medium and high levels of attributes. In fact, this sequence chart is the glyph-based visualization method as introduced in [37]. Glyphs are graphical entities suitable for the visualization of attributes and data values [38, 39], and thus improve user's experience.

The symbols used for variables and levels on the sequence charts are shown in Fig. 6, where the levels' symbols are represented for low, medium, and high intensity levels, and the symbols for variables are shown the cyclist's speed (maximum speed, speed standard deviation, average speed), its heart-rate (maximum heart-rate, minimum heart-rate, average heart-rate, heart-rate standard deviation), and its altitude standard deviation.

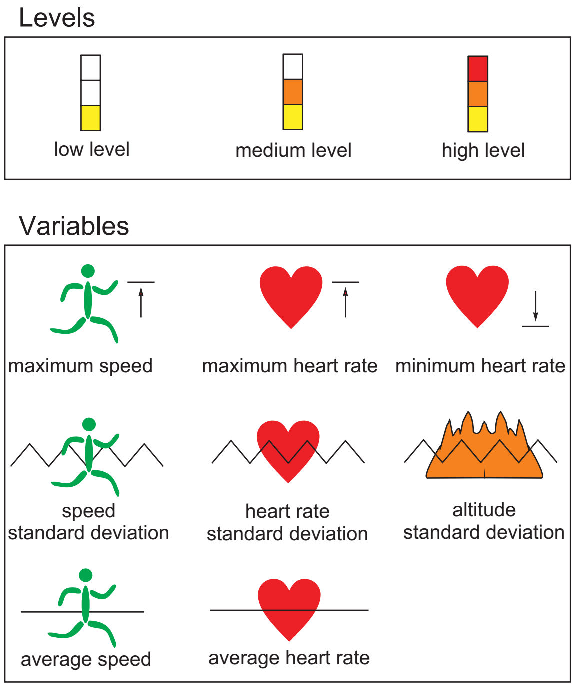

Fig. 6. Symbols for variables and levels.

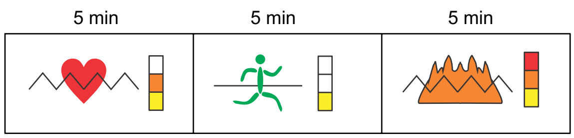

Fig. 7. Example of sequence chart.

An example of a sequence chart representing the sequential pattern from Fig. 4, is shown in Fig. 7. Here the levels are taken from Table 1. As can be seen from the figure, the cyclist had a first 5 minutes heart-rate standard deviation at the medium level, the next 5 minutes average speed at the low level, and the last 5 minutes altitude standard deviation at the high level.

5. Experiments and results

The aim of our experimental work was to show that sequential pattern mining could also be useful on non market-basket types of data, which is not as trivial for discovering sequential patterns. Data analysis is performed on the basis of data gathered by a real cyclist using the Garmin Forerunner sports watch over a longer period of time (e.g., one year). The purpose of the experiments was to show that these cycling data can be used with the sequential pattern mining algorithm to discover interesting sequential patterns during training data. The proposed method was experimentally applied to the TCX dataset in order to discover the sequential patterns with the more significant trends, i.e., with the lowest trend interestingness measure. Only speed, heart-rate, altitude, and time attributes were selected for the mining process. Such sequential patterns together with trend plots are suitable for determining the form progress of a cyclist during the training period.

The parameters during the experiments were set as follows. The minimal support was set at 0.3, the maximal size of the elements within a sequence at 3, the minimal and maximal sizes of sequence at 3 and 10, respectively, and the maximal gap between elements in the sequence at 1. This means that rest periods in the sequences are excluded. All parameters were determined empirically in order to produce promising results on the selected datasets (61,046 sequential patterns were discovered for the selected parameters). In the experiments, the real TCX datasets obtained from professional cyclist between the 4th of July 2011 and 24th of May 2012 were taken into account.

For data processing, we used language R [40] of Version 3.0.2 with the packages arulesSequences 0.2-6 [41] that implements SPADE with constraints, arules 1.1-5 [42] for the purpose of discretization, zoo 1.7-11 [43] for time and date transformation, stats 3.0.2 [40] for linear model estimation, and sqldf 0.4-7.1 [44] for more convenient data manipulation. Experiments were conducted on the MS Windows 7 operating system.

Variables in the dataset were discretized by equal-frequency as shown in Table 2. In fact, four experiments were conducted in which the sequential patterns were searched for:

– the more significantly increasing duration trends,
– the more significantly increasing daily trends,
– the more significantly decreasing duration trends, and
– the more significantly decreasing daily trends.

In each of the experiments justification for one sequential pattern was provided. The other sequential patterns were interpreted the same within each experiment. The following interpretations for all four types of patterns applied. Sequential patterns with the more significantly increasing duration trends represent patterns which are hard to keep the more often during training time. Sequential patterns with the more significantly increasing daily trends represent patterns which are hard to achieve only once during the whole training day. On the other hand, sequential patterns with the

Table 2
Levels of variables in dataset discretized by equal-frequency

Variables	Low level	Medium level	High level
Average speed	[0.00, 5.06)	[5.06, 7.48)	[7.48, 33.45]
Maximal speed	[0.00, 9.05)	[9.05, 11.71)	[11.71, 36.67]
Speed standard deviation	[0.00, 1.44)	[1.44, 2.33)	[2.33, 12.54]
Average heart-rate	[54, 122)	[122 ,139)	[139, 181]
Maximal heart-rate	[69, 142)	[142, 159)	[159, 200]
Minimal heart-rate	[37, 93)	[93, 116)	[116, 179]
Heart-rate standard deviation	[0.0, 5.8)	[5.8, 13.5)	[13.5, 50.0]
Altitude standard deviation	[0.00, 8.28)	[8.28, 21.09)	[21.09, 92.11]

Table 3
Top 3 sequential patterns of at least size 3 with the more significantly increasing duration trends

	Sequential pattern	p-value	Support
S1	{min.HR.=[116, 179]}, {avg.Speed.=[5.06, 7.48)}, {min.HR.=[93, 116)}	0.0012	0.3525
S2	{avg.HR.=[139, 181)}, {avg.Speed.=[5.06, 7.48)}, {avg.HR.=[122, 139)}	0.0023	0.3381
S3	{min.HR.=[37, 93)}, {avg.HR.=[122, 139)}, {max.HR.=[142, 159), min.HR.=[116,179]}	0.0025	0.3094

more significantly decreasing duration trends represent patterns for which we need several months of training to maintain them less frequently. Similarity applies for those sequential patterns with the more decreasing daily trend, when it is difficult to eliminate them during the training day. Duration trends and daily trends can also be interpreted together. If the significantly increasing duration trend is more significant than the increasing daily trend, this means that is easier to achieve this sequential pattern once than several times per training day at the initial beginning of the training period.

In the next subsections, those typical sequential patterns are presented that were discovered during tests.

5.1. *Sequential patterns with the more significantly increasing duration trends*

In this experiment, we mined for sequential patterns with the more significantly increasing duration trends, which are normally hard to keep at the same level. The results are presented in Table 3, where the top three sequential patterns (i.e., S1, S2, and S3) of a size of at least 3 and the more significantly increasing duration trends are shown.

Fig. 8 shows a sequential pattern S3, where in the first five minutes the cyclist contained low level of minimal heart-rate, the second five minutes had a medium

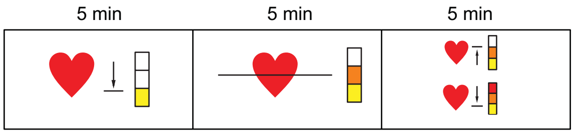

Fig. 8. Sequence chart for sequential pattern S3.

a level of average heart-rate, and the third five minutes medium level of maximal heart-rate as well as high level of minimal heart-rate. This sequential pattern occurred in 35.25% of the training days as determined by support value from Table 3. In Fig. 9, the trend plots for sequential pattern S3 can be seen, where it is noticed that this sequential pattern never appeared during the fifth and sixth months, namely in November and December, due to lack of training. These two months were not considered when calculating the trend interestingness measure. The increasing trend shows that the sequential pattern S3 typically describes the progress of the cyclist. The increasing daily trend represents a trend where the pattern occurs at least once over more and more training days. The increasing duration trend represents a trend where the pattern occurs more and more during the whole training time. Therefore in the case where the duration trend is more significant than the daily trend it is easier to achieve S3 once during a training day. Several months of training is required to achieve S3 more than once in the training day. The third five minutes interval is especially interesting because it contains the medium level of maximal heart-rate as well as high level of minimal heart-rate at the same time. Consequently the cyclist kept his/her heart-rate in that interval within the range of 116-159. This sequential pattern indicates decreasing of the heart-rate that is typically for athletes in training.

5.2. *Sequential patterns with the more significantly increasing daily trends*

The goal of this experiment was to discover the sequential patterns with the more significantly

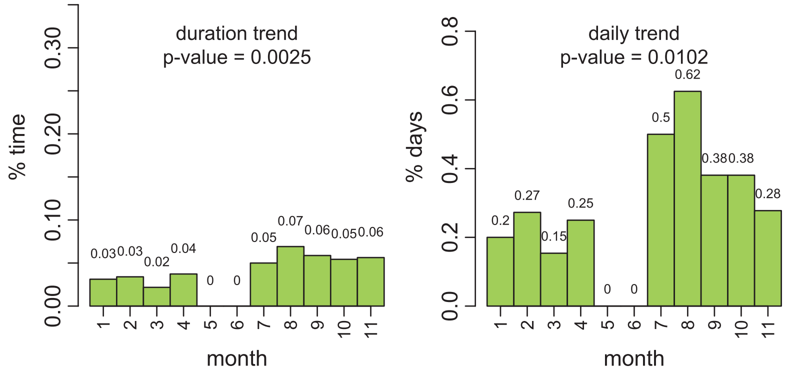

Fig. 9. Trend plots for sequential pattern S3.

Table 4
Top 3 sequential patterns of at least size 3 with the more significantly increasing daily trends

	Sequential pattern	p-value	Support
S4	{avg.HR.=[122,139), stdev.HR.=[13.5,50.0]}, {stdev.HR.=[13.5, 50.0], stdev.Speed.=[2.33, 12.54]}, {avg.HR.=[139, 181]}	0.0013	0.3094
S5	{avg.HR.=[139, 181]}, {avg.HR.=[122, 139) stdev.HR.=[13.5, 50.0]}, {stdev.Speed.=[2.33, 12.54]}	0.0014	0.3453
S6	{avg.HR.=[122, 139), stdev.HR.=[13.5, 50.0]}, {stdev.Speed.=[2.33, 12.54]}, {avg.HR.=[139, 181]}	0.0014	0.3237

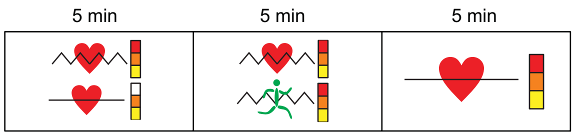

Fig. 10. Sequence chart for sequential pattern S4.

Table 5
Top 3 sequential patterns of at least size 3 with the more significantly decreasing duration trends

	Sequential pattern	p-value	Support
S7	{stdev.Alt.=[8.28, 21.09)}, {stdev.HR.=[0.0, 5.8)}, {stdev.HR.=[0.0, 5.8), max.Speed.=[0.00, 9.05), stdev.Speed.=[0.00, 1.44)}	0.0024	0.3237
S8	{stdev.Alt.=[8.28, 21.09)}, {stdev.HR.=[0.0, 5.8)}, {max.Speed.=[0.0, 9.05), stdev.Speed.=[0.00, 1.44)}	0.0032	0.3741
S9	{stdev.Alt.=[8.28, 21.09)}, {stdev.HR.=[0.0, 5.8)}, {stdev.HR.=[0.0, 5.8), max.Speed.=[0.00, 9.05)}	0.0033	0.3309
		0.0033	0.3309

increasing daily trends that is typically hard to achieve, in practice. The results of the experiments are presented in Table 4, where the top three sequential patterns (i.e., S4, S5, and S6) of size at least 3 with the more significantly increasing daily trends are shown.

In Fig. 10 a sequential pattern S4 is shown, where during the first five minutes a cyclist contained a high level of heart-rate standard deviation and also a medium level of average heart-rate, the second five minutes had a high level of heart-rate standard deviation and a high level of speed standard deviation, and the third five minutes a high level of average heart-rate. Fig. 11 shows the trend plots for sequential pattern S4. Typically, this sequential pattern appears during later period of trainings, when the specific cyclist is near to his/her optimal form.

This sequential pattern is typical for interval training which contains high levels of heart-rate standard deviation and high levels of speed standard deviation.

5.3. Sequential patterns with the more significantly decreasing duration trends

The aim of this experiment was to discover sequential patterns with the more significantly decreasing duration trends that are typically hard to maintain less frequently. The results of this experiment are presented in Table 5, where the top three sequential patterns (i.e., S7, S8, and S9) of sizes at least 3 with the more significantly decreasing duration trends are shown.

Sequential pattern S7 is shown in Fig. 13, where during the first five minutes, the altitude standard deviation was at medium level, the second five minutes the

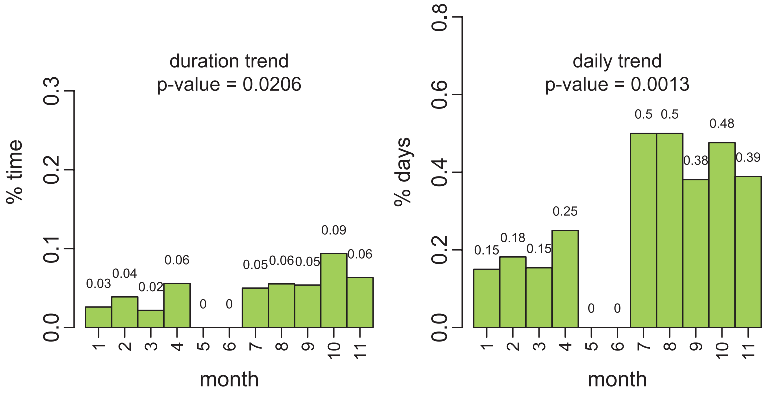

Fig. 11. Trend plots for sequential pattern S4.

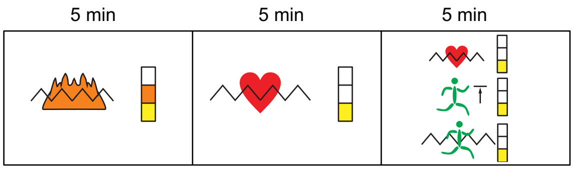

Fig. 12. Sequence chart for sequential pattern S7.

cyclist had a low level of heart-rate standard deviation, and the third five minutes had a low level of maximal speed, speed standard deviation and also heart-rate standard deviation. Fig. 12 contains the trend plots for sequential pattern S7. Typically, this sequential pattern appears less within training days after a lot of training activity.

This sequential pattern is typical for untrained athletes on steady state training, where low levels of speed standard deviation and also low levels of heart-rate standard deviation normally occur.

5.4. Sequential patterns with the more significantly decreasing daily trends

In this experiment, we looked for the sequential patterns with the more significantly decreasing daily trends, which are hard to avoid, in practice. The results of the experiments are presented in Table 6, where the top three sequential patterns (i.e., S10, S11, and S12) of sizes at least 3 with the more significantly decreasing daily trends are shown.

The sequential pattern S12 is shown in Fig. 15, where during the first five minutes, the cyclist had a low level of heart-rate standard deviation, next five minutes he had a low level of maximal speed as well as a low level of speed standard deviation, and the next five minutes a low level of heart-rate standard deviation. Fig. 14 shows the trend plot for sequential pattern S12. This sequential pattern appears within less training days and also less within each training day after a lot of training activity.

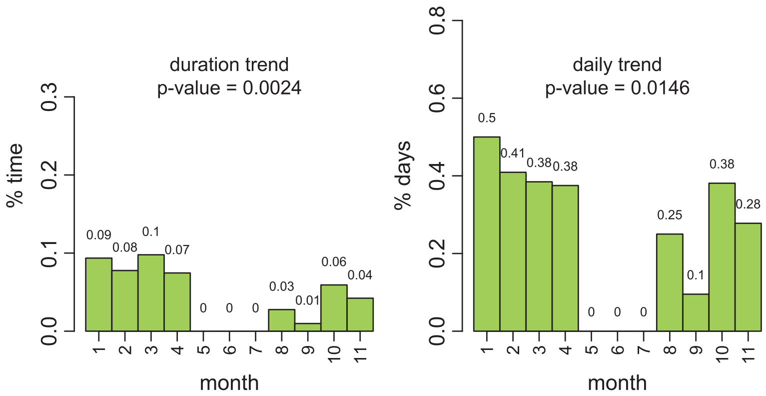

Fig. 13. Sequence chart for sequential pattern S7.

Table 6
Top 3 sequential patterns of at least size 3 with the more significantly decreasing daily trends

	Sequential pattern	*p*-value	Support
S10	{stdev.HR.=[0.0, 5.8)}, {max.Speed.=[0.00, 9.05)}, {stdev.HR.=[0.0, 5.8)}	0.0033	0.4388
S11	{max.Speed.=[9.05, 11.71)}, {stdev.HR.=[5.8, 13.5)}, {avg.Speed.=[5.06, 7.48), stdev.HR.=[5.8, 13.5)}	0.0035	0.3093
S12	{stdev.HR.=[0.0, 5.8)}, {max.Speed.=[0.00, 9.05), stdev.Speed.=[0.00, 1.44)}, {stdev.HR.=[0.0, 5.8)}	0.0036	0.4101

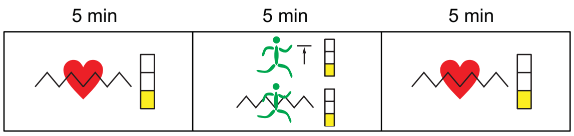

Fig. 14. Sequence chart for sequential pattern S12.

Typically, this sequential pattern is characteristic of untrained athletes because it contains low levels of maximal speed, low levels of speed standard deviation, as well as low levels of heart-rate standard deviation. In contrast, a high level of maximal speed or speed standard deviation is characteristics of trained athletes.

6. Discussion

In our study the TCX data obtained by a single cyclist collected over several months were used to mine frequent sequential patterns. Duration and daily trend interestingness measures were proposed and some sequential patterns with the more significantly increasing and decreasing trends were visualized and interpreted. The trend analysis during this study was based on the idea of discovering the trends as proposed in [45], where trends for interestingness measures of association rules were contrasted between different subgroups of patients. The sequential patterns discovered with the proposed technique are interesting for sports analysts and are promising.

After six-months of regular endurance training of moderate intensity the decrease of heart-rate is common and evident [46]. The decrease indicates that the heart has achieved higher endurance through training [46]. A trained heart performs less work with lowering heart-rate but altering stroke volumes for the same absolute workload. Such adaptation occurs because of certain training induced changes of cardiac function and the autonomic system. Carter et al. [47] explained that heart-rate is controlled by an autonomic nervous system with parasympathetic and sympathetic activities [48]. Sympathetic activity is dominant during rest and short-term training and leads to increase in heart-rate. Long-term endurance training enhances parasympathetic activity which decreases heart-rate. Endurance trained athletes have a decreased submaximal exercise heart-rate due to reduced sympathetic activity and increased parasympathetic activity for a given sub-maximal work.

A sequential pattern S3, which had significantly increasing duration trend (*p*-value = 0.0025) is a good indicator of heart function adaptation. The decrease in minimal heart-rate in the first five minutes of the sequential pattern indicates a more economical response during less strenuous exercise. The second five minutes interval indicates the middle level of average heart-rate; however the third five minutes interval shows

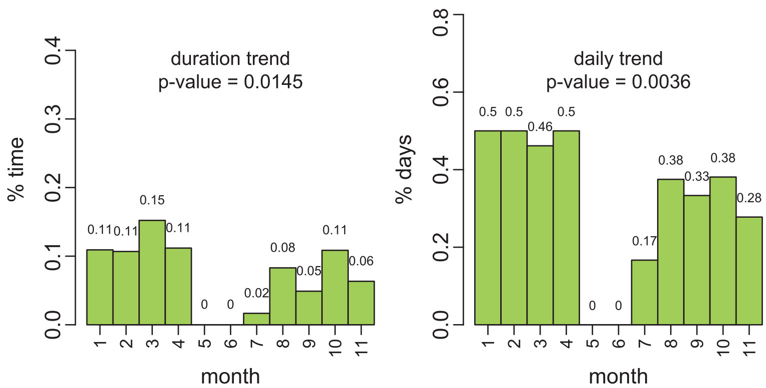

Fig. 15. Trend plots for sequential pattern S12.

the high level of minimal heart-rate and the middle level of maximal heart-rate (116-159 BPM). In moderately high-intensity exercise, adaptation of the heart function leads to non extreme peaks in heart-rate that is more typical for trained athletes.

A sequential pattern S10, which has significantly decreasing daily trend (p-value = 0.0033) is therefore less and less common. The first and third five minutes intervals contain low levels of heart-rate standard deviation and the second five minutes interval contains low levels of maximal speed. This pattern is typical for non-adapted heart and non-trained athletes.

7. Conclusion

In the presented work, sports data obtained using a tracking device and heart-rate monitor were used for discovering interesting sequential patterns and the novel trend interestingness measure p-value was introduced obtained as a result of permutation test. Duration and daily temporal trends have been presented together with sequence charts to present sequential patterns for users in more convenient and user-friendly ways. Daily trend with a significantly increasing trend characterizes the sequential pattern which is hard to achieve. On the other hand, the duration trend with a significantly increasing trend characterizes the sequential pattern which is more often hard to maintain, but not necessarily hard to achieve for beginners.

This information may be useful for trainers when monitoring the progress of an athlete. According to the progress, trainers may decide if their athletes are prepared for a competition. Their mined significant sequential pattern should contain items typical for profession athletes (e.g. significantly increasing trends for sequential patterns contain middle levels of average heart-rate). Amateur and hobby athletes may use the presented technique for analyzing their progress, to share sport results with friends and compare these with others, which could motivate them to be more concerned about sports activities [9]. Finally, the proposed technique can also be applied to other sports where wearable tracking and other mobile devices can be applied e.g., skiing, football, basketball, hockey, motor sports and even swimming [49].

More features can be selected for further research, e.g., calories burned, ambient temperature, humidity. The idea of sliding windows might be considered to discover sequential patterns not only over predetermined intervals. More data from different professional users could be analyzed and significant sequential patterns typical for them would be mined. The same sequential patterns for different users could be compared according to their trends, where common attribute intervals need to be determined.

References

[1] A. Holzinger, Human-computer interaction and knowledge discovery (hci-kdd): What is the benefit of bringing those two fields to work together? *In Availability, Reliability, and Security in Information Systems and HCI*, Springer, 2013, pp. 319–328.

[2] I. Fister, D. Fister, S. Fong and I. Fister Jr., Widespread mobile devices in applications for real-time drafting detection in triathlons, *Journal of Emerging Technologies in Web Intelligence* **5**(3) (2013), 310–321.

[3] R. Prasad and M. Ruggieri, *Applied Satellite Navigation Using GPS, GALILEO, and Augmentation Systems*. Artech House, Boston, US, 2005.

[4] B. Knechtle, U. Barandun, P. Knechtle, M.A. Zingg, T. Rosemann and C.A. Rüst, Prediction of half-marathon race time in recreational female and male runners, *Springer Plus* **3**(1) (2014), 248.

[5] A.J. Blazevich, Sports biomechanics: The basics: Optimising human performance. A&C Black, 2013.

[6] G. Oliveira, J. Comba, R. Torchelsen, M. Padilha and C. Silva, Visualizing running races through the multivariate time-series of multiple runners. *In Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI Conference on*, 2013, pp. 99–106. IEEE.

[7] P.A. Legg, D.H.S. Chung, M.L. Parry, M.W. Jones, R. Long, I.W. Griffiths and M. Chen, Matchpad: Interactive glyph-based visualization for real-time sports performance analysis. *In Computer Graphics Forum*, volume 31, Wiley Online Library, 2012, pp. 1255–1264.

[8] F. Buttussi and L. Chittaro, Mopet: A context-aware and user-adaptive wearable system for fitness training, *Artificial Intelligence in Medicine* **42**(2) (2008), 153–163.

[9] A. Holzinger, S. Dorner, M. Födinger, A.C. Valdez and M. Ziefle, Chances of increasing youth health awareness through mobile wellness applications. *In HCI in Work and Learning, Life and Leisure*, Springer, 2010, pp. 71–81.

[10] R. Agrawal and R. Srikant, Mining sequential patterns. *In Data Engineering, 1995 Proceedings of the Eleventh International Conference on*, IEEE, 1995, pp. 3–14.

[11] R. Agrawal, T. Imieliński and A. Swami, Mining association rules between sets of items in large databases, *In ACM SIGMOD Record* **22** (1993), 207–216. ACM.

[12] R. Agrawal, R. Srikant, et al., Fast algorithms for mining association rules, *In Proc 20th Int Conf Very Large Data Bases, VLDB*, volume 1215, 1994, pp. 487–499.

[13] R. Srikant and R. Agrawal, Mining sequential patterns: Generalizations and performance improvements, Springer, 1996.

[14] F. Masseglia, F. Cathala and P. Poncelet, The psp approach for mining sequential patterns. *In Principles of Data Mining and Knowledge Discovery*, Springer, 1998, pp. 176–184.

[15] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal and M.-C. Hsu, Freespan: frequent pattern-projected sequential pattern mining. *In Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2000, pp. 355–359. ACM.

[16] J. Han, J. Pei, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal and M.C. Hsu, Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. *In Proceedings of the 17th International Conference on Data Engineering*, 2001, pp. 215–224.
[17] M.J. Zaki, Spade: An efficient algorithm for mining frequent sequences, *Machine Learning* **42**(1-2) (2001), 31–60.
[18] J. Ayres, J. Flannick, J. Gehrke and T. Yiu, Sequential pattern mining using a bitmap representation. *In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2002, pp. 429–435. ACM.
[19] Z. Yang and M. Kitsuregawa, Lapin-spam: An improved algorithm for mining sequential pattern. *In Data Engineering Workshops, 2005 21st International Conference on*, 2005, pp. 1222–1222. IEEE.
[20] X. Yan, J. Han and R. Afshar, Clospan: Mining closed sequential patterns in large datasets. *In Proc 2003 SIAM IntâĂŹl Conf Data Mining (SDMâĂŹ03)*, 2003, pp. 166–177.
[21] J. Wang and J. Han, Bide: Efficient mining of frequent closed sequences. *In Data Engineering, 2004 Proceedings 20th International Conference on*, 2004, pp. 79–90. IEEE.
[22] A. Gomariz, M. Campos, R. Marin and B. Goethals, Clasp: An efficient algorithm for mining frequent closed sequences, *In Advances in Knowledge Discovery and Data Mining*, Springer, 2013, pp. 50–61.
[23] A. Pitman and M. Zanker, Insights from applying sequential pattern mining to e-commerce click stream data. *In Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, 2010, pp. 967–975. IEEE.
[24] E. Baralis, G. Bruno, S. Chiusano, V.C. Domenici, N.A. Mahoto and C. Petrigni, Analysis of medical pathways by means of frequent closed sequences. *In Knowledge-Based and Intelligent Information and Engineering Systems*, 2010, pages 418–425. Springer.
[25] Y.-P. Wu, K.-P. Wu and H.-M. Lee, Stock trend prediction by sequential chart pattern via k-means and aprioriall algorithm. *In Technologies and Applications of Artificial Intelligence (TAAI), 2012 Conference on*, 2012, pp. 176–181. IEEE.
[26] J. Han, M. Kamber and P. Jian, Data mining: concepts and techniques, Morgan kaufmann, 2006.
[27] G.-E. Yap, X.-L. Li and S.Y. Philip, Effective next-items recommendation via personalized sequential pattern mining. *In Database Systems for Advanced Applications*, 2012, pp. 48–64. Springer.
[28] S. Concaro, L. Sacchi and R. Bellazzi, Temporal data mining methods for the analysis of the ahrq archives, *Proc Am Med Inform Assoc*, (2007).
[29] HCUP Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality, Rockville, MD. www.hcup-us.ahrq.gov/nisoverview.jsp, 2005.
[30] Y.-H. Hu, Y.-L. Chen and K. Tang, Mining sequential patterns in b2b environment, *Journal of Information Science*, 2009.
[31] I.A.R. Li, A.K. Dey and J. Forlizzi, Monitoring and feedback to increase awareness of exercise activities. *In Ubicomp 2005 Workshop on Monitoring, Measuring and Motivating Exercise: Ubiquitous Computing to Support Physical Fitness*. Citeseer, 2005.
[32] E.M. Tapia, S.S. Intille, W. Haskell, K. Larson, J. Wright, A. King and R. Friedman, Real-time recognition of physical activities and their intensities using wireless accelerometers and a heart rate monitor. *In Wearable Computers, 2007 11th IEEE International Symposium on*, 2007, pp. 37–40. IEEE.
[33] C.G. Wylie and P. Coulton, Persuasive mobile health applications. *In Electronic Healthcare*, Springer, 2009, pp. 90–97.
[34] I. Pernek, G. Stiglic and P. Kokol, How hard am i training? using smart phones to estimate sport activity intensity. *In Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on*, IEEE, 2012, pp. 65–68.
[35] S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin and M. Srivastava, Biketastic: Sensing and mapping for better biking. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 2010, pp. 1817–1820.
[36] S. Kotsiantis and D. Kanellopoulos, Discretization techniques: A recent survey, *GESTS International Transactions on Computer Science and Engineering* **32**(1) (2006), 47–58.
[37] M.O. Ward, A taxonomy of glyph placement strategies for multidimensional data visualization, *Information Visualization* **1**(3-4) (2002), 194–210.
[38] H. Müller, R. Reihs, K. Zatloukal and A. Holzinger, Analysis of biomedical data with multilevel glyphs, *BMC Bioinformatics* **15**(Suppl 6) (2014), S5.
[39] C. Turkay, F. Jeanquartier, A. Holzinger and H. Hauser, On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics. *In Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, 2014, pp. 117–140. Springer.
[40] R. Developement Core Team et al., A language and environment for statistical computing, R, 2005.
[41] C. Buchta, M. Hahsler, and with contributions from Daniel Diaz. arulesSequences: Mining frequent sequences, R package version 0.2-4, 2013.
[42] M. Hahsler, B. Grün and K. Hornik, A computational environment for mining association rules and frequent item sets, 2005.
[43] A. Zeileis and G. Grothendieck, zoo: S3 infrastructure for regular and irregular time series, *Journal of Statistical Software* **14**(6) (2005), 1–27.
[44] G. Grothendieck, sqldf: Perform SQL Selects on R Data Frames, R package version 0.4-6.4, 2012.
[45] G. Hrovat, G. Stiglic, P. Kokol and M. Ojsteršek, Contrasting temporal trend discovery for large healthcare databases, *Computer Methods and Programs in Biomedicine* **113**(1) (2014), 251–257.
[46] J.H. Wilmore, D.L. Costill, et al., Physiology of sport and exercise, *Human Kinetics Publishers* (1994).
[47] J.B. Carter, E.W. Banister and A.P. Blaber, Effect of endurance exercise on autonomic control of heart rate, *Sports medicine* **33**(1) (2003), 33–46.
[48] A.C. Guyton and J.E. Hall, *Medical physiology*, Saunders, 1961.
[49] F. Dadashi, G.P. Millet and K. Aminian, A bayesian approach for pervasive estimation of breaststroke velocity using a wearable imu, *Pervasive and Mobile Computing*, 2014.