Discovering Temporal Knowledge in Multivariate Time Series

Fabian Mörchen and Alfred Ultsch

Data Bionics Research Group
University of Marburg, D-35032 Marburg, Germany

Abstract. An overview of the *Time Series Knowledge Mining* framework to discover knowledge in multivariate time series is given. A hierarchy of temporal patterns, which are not a priori given, is discovered. The patterns are based on the rule language *Unification-based Temporal Grammar*. A semiotic hierarchy of temporal concepts is build in a bottom up manner from multivariate time instants. We describe the mining problem for each rule discovery step. Several of the steps can be performed with well known data mining algorithms. We present novel algorithms that perform two steps not covered by existing methods. First results on a dataset describing muscle activity during sports are presented.

1 Introduction

Many approaches in time series data mining concentrate on the compression of univariate time series (patterns) down to a few temporal features. The aim is often to speed up the search for *known* patterns in a time series database (see Hetland (2004) for an overview). The introduced techniques for time series abstraction and the accompanying similarity measures can often be used in other contexts of data mining and knowledge discovery, e.g. for searching *unknown* patterns or rules. Most rule generation approaches search for rules with a known consequent, that is some unknown pattern predicting a predefined event (Povinelli (2000)). In addition, the form of the possible patterns is often restricted by rule language syntax (see Hetland and Saetrom (2002) for a discussion). Very few approaches search for rules with an unknown antecedent part *and* an unknown consequent part (Saetrom and Hetland (2003), Höppner (2001)). Finally, few publications explicitly consider multivariate time series (Höppner (2002)).

Knowledge Discovery is the mining of previously unknown rules that are useful, understandable, interpretable, and can be validated and automatically evaluated (Ultsch (1999)). It is unlikely that one method will maintain good results on all problem domains. Rather, many data mining techniques need to be combined for this difficult process. In Guimaraes and Ultsch (1999) some early results of understandable patterns extracted from multivariate times series were presented. Here, we want to describe our new hierarchical time series rule mining framework Time Series Knowledge Mining (TSKM).

The rest of this paper is structured as follows. The data from sports medicine is described in Section 2. The temporal concepts expressible by the rule language are explained in Section 3 using examples from the application. Section 4 defines the steps of the framework and gives details on two novel algorithms. The merits of the application, possible extensions of our work, and related methods are discussed in Section 5. Section 6 summarizes the paper.

2 Data

The TSKM method is currently applied to a multivariate time series from sports medicine. Three time series describe the activity of the leg muscles during In-Line Speed Skating measured with surface EMG (Electromyography) sensors. The current leg position is described by three angle sensors (Electrogoniometer), attached at the ankle, the knee, and the hip. Finally, there is a time series produced by an inertia switch, indicating the first ground contact.

3 Unification-based Temporal Grammar

The Unification-based Temporal Grammar (UTG) is a rule language developed especially for the description of patterns in multivariate time series (Ultsch (1996)). Unification-based Grammars are an extension of context free grammars with side conditions. They are formulated with first order logic and use unification. The UTG offers a hierarchical description of temporal concepts. This opens up unique possibilities in relevance feedback during the knowledge discovery process and in the interpretation of the results. An expert can focus on particularly interesting rules and discard valid but known rules before the next level constructs are searched. After obtaining the final results, an expert can zoom into each rule to learn about how it is composed and what it's meaning and consequences might be.

At each hierarchical level the grammar consists of semiotic triples: a unique symbol (syntax), a grammatical rule (semantic), and a user defined label (pragmatic). The unique symbols can be generated automatically during the mining process. The labels should be given by a domain expert for better interpretation. Due to lack of space we will only briefly describe the conceptual levels of the hierarchy (see also Figure 1) along with an example from the application. The basic ideas of the UTG were developed in Ultsch (1996) and applied in Guimaraes and Ultsch (1999). For a detailed description see Ultsch (2004).

A *Primitive Pattern* is a temporal atom with unit duration. It describes a state of the time series at the smallest time scale. For the muscle activity we found 3 to 5 states corresponding to subsets of *very low*, *low*, *medium*, *high*, and *very high*. For the leg position six typical sport movement phases, namely *stabilization*, *forward gliding*, *pre-acceleration* (of center of gravity), *preparation* (of foot contact), *foot placement*, *push-off*, and *leg swing* were

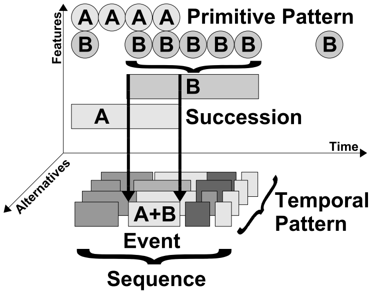

Fig. 1. UTG concepts

```
An Event is a 'Weight Transfer' if
  'Foot contact' is 'on'
coincides with
  'Movement' is 'glide'
coincides with
  'Medial Gastrocnemius' is 'high'
coincides with
  'Vastus Medialis' is 'high'
coincides with
  'Gluteus Maximus' is 'high'
.
```

Fig. 2. UTG Event rule

identified. The labeling (pragmatic) needs to be done in close cooperation with an expert to ensure meaningful results.

A *Succession* introduces the temporal concept of duration. It represents a time interval where nearly all time points have the same Primitive Pattern label. Short interruptions (*Transients*) of an otherwise persisting state should be removed. For the muscle activity interruptions shorter than 50ms were discarded because a change of the state at this time scale is physiologically not plausible. The movement phases are much longer in general, interruptions up to 100ms were removed.

An *Event* represents the temporal concept of coincidence. It represents a time interval where several Successions overlap. If the start points of all overlapping Succession are approximately equal and the same is true for the end points, the Event is called *synchronous.* The Events present in the skating data relate the current movement phase and the position of the foot to the activation of the muscles *at the same time.* One Event in the skating data corresponded to all three muscles being highly active during the forward gliding phase with the foot on the ground. This Event was labeled by the expert as the weight transfer from one leg to the other (see Figure 2). The five most frequent Events were labeled by the expert as follows: *active gliding* (G), *relaxation* (R), *anticipation* (A), *weight transfer* (W), *initial gliding* (I).

A *Sequence* introduces the temporal concept of order. A Sequence is composed of several Events occurring sequentially, but not necessarily with meeting end and start points. The three most frequent Sequences were (G,R,A), (G,R,A,W), and (G,R,A,W,I). They all have the same prefix (G,R,A) corresponding to the *contraction & relaxation* phase. The Events W and I complete the typical skating motion cycle, but are not always recognized due to measurement errors in the foot contact sensor.

A *Temporal Pattern* is the summary of several Sequences by allowing a set of Events at some positions of the pattern. Temporal Patterns represent the non-temporal concept of alternative. Since all Sequences were quite similar in this application, they were merged into a single Temporal Pattern describing the *typical motion cycle* during Inline-Speed Skating.

4 Time Series Knowledge Mining

The temporal knowledge discovery framework Time Series Knowledge Mining (TSKM) aims at finding interpretable symbolic rules describing interesting patterns in multivariate time series. We define the data models, mining task, and algorithms for each level of the framework. The levels correspond to the temporal concepts of the UTG and include some additional steps (see Figure 5. Some tasks can be solved with well known data mining algorithms, while other require new algorithms.

Aspects: The starting point of the TSKM is a multivariate time series, usually but not necessarily uniformly sampled. An expert should divide the features of the time series into possibly overlapping groups that are related w.r.t. the investigated problem domain. Each feature subset is called an Aspect and should be given a meaningful name. In the absence of such prior knowledge, one Aspect per time series can be used. Each Aspect is treated individually for the first steps of the framework.

Preprocessing and feature extraction techniques are applied to each Aspect or even each time series individually. This is a highly application dependent step.

Finding Primitive Patterns: The task of finding Primitive Patterns is the reduction of the time series to a series of states. The input data is a real or vector valued time series, the output is a time series of symbols for each atomic time interval. It is important, that each symbol is accompanied by a rule and a linguistic description to complete the semiotic triple.

Many discretization techniques can be used to find Primitive Patterns for univariate Aspects. Simple methods aggregate the values using histograms. Additionally down-sampling can be performed by aggregation over a time window, e.g. Lin et al. (2003). The symbols for the bins can easily be mapped to linguistic descriptions like *high* or *low*. A first order description method describes the current trend of a time series, e.g. Kadous (1999). Second order descriptions additionally incorporate the second derivative of the signal to distinguish convex from concave trends, e.g Höppner (2001).

For Aspects spanning several time series we propose to use clustering and rule generation on the spatial attributes. If the process alternates between several regimes or states, these regions should form clusters in the high dimensional space obtained disregarding the time attribute. In Guimaraes and Ultsch (1999) and for the identification of the skating movement phases Emergent Self-Organizing Maps (ESOM)(Ultsch (1999)) have been used to identify clusters. The rules for each cluster were generated using the Sig* Algorithm (Ultsch (1991)). The ESOM enables visual detection of outliers and arbitrarily shaped clusters and Sig* aims at understandable descriptions of the Primitive Patterns.

Finding Successions: The input data for finding Successions is a univariate symbolic time series of Primitive Patterns, the output consist of a univariate series of labeled intervals. The merging of consecutive Primitive Pat-

```
$i := 2$
while $i < n$
    // check symbols and duration
    if $(s_{i-1} = s_{i+1})$ and $(d_i \leq d_{max})$
    and $(d_i \leq r_{max} * (d_{i-1} + d_{i+1}))$
        // merge 3 intervals
        $d_i - 1 := \sum_{j=i-1}^{i+1} d_j$
        $\forall k \in \{i, i+1\}\ d_k := 0$
        $i := i + 2$
    else
        $i := i + 1$
    end if
end while
// remove zero durations
$S := S \setminus \{(t_i, d_i, s_i) \in S | d_i = 0\}$
```

Fig. 3. SequentialTransientFilter

```
$i := 2$
while $i < n$
    $s := i$
    // search end
    while S(i)=S(i-1)
        $i := i + 1$
    end while
    // check duration
    if $i - s \geq min_d$
        add Event on $[s, i - 1]$
    end if
    $i := i + 1$
end while
```

Fig. 4. FullEvents

terns into a Succession is straight forward. But with noisy data there are often interruptions of a state (*Transients*). Let a Succession interval be a triple of a start point t, a duration d, and a symbol s. Let the input Successions be $S = \{(t_i, d_i, s_i)\ i = 1..n\}$ with $t_i + d_i \leq t_{i+1}$ and $s_i \neq s_{i+1}$. Let d_{max} be the maximum absolute duration and r_{max} the maximum relative duration of a Transient. For the removal of Transients we propose the *SequentialTransientFilter* algorithm shown in Figure 3.

The time complexity of the algorithm is $O(n)$. A good choice for r_{max} is 0.5, i.e. the gap is allowed to be at most half as long as the surrounding segments together. The d_{max} parameter has to be chosen w.r.t to the application. Often, some knowledge on the minimum duration of a phenomena to be considered interesting is available.

Finding Events: Events represent the concept of coincidence, thus in this step all Aspects are considered simultaneously. The input data is a multivariate series of labeled intervals (Successions) and the output is a univariate series of labeled intervals (Events). Let S be a $k \times n$ matrix containing the symbols of the Successions from k Aspects at n time points. We use $S(i)$ for the i-th column of S and $S(i) = S(j)$ for element-wise equality. Let min_d be the minimum duration of an Event. The algorithm *FullEvents* shown in Figure 4 discovers all Events where Successions from all Aspects coincide.

The time complexity of the algorithm is $O(n)$. The d_{max} parameter can be chosen similar the maximum duration of Transients when finding Successions. The post-processing to identify synchronous Events is rather straight forward. For each Event the maximum difference between all start points of the participating Successions are checked against a threshold and the same is done for the end points. Additionally, the SequentialTransientFilter algorithm can be applied to the resulting Event series.

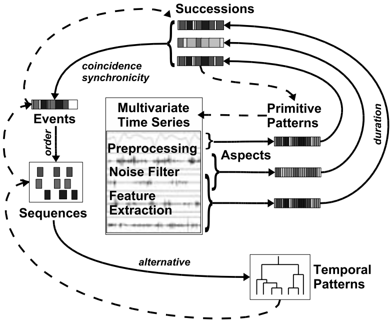

Fig. 5. The TSKM process

Finding Sequences: For the step of finding Sequences there is a large number of algorithms that could be utilized. The input data is a univariate series of labeled intervals and the output is a set of subsequences thereof. For moderately sized dataset we use a suffix trie (e.g. Vilo (1998)). Compared to a suffix tree, all egde labels in a trie have length one. Tries are larger, but can easier be queried for patterns with wild-cards. The trie stores all subsequences up to a maximum length and can be queried with frequency and length thresholds to find the most interesting patterns. For larger datasets more scalable and robust techniques from sequential pattern mining, e.g. Yang et al. (2002), can be used.

Finding Temporal Patterns: The Sequences often overlap. The last step of the framework is finding generalized Sequences, called Temporal Patterns. We propose to use clustering based on a string metric to find groups of similar Sequences. The Temporal Pattern can be generated by merging the patterns in a cluster using groups of symbols at positions where the patterns do not agree. We have successfully used hierarchical clustering based on the string edit distance with a dendrogram visualization.

5 Discussion

The Temporal Pattern found in the skating data provided new insights for the expert. The symbolic representation offers better interpretation capabilities on the interactions of different skeletal muscles than the raw EMG data. We identified the most important cyclical motion phases. The rule describing this phase can be expanded to provide more details. At the level of Temporal Patterns there is a Sequence of Events allowing some variations. Each Event is associated with a rule listing the coinciding muscle and movement states in form of the underlying Successions. Each movement Succession is linked to a Primitive Pattern with a rule describing the range of hip, knee, and ankle angles observed during this state. We plan to compare the patterns between several skaters and running speeds to investigate possible differences. Based on background knowledge about the performance of the individual skaters this can lead to strategies for individualized training optimization.

One could criticize the manual interaction needed at some levels of our mining process, but we feel that a fully automated knowledge discovery is not desirable. We see the hierarchical decomposition into single temporal concepts as a great advantage of the TSKM. The separate stages offer unique possibilities for the expert to interpret, investigate and validate the discovered rules at different abstraction levels. The search space for the algorithms is smaller than when mining several concepts at the same time. Also, a large set of different algorithms can be plugged in the framework, e.g. segmentation to discover Successions or Hidden Markov Models to obtain Primitive Pattern to name just a few.

Usually only frequent Events and Sequences are kept while rare occurring patterns are discarded from further processing. Depending on the application, rare pattern might be important, however, and ranking should be done by a different interestingness measure.

While the data model for Events is currently univariate, we are experimenting with algorithms allowing overlapping Events involving less Aspects. However, this increases the number of Events found and makes mining Sequences more problematic.

There are only very few methods for rule discovery in multivariate time series. Last et al. (2001) use segmentation and feature extraction per segment. Association rules on adjacent intervals are mined using the Info-Fuzzy Network (IFN). The rule set is reduced using fuzzy theory. Höppner mines temporal rules in sequences of labeled intervals (Höppner (2001), Höppner (2002)), also obtained by segmentation and feature extraction. Patterns are expressed with Allen's interval logic (Allen (1983)) and mined with an Apriori algorithm. A comparison to the TSKM method on a conceptual and experimental level is planned.

6 Summary

We have presented our time series knowledge extraction framework TSKM. The hierarchal levels of the underlying rule language UTG cover the temporal concepts duration, coincidence, synchronicity and order at successive levels. Rules from each level are accompanied by linguistic descriptions, thus partial results can be interpreted and filtered by experts. We proposed algorithms for the mining stages including two new algorithms for mining duration and coincidence. First results of an application in sports medicine were mentioned.

Acknowledgements

We thank Dr. Olaf Hoos, Department of Sports Medicine, Philipps-University Marburg, for the data and interpretation of the mining results.

References

ALLEN, J. F. (1983): Maintaining knowledge about temporal intervals. *Comm. ACM, 26(11), 832–843.*

GUIMARAES, G. and ULTSCH, A. (1999): A method for temporal knowledge conversion. In: D. J. Hand, J. N. Kok and M. R. Berthold (Eds.): *Proc. of the 3rd Int. Symp. on Advances in Intelligent Data Analysis.* Amsterdam, Springer, 369–380.

HETLAND, M. L. (2004): A survey of recent methods for efficient retrieval of similar time sequences. In: M. Last, A. Kandel and H. Bunke (Eds.): *Data Mining in Time Series Databases.* World Scientific, 23–42.

HETLAND, M. L. and SAETROM, P. (2002): Temporal rule discovery using genetic programming and specialized hardware. In: A. Lotfi, J. Garibaldi, and R. John (Eds.): *Proc. of the 4th Int. Conf. on Recent Advances in Soft Computing (RASC)*, 182–188.

HÖPPNER, F. (2001): Discovery of temporal patterns – learning rules about the qualitative behaviour of time series. In: *Proc. of the 5th European PKDD.* Springer, Berlin, 192–203.

HÖPPNER, F. (2002): Learning dependencies in multivariate time series. *Proc. of the ECAI'02 Workshop on Knowledge Discovery in (Spatio-) Temporal Data, Lyon, France*, 25–31.

KADOUS, M. W. (1999): Learning comprehensible descriptions of multivariate time series. In: *Proc. 16th International Conf. on Machine Learning*, 454–463.

LAST, M., KLEIN, Y. and KANDEL, A. (2001): Knowledge discovery in time series databases. *IEEE Transactions on Systems, Man, and Cybernetics (31B).*

LIN, J., KEOGH, E., LONARDI, S. and CHIU, B. (2003): A symbolic representation of time series, with implications for streaming algorithms. In: *Proc. 8th ACM SIGMOD workshop DMKD 2003*, 2–11.

POVINELLI, R. J. (2000): Identifying temporal patterns for characterization and prediction of financial time series events. In: *Proc. International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining: TSDM 2000, Lyon, France*, 46–61.

SAETROM, P. and HETLAND, M. L. (2003): Unsupervised temporal rule mining with genetic programming and specialized hardware. In: *Proc. Int. Conf. on Machine Learning and Applications, ICMLA*, 145–151.

ULTSCH, A. (1991): Konnektionistische Modelle und ihre Integration mit wissensbasierten Systemen. *Forschungsbericht Informatik Nr. 396, Universität Dortmund, Habilitationsschrift.*

ULTSCH, A. (1996): Eine unifikationsbasierte Grammatik zur Beschreibung von komplexen Mustern in multivariaten Zeitreihen. *personal communication*

ULTSCH, A. (1999): Data Mining and Knowledge Discovery with Emergent Self-Organizing Feature Maps for Multivariate Time Series. In: Oja, E. and Kaski, S. (Eds.): *Kohonen Maps*, Elsevier, New York, 33–46.

ULTSCH, A. (2004): Unification-based temporal grammar. *Technical Report No. 37, Philipps-University Marburg, Germany.*

VILO, J. (1998): Discovering frequent patterns from strings. *Technical Report C-1998-9, Department of Computer Science, University of Helsinki.*

YANG, J., WANG, W., YU, P.S. and HAN, J. (2002): Mining long sequential patterns in a noisy environment. In: *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, 406–417.