

ASSOCIATION RULE MINING IN MULTIPLE, MULTIDIMENSIONAL TIME SERIES MEDICAL DATA

Gaurav N. Pradhan

Department of Computer Science
Arizona State University
Tempe, AZ 85287.

B. Prabhakaran

Department of Computer Science
University of Texas at Dallas
Richardson, TX 75080.

ABSTRACT

Time series pattern mining (TSPM) finds correlations or dependencies in same series or in multiple time series. When the numerous instances of multiple time series data are associated with different quantitative attributes, they form a multiple multi-dimensional framework. In this paper, we consider real-life time series data of muscular activities of human participants obtained from multiple Electromyogram (EMG) sensors and discover patterns in these EMG data streams.

Each EMG data stream is associated with quantitative attributes such as energy of the signal and onset time which are required to be mined along with EMG time series patterns. We propose a two-stage approach for this purpose: in the first stage, our emphasis is on discovering frequent patterns in multiple time series by doing sequential mining across time slices. And in the next stage, we focus on the quantitative attributes of only those time series that are present in the patterns discovered in the first stage.

Our evaluation with large sets of time series data from multiple EMG sensors demonstrate that our two-stage approach speeds up the process of finding association rules in such multidimensional environment as compared to other methods and scales up linearly in terms of number of time series involved. Our approach is generic and applicable to any multiple time series dataset format.

Index Terms— Association rules, multidimensional data, electromyogram, prosthetics.

1. INTRODUCTION

Analyzing multiple time series and multidimensional databases has several real-world applications such as medical, finance, and engineering. The main interest lies in discovering the structural and temporal relationships or the *dependencies* hidden inside these data streams. Dependencies give the information on the frequent co-occurrences of events in multiple streams of time series data. These dependencies are expressed in the form of *association rules*. In this paper,

we focus on mining association rules between muscular activities recorded by the surface electromyographic (EMG) sensors placed on different parts of human body. These types of association rules related to muscular patterns from different parts of body can be effectively used in designing/training prosthetic devices, and monitoring the longitudinal improvement for the patients in rehabilitative environment by giving real-time bio-feedback information.

Rule: $\langle S_F, onset = 13 \rangle \wedge \langle S_E, onset = 10 \rangle \rightarrow \langle S_G, onset = -9 \rangle$ (support=3.22%, conf.=100%)

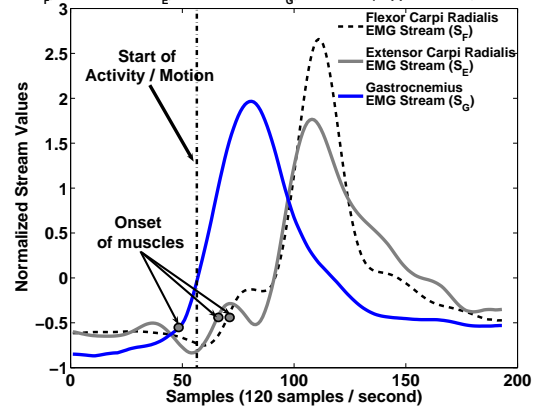


Fig. 1. An example of pattern discovery within multiple time series for “Raise-Arm” activity.

As EMG is a non-stationary signal, we represent it as a linear envelope [1], which is computed by passing low-pass filter through a full-wave rectification of the original EMG signal. Further, each EMG time series has important associative quantitative features such as onset timings and energies that are useful to analyze the behavior of muscles. A real-example of a rule from EMG database is shown in Figure 1 and is represented as follows:

$$\langle Pattern(S_{Flexor}), Onset = 13ms \rangle \wedge \langle Pattern(S_{Extensor}), Onset = 10ms \rangle \Rightarrow \langle Pattern(S_{Gastrocnemius}), Onset = -9ms \rangle \quad (1)$$

We can read this rule as follows: *While raising arms, if the flexor carpi radialis muscle with pattern S_{Flexor} has an onset of approximately 13ms and extensor carpi radialis muscle*

with pattern $S_{Extensor}$ has an onset of approximately 10ms after start of activity, then onset of gastrocnemius leg muscle with pattern $S_{Gastrocnemius}$ had been activated 9ms before the start of activity. This rule reveals the fact that, when person gets ready to raise his/her arms, just few milliseconds before, his/her legs are prepared with muscular contraction to maintain balance while raising arms. To find such kind of hybrid rules, previous approaches [2, 3] that computed the data-cubes by aggregating each and every combination of the multiple dimensions may be too time consuming and inefficient in terms of storage.

Approach : We propose a two-stage approach to mine high confidence patterns/rules in multiple, multi-attribute time series data. In the first stage, our emphasis is on discovering frequent patterns in multiple time series by doing sequential mining across time slices using an apriori technique. And in the next stage, we perform multi-dimensional pattern mining on the attributes of only those time series that are involved in the discovered, frequent time series patterns.

2. RELATED WORK

The traditional association rule mining algorithms to recognize frequent events in form of item-sets were built on quantitative databases such as market basket. [4] were among the first to address the problem of sequential pattern mining across similar data from transaction database, based on the Apriori property [5, 6].

The work on association rules was extended from sequential patterns to time series in [7], where authors proposed a rule discovery that finds the temporal relationships from time series using subsequence clustering. In [2, 3], authors explored data cube-based rule mining algorithms on sequential multidimensional databases, where each tuple/transaction consisted of one sequence with multi-dimensional features. In the area of multi-dimensional data sets, in [8], authors discussed a multidimensional data model, in which the multidimensional data was viewed as a value in the multidimensional space. Based on this model, efficient data mining have been performed using data cubes based on aggregates of dimensions were computed in [9, 10]. Decision tree mining is another well studied data mining problem and over the years many techniques have been designed to construct decision trees for mining the patterns in the streaming data [11, 12, 13].

3. MULTISTREAM PATTERN MINING FRAMEWORK

From the mining perspective, for each motion, we have *multiple time series information* and *multi-attribute information* that need to be processed in order to discover important relationships between acting muscles while doing movements or exercises.

3.1. Multiple time series information

Mining multiple time series pattern is difficult, but if they are modulated into strings of representative symbols using time slices, interesting patterns can be discovered and moreover, mining will be easier. In multi-block of *multiple time series information* of Figure 2, we have shown a string of symbols for each time series of EMG sensors for 5 motions. The original, equi-length time series data of 4 EMG sensors [biceps brachii (biceps), triceps brachii (triceps), flexor carpi radialis (flexor), and extensor carpi radialis (extensor)] are spliced into 3 time slices and hence each time series is represented in the form of 3 symbols.

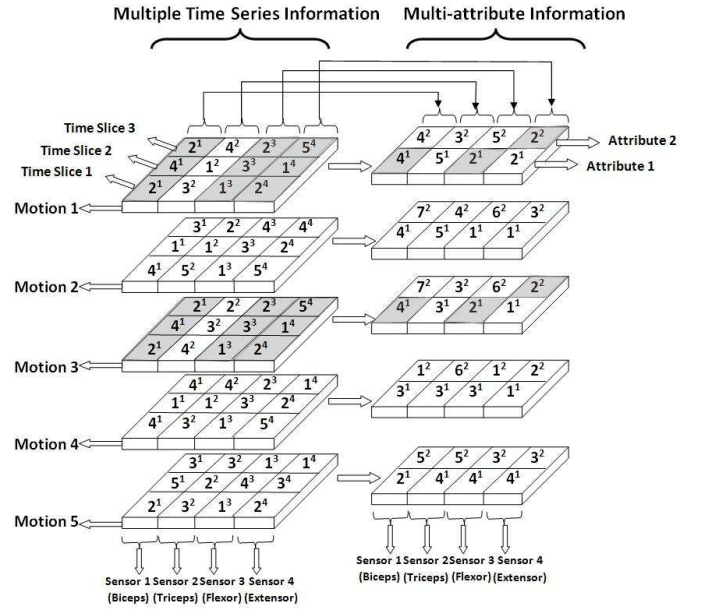


Fig. 2. A discretized version of multiple time series information with its associated multi-attribute information for the database of EMG sensors. (Note: A frequent pattern is shown in dark faces)

3.2. Multi-attribute Information

As discussed earlier, each EMG stream is associated with the quantitative attributes that are distributed in the continuous range of values. For fast and efficient association rule mining, it is necessary to transform these continuous distribution of attributes A_1, \dots, A_Q into discrete representation such as *symbols* with equiprobability. The discretized version of the multi-attribute information for each sensor in each motion is shown in Figure 2 (for brevity, only two attributes are shown). For discretized multi-attribute symbols, super-script j indicates the j^{th} attribute.

3.3. Discovering Association Rules

To discover the structural patterns in multiple time series data along with their attributes, we divide the association rule min-

ing process into two stages. First, we mine frequent time series patterns in multiple time series information using sequential *apriori* algorithm [6]. This sequential approach helps in pruning the patterns that are non-frequent in preceding time-slice, as they will always be non-frequent for the next iterations corresponding to subsequent time-slices. Next, we mine frequent attributes from the *projected multi-attribute information* corresponding to the frequent time series patterns. The *projected multi-attribute information* contains only the attributes that are associated with the time series present in the discovered frequent patterns. The association rules for the given *support threshold* can be derived by combining frequent time series patterns and corresponding frequent attribute patterns.

4. PERFORMANCE EVALUATION

4.1. Test environment and datasets

All our experiments were performed and tested on a PC with Pentium IV 2.6GHz processor with 1.5GB main memory under the Windows XP operating system. We collected EMG time series data from 20 different participants that were uniformly distributed across ages 20-80, with initial consent. As the behavior (movement dynamics) and goal-oriented purpose of each experiment/motion is different, we needed to discover different sets of association rules corresponding to each experiment. In following Section 4.2, we show the performance of our technique on an experiment where person raises his both arms on reacting to a visual cue. With 120 samples per sec. and nearly 2 seconds for each trial of *raise* motion, we have approximately 50000 samples each from different EMG sensors (total of 600,000 samples from 20 participants). Each motion is represented by maximum of 12 synchronous time series data, which form the *multiple time series information*. To form *multi-attribute information*, we extracted quantitative attributes like *onset* and *energy* from the each time series of EMG sensor as discussed in Section 1.

4.2. Experimental results

We evaluated the performance of our approach by varying support thresholds for mining association rules in EMG sensor database that consisted of multiple time series and multi-attribute information. For comparison purpose, we tested our data set with bottom-up cube computation (BUC) [14] for finding association rules. Since BUC worked on multi-dimensional data and also took advantage of minimum support pruning property, it was interesting to see the comparative results with our proposed approach. As seen in Figure 3, our two-stage proposed approach always runs faster than BUC by more than an order for all tested support thresholds.

Figure 4 shows the comparison between quality of the mined associative rules by our approach with BUC, based on the 100% confidence rules among the total output rules for the tested support thresholds. Based on our observation in the

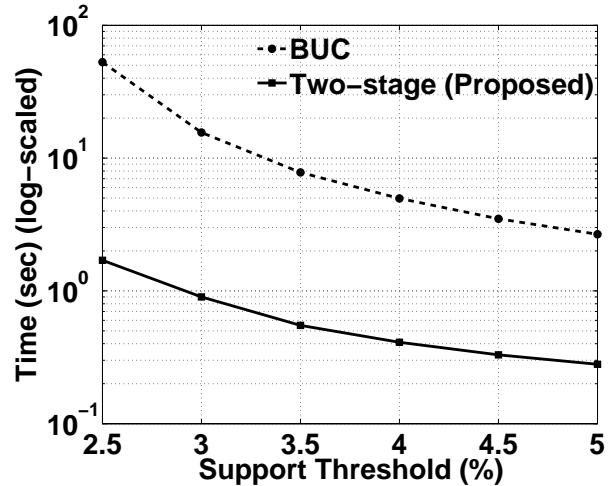


Fig. 3. Comparison of running time with BUC across support thresholds.

graph shown in Figure 4, we have high confidence rules with low support (2.5-3.5%), which are important to discover rare but specific muscular activities/conditions in participants. On other hand, we have very general rules with high support (4-5%) but average confidence which are necessary to monitor the generic behavior of the muscular activities among participants. Thus, the rules that are interesting to the user depends on his/her requirements and our approach efficiently provides the rules depending on his/her needs.

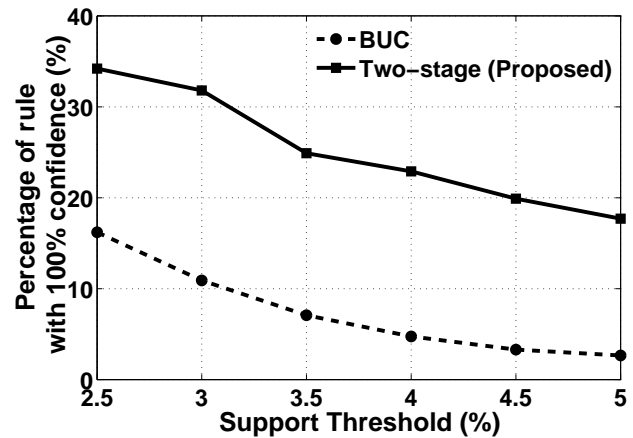


Fig. 4. Comparison of percentage of rules with 100% confidence

Further, as we get lots of rules for different sets of experiments with varying confidences and support, we used the metric called *J-measure* proposed in [15] that gave us the quantitative measure of information content present in rules, using the ideas of information theory. *J-measure* balances confidence and support, and moreover, is also the simplified measurement as it is dependent on the frequencies of the item-sets present in the corresponding rule. For a mined rule $Y \Rightarrow X$,

corresponding J-measure is given as,

$$J(Y \Rightarrow X) = p(y) * \left[p(x|y) \cdot \log \left(\frac{p(x|y)}{p(x)} \right) + (1 - p(x|y)) \cdot \log \left(\frac{(1 - p(x|y))}{(1 - p(x))} \right) \right] \quad (2)$$

J-measure is the product of two terms: $p(y)$, which is probability of the antecedent of the rule (a measure of *hypothesis simplicity*) and term in square brackets that gives the cross entropy (a measure of the *goodness of fit* between rule and data). Further information on J-measure is given in [15]. High J-measure indicates an important rule, but a rule with high confidence will not have a high J-measure if the corresponding support is very low.

Figure 5 shows the comparison of the corresponding J-measure for rules with 100% confidence. We achieved higher rate of information content for all tested support thresholds, which suggests the better informative rules as compared to BUC approach. Also, from Figure 5, lower support thresholds gives high amount of redundant rules, which reduces the average information content.

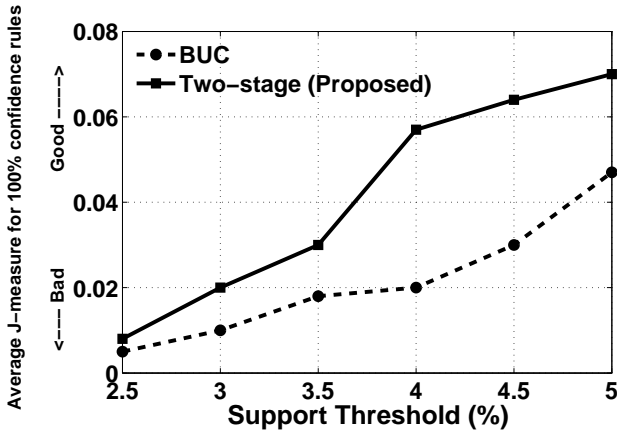


Fig. 5. Average J-measure for rules with 100% confidence.

5. CONCLUSIONS

In this paper, we introduced an efficient technique to discover hybrid-multidimensional associative rules in synchronous, multiple time series database, where each time series is associated with quantitative features or attributes. By conducting an extensive set of experiments on real-life data set such as electromyogram we have shown the effectiveness of the algorithm design. Our approach runs over an order of magnitude faster and gives high-confident association rules than bottom-up computation technique for multi-dimensional pattern mining. It also has linear scalability in terms of the number of time series present in database.

Though we have tested our approach using EMG database, the techniques should be applicable to any real data sets involving multiple sequence or time series patterns that are associated with other discrete attributes. Such circumstances

form multi-attribute information for each sequence or series. The direct advantage of the proposed method for mining on EMG streams is in the bio-medical fields such as prosthetic designs, physical medicines, and rehabilitations.

6. REFERENCES

- [1] D. Gordon E. Robertson, Graham E. Caldwell, and Joseph Hamill, *Research Methods in Biomechanics*, 2004.
- [2] Micheline Kamber, Jiawei Han, and Jenny Chiang, "Metarule-guided mining of multi-dimensional association rules using data cubes," in *Proc. of the KDD conf.*, 1997, pp. 207–210.
- [3] Helen Pinto, Jiawei Han, Jian Pei, Ke Wang, Qiming Chen, and Umeshwar Dayal, "Multi-dimensional sequential pattern mining," in *Proc. of CIKM*, 2001, pp. 81–88.
- [4] Rakesh Agrawal and Ramakrishnan Srikant, "Mining sequential patterns," in *Proc. of the 11th ICDE'*, Taipei, Taiwan, 1995, pp. 3–14.
- [5] Rakesh Agrawal, Tomasz Imielinski, and Arun N. Swami, "Mining association rules between sets of items in large databases," in *Proc. of the ACM SIGMOD*, Washington, D.C., 26–28 1993, pp. 207–216.
- [6] Rakesh Agrawal and Ramakrishnan Srikant, "Fast algorithms for mining association rules," in *Proc. of the 20th VLDB conf.* 12–15 1994, pp. 487–499, Morgan Kaufmann.
- [7] Gautam Das, King-Ip Lin, Heikki Mannila, Gopal Renganathan, and Padhraic Smyth, "Rule discovery from time series," in *Proc. of KDD*, 1998, pp. 16–22.
- [8] Surajit Chaudhuri and Umeshwar Dayal, "An overview of data warehousing and olap technology," *SIGMOD Rec.*, vol. 26, no. 1, pp. 65–74, 1997.
- [9] Venky Harinarayan, Anand Rajaraman, and Jeffrey D. Ullman, "Implementing data cubes efficiently," *SIGMOD Rec.*, vol. 25, no. 2, pp. 205–216, 1996.
- [10] Sameet Agarwal, Rakesh Agrawal, Prasad M. Deshpande, Ashish Gupta, Jeffrey F. Naughton, Raghu Ramakrishnan, and Sunita Sarawagi, "On the computation of multidimensional aggregates," in *Proc. of 22nd VLDB conf.*, 1996, pp. 506–521.
- [11] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proc. of KDD*, 2000, pp. 71–80.
- [12] Rouming Jin and Gagan Agrawal, "Efficient decision tree construction on streaming data," in *Proc. of the 9th ACM SIGKDD conf.*, New York, NY, USA, 2003, pp. 571–576.
- [13] Johannes Gehrke, Venkatesh Ganti, Raghu Ramakrishnan, and Wei-Yin Loh, "Boat optimistic decision tree construction," in *Proc. of ACM SIGMOD*, New York, NY, USA, 1999, pp. 169–180.
- [14] Kevin Beyer and Raghu Ramakrishnan, "Bottom-up computation of sparse and Iceberg CUBE," in *Proc. of ACM-SIGMOD'99*, Philadelphia, PA, June 1999, pp. 359–370.
- [15] P. Smyth and R. M. Goodman, "An information theoretic approach to rule induction from databases," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 4, pp. 301–316, 1992.