International Journal of Computer Science Engineering & Technology
IJCSET
Approved by
National Science Library (NSL)
National Institute of Science Communication and Information Resources (NISCAIR)
Council of Scientific and Industrial Research (CSIR)- New Delhi India .
ISSN :2455-9091

EVALUATION ON DIFFERENT SEQUENTIAL PATTERN ANALYSIS ALGORITHMS

[1] Priyadharshini. S.P, [2] Dr.M.Hemalatha,
[1] Ph.D Research Scholar, [2] Associate Professor
[1] Bharathiar University, [2] Sri Ramakrishna College of Arts & Science,
[1, 2] Coimbatore,India.

ABSTRACT- Sequential pattern analysis is use full for Developing novel approaches for sequential pattern analysis with applications in dynamic business environments, including operation and management tasks in healthcare industry as well as B2B (Business-to-Business) marketing. Then, the "skeleton" of the graph serves as a higher granularity on which hidden temporal patterns are more likely to be identified. In the meantime, the embedding topology of the graph allows us to translate the rich temporal content into a metric space. This opens up new possibilities to explore, quantify, and visualize sequential data. Our approach has shown to provide substantial improvements over the state-of-theart methods in challenging tasks of sequential pattern mining and sequence clustering. Evaluation on a Business-to-Business (B2B) marketing application demonstrates that our approach can effectively discover critical buying paths from noisy customer event data. This paper discuss about different types of Sequence Pattern Algorithms (GSP,SPAN, PrefixSpan, SPADE).

Keywords: [Sequence Pattern Mining, Clustering, Recovery, Utility.]

1. INTRODUCTION

A temporal skeletonization approach to proactively reduce the representation of sequences, so as to expose their hidden temporal structures. Our goal is to make temporal structures of the sequences more concise and clarified, and thus more prone to discovery. Our basic assumption is the existence of symbolic events that tend to aggregate temporally. Then, by identifying temporal clusters and mapping each symbol to the cluster it belongs to, we can reduce not only the cardinality of sequences but also their temporal variations. This allows us to find interesting hidden temporal structures which are otherwise obscured in the original representation. Exploring temporal clusters from a large number of sequences can be challenging. To achieve this, we have resorted to graph-based manifold learning. The basic idea is to summarize the temporal correlations in the data in an undirected graph. The "skeleton" of the graph (i.e., the temporal clusters) can then be extracted through the graph Laplacian, which serves as a higher granularity where hidden temporal patterns are more likely to be identified. A nice interpretation of such temporal grouping is that when individual symbols are replaced by their cluster labels, the averaged smoothness of all sequences is maximized. Intuitively, this can greatly improve the possibility of finding significant sequential patterns, as we shall observe empirically. In addition, the embedding topology of the graph allows us to translate the rich

temporal content of symbolic sequences into a metric space for further analysis and visualization. Compared with existing methods that attempt to reduce the cardinality via clustering, our approach does not require specific knowledge about the items. Instead, it caters directly to the temporal contents of given data sequences. To the best of our knowledge, using the temporal correlations to perform clustering and reduction of representation is a novel approach in sequential pattern mining. While these methods have been successfully applied in some application scenarios, there are some emerging issues to be addressed when we face the overwhelming scale and the heterogeneous nature of the sequential data. First, in some applications, it might be difficult to obtain the knowledge of symbols. For example, many sequential data simply use an arbitrary coding of events either for simplicity or security reasons. Second, there are circumstances where it is difficult to define distance among symbols, and therefore clustering becomes impractical. For example, it is unclear how to define the distance between actions customers have taken in their purchasing process. Finally, the biggest concern is that the grouping in these methods is performed irrespective of the temporal content. As a result, these methods may not be able to find statistically relevant temporal structures in sequential data. Therefore, there is a need to develop a new vision and strategy for sequential pattern mining.
Temporal skeletonization can be deemed as a transformation that maps the temporal structures of sequences into the topologies of a graph. Such a dual perspective provides not only more insights on pattern mining, but also brings powerful new tools for analysis and visualization. For example, many techniques in graph theories can be used to analyze symbolic sequences, which appear as random walks on the created graph. On the other hand, due to the explicit embedding, symbolic sequences are represented as numerical sequences or point clouds in the Euclidean space, for which visualization becomes much more convenient. Experimental results on real-world data have shown that the proposed approach can greatly alleviate the problem of curse of cardinality for the challenging tasks of sequential pattern mining and clustering. Also, we show that it is convenient to visualize sequential patterns in the Euclidean space by temporal skeletonization. In addition, the case study on a Business to-Business (B2B) marketing application demonstrates that our approach can effectively identify critical buying paths from noisy marketing data. The eigenvector of the graph Laplacian corresponding to the second smallest eigenvalue. In practice, one usually computes several (e.g., d) eigenvectors as columns in $y \in R|S|\times d$. The useful eigenvectors of the graph Laplacian not only provide a relaxed solution of temporal clusters, but also more interestingly, naturally connect to the manifold embedding of the graph. Note that the eigenvectors of the graph can be deemed a low-dimensional embedding, in which the proximity relation among objects preserves that in the original space. Since the similarity measurements in Wij of the graph reflect the temporal closeness of the events, the embedding eigenvectors of the graph will also inherit this configuration. Namely, if two symbols, e_i and e_j are temporally more related, their distance will also be small in the embedded space. In this way, our approach provides a direct platform for visualizing the temporal structures of sequential data.

2. LITERATURE SURVEY

Yan et al. Yan and Chen developed a mobile App recommender system, named Appjoy, which is based on user's App usage records to build a preference matrix instead of using explicit user ratings. Also, to solve the sparsity problem of App usage records, Shi et al. Shi and Ali studied several recommendation models and proposed a content based collaborative filtering model, named Eigenapp, for recommending Apps in their Web site Getjar. Karatzoglou et al. proposed a novel context-aware collaborative filtering algorithm based on tensor factorization for mobile App recommendation, which named Djinn model. Indeed, detecting the rating and

comment spam is also an important application of recommender systems. For example, Lim et al. have identified several representative behaviors of review spammers and model these be haviors to detect the spammers. Wu et al. have studied the problem of detecting hybrid shilling attacks on rating data based on the semi-supervised learning algorithm. Xie et al. have studied the problem of singleton review spam detection. Specifically, they solved this problem by detecting the co-anomaly patterns in multiple review comments based time series. Although most of the previous studies leveraged the popularity information in their applications, none of them can comprehensively model the popularity observations. To this end, in this work we proposed a PHMM model for popularity modeling of mobile Apps, which can be exploited for most of above applications. Maiorana et al. have applied the HMM into biometrics with application of online signature recognition. Yamanishi and Maruyama proposed to leverage HMM for network failure detection by estimating the anomaly sequences of system logs. Different with above works, in this work, we introduce a novel application, namely popularity modeling for mobile Apps, by extending the HMM model with multiple popularity observations.

3. B2B PURCHASE PATTERN ANALYSIS

Since B2B purchases are often involved with strategic development of the company, and as a result, extra cautions and extensive research efforts have to be taken in making such investment, the decision process of customers in purchasing certain products or services is much more complicated than that in our daily purchasing activities. Thus, it is of significant business value if we can discover characteristic and critical buying paths from observations. These can be used to recommend directed advertising campaigns so as to increase potential profits and also reduce the marketing cost. In addition, we would also like to visually display the buying processes of the customers. By doing this, we can better understand the customer behavior patterns and accordingly develop promising marketing strategies. Apply our method to find critical buying paths of Business-to-Business (B2B) buyers from historical customer event sequences. Since B2B purchases are often involved with strategic development of the company, and as a result, extra cautions and extensive research efforts have to be taken in making such investment, the decision process of customers in purchasing certain products or services is much more complicated than that in our daily purchasing activities. Thus, it is of significant business value if we can discover characteristic and critical buying paths from observations. These can be used to recommend directed advertising campaigns so as to increase potential profits and also reduce the marketing cost. In addition, we would also like to visually display the buying processes of the customers. By doing this, we can better understand the behavior patterns of B2B customers and accordingly develop promising marketing strategies.

3.1 EMPIRICAL EVALUATION

A. SYNTHETIC DATA

We have simulated symbolic sequential data composed of stages of events. We define 5 stages {A, B, C, D, E}, where each contains 25 symbols. Then, we create 5000 sequences that are of two patterns. The first 2500 sequences mainly follow stage pattern A → B → C → D; the other 2500 sequences follow B → E → C. The simulation proceeds as follows. After deciding which stage to sample from based on the two patterns, we randomly pick d symbols from that stage, where d is a random integer. Then, we inject the selected symbols into the sequence, and continue to the next stage in the pattern. Indeed, such a simulation process is equivalent to a standard where 5 stages correspond to 5 hidden states and symbols within each stage correspond to observations. Let the transition probability from each stage to itself be p, and that to the next stage (as specified in the two patterns) be 1 − p. Then, the stage duration d follows a geometric

distribution d ∼ (1 − p)pd−1, with the expected value E[d] = 1/1−p . To have significant stage-wise patterns in the produced sequences, we have used a large probability p = 14/15 , leading to E[d] = 15. In other words, on average, we randomly pick 15 symbols for each stage.

B. BASELINES

First, we apply state-of-the-art Frequent Sequence Mining (FSM) algorithms, including GSP, SPADE, PrefixSpan, SPAM. The results in Figure 1 show that, when desired pattern support drops, the time consumption of these algorithms grow super-exponentially, indicating the difficulties introduced by the large numbers of symbols. The number of detected patterns also becomes explosive, most of which are non-informative and provide no clear insight of the underlying sequence generating processes. In comparison, using the temporal clusters identified via our approach, the mining process succeeds quickly in one second.

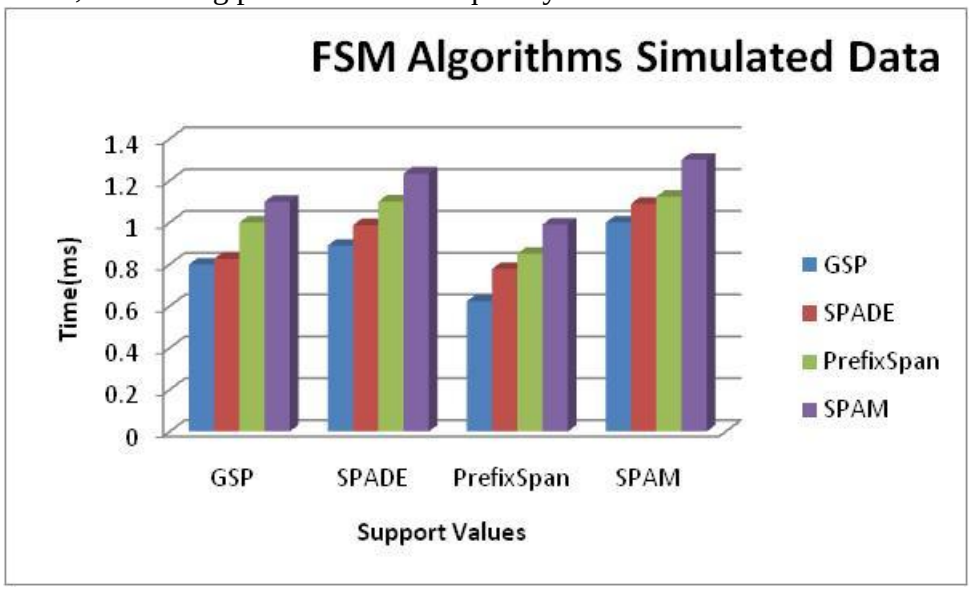

Figure 1: Simulations of Frequent Sequence Algorithms

Task	Pattern Mining	Sequence Clustering	Stage Recovery
Method	FSM	HMM	HMM
Precision	0.725	0.99	0.49
Recall	0.174	0.99	0.45

Table 1: Utility Comparison

In addition to the improvement on efficiency, we also compare the pattern mining results on the original and the re-encoded sequences via our method in Table I. For the task of pattern mining, we compute the precision and recall of the discovered patterns against the ground truth. The results show that when working on the raw data, FSM performs poorly with an Fmeasure around 0.281. In contrast, after re-encoding using our approach, it can lead to an 100% accuracy.

CONCLUSION

The key idea is to map the temporal structures of sequences into the topologies of a graph in a way that the temporal contents of the sequential data are preserved in the so-called temporal graph. Indeed, the embedding topology of the graph can allow to translate the rich temporal content into the metric space. Such a transformation enables not only sequential pattern mining

at a more informative level of granularity, but also enables new possibilities to explore, quantify, and visualize statistically relevant temporal structures in the metric space.

REFERENCES

[1] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. Data Mining and Knowledge Discovery, 15(1): 55–86, 2007.

[2] Chuanren Liu, Yong Ge, Hui Xiong, Keli Xiao, Wei Geng, and Matt Perkins. Proactive workflow modeling by stochastic processes with application to healthcare operation and management. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014.

[3] Chuanren Liu, Kai Zhang, Hui Xiong, Geoff Jiang, and Qiang Yang. Temporal skeletonization on sequential data: patterns, categorization, and visualization. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 1336–1345. ACM, 2014.

[4] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. Advances in neural information processing systems, 2002.

[5] Lane MD Owsley, Les E Atlas, and Gary D Bernard. Automatic clustering of vector time-series for manufacturing machine monitoring. In Acoustics, Speech, and Signal Processing, 1997. ICASSP-97.

[6] Ramakrishnan Srikant and Rakesh Agrawal. Mining generalized association rules. In VLDB, volume 95, pages 407–419, 1995.

[7] Ramakrishnan Srikant and Rakesh Agrawal. Mining sequential patterns: Generalizations and performance improvements. In EDBT, volume 1057 of Lecture Notes in Computer Science, pages 3–17. Springer, 1996.

[8] Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 290:2319–2323, 2000.

[9] Shuicheng Yan, Dong Xu, Benyu Zhang, Hong-Jiang Zhang, Qiang Yang, and Stephen Lin. Graph embedding and extensions: a general framework for dimensionality reduction. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 29(1):40–51, 2007.

[10] Mohammed J Zaki. Spade: An efficient algorithm for mining frequent sequences. Machine learning, 42(1-2): 31–60, 2001.