A Survey on Frequent Pattern Mining Techniques in Sequence Data Sets

Kirti Mirgal
PG Student
Department of Computer Engineering
Pillai Institute of Information Technology, Engineering, Media Studies and Research, Panvel, India

Dr. Satishkumar Varma
Associate Professor
Department of Information Technology
Pillai Institute of Information Technology, Engineering, Media Studies and Research, Panvel, India

Abstract

Finding interesting patterns from large amounts of data is an important task of data mining. There are many data mining tasks, such as classification, clustering, association rule mining, and sequential pattern mining. Sequential pattern mining extracts frequent subsequences from a sequence database which is helpful in making predictions, improving usability of systems, detect events and in making strategic product decisions in applications such as web user analysis, stock trend prediction, DNA sequence analysis. Subsequences can be contiguous and non-contiguous. Moreover the mining algorithm is classified into three categories: periodic patterns, statistically patterns, and approximate patterns. This paper discusses about few such pattern mining algorithms.
Keywords: CloSpan, cSPADE, Data Mining, Random Projections, Sequential Pattern Mining, Subsequences

I. INTRODUCTION

Sequences are an important kind of pattern which occur frequently in many fields such as medical, business, financial, customer behavior, educations, security, and other applications. Sequential pattern mining algorithms can be applied on various types of data such as transaction databases, sequence databases, streams, strings, spatial data, graphs, etc. The analysis of the data in these applications needs to be carried out in different ways to satisfy different application requirements, and it needs to be carried out in an efficient manner [1].

Sequential pattern mining handles data in large sequential data sets. It has gained popularity in in marketing in retail industry, biomedical research, DNA sequence patterns, financial industry. Most common applications are discovery of motifs in DNA sequences, identifying interesting share price movements in financial industry, analysis of web log for web usage, customer shopping sequences and the investigation of scientific or medical processes and so on.

Sequential pattern mining is find the relationships between occurrences of sequential events for looking for any specific order of the occurrences. In the other words, sequential pattern mining aims at finding the frequently occurred sequences to analyse the data or predict future data or mining periodical patterns [2][3]. It uses support as the criteria to evaluate frequency but is not efficient to discover some patterns. The problem of mining sequential pattern can be partitioned into three categories: periodic patterns, statistically patterns, and approximate patterns [4]. In this paper we will discuss a few approximate pattern mining techniques such as CloSpan[5] , cSPADE [6], and Random projections[7].

II. CATEGORIES OF SEQUENTIAL PATTERN MINING

The problem of mining sequential pattern can be partitioned into three categories: periodic patterns, statistically patterns, and approximate patterns [4].

A. Periodic Patterns

This model is not flexible and it is unable to find patterns whose occurrences are asynchronous [8]. Periodicity detection in time series database is an important data mining problem and has a number of applications. For example, "The gold prices increase every weekend" is a periodic pattern. As this model is restrictive it may fail to detect some interesting pattern if its occurrence is misaligned due to noise events. A pattern can be partially represented to provide a more flexible model. For example, pattern length four (I_1, *, *) is a partial pattern indicating that the frist symbol must be I_1. Behaviour of the system may change over time and some patterns may not be present all the time.

Namely two parameters min-rep and max-dist, are used to specify the minimum number of occurrences required within each subsequences and the maximum disturbance between any two successive subsequences. Given a sequence S, the parameters min-rep and max-dist and the maximum period length L_{max}, we can find the valid subsequences that have the most repetitions for each valid pattern whose period length does not exceed L_{max} in three phases. If parameters are not set properly, noise may be qualified as a pattern. The following three phases outline algorithm for mining periodic patterns in brief [9].

The first phase: For each symbol I, the distance between any two occurrences of I are examined and then for each period l, the set of symbols whose number of times are at least min-rep are sent to the next phase. Since there are a huge number of candidates, a pruning method is needed to reduce it.

The second phase: In this phase, the single patterns (1-pattern) are generated. For each period l and each symbol I a candidate pattern (I,*,*,…,*) is formed that number of symbol * is (l-1).

The third phase: After discovering the single patterns in previous phase, i-patterns are generated from the set of valid (i-1)-patterns and then these patterns are validated. In this phase, we can apply some heuristics. For example, it is obvious that if a pattern is valid, then all of its generalizations are valid. Pattern (I_1, I_2*) is a generalization of pattern (I_1,I_2,I_3).

B. Statistically Significant Patterns

For sequential patterns support and confidence are the two important measures. The support evaluates frequencies of the patterns and the confidence evaluates frequencies of patterns in the case that sub-patterns are given. In some applications, the frequent occurrence of a pattern may not be of significance importance and interesting as the few occurrences of an expected rare pattern. This pattern called surprising pattern instead of frequent pattern. The information gain metric used in the information theory field, may be useful to evaluate the degree of surprise of the pattern [10]. Target is finding set of patterns that have information gain higher than minimum information gain threshold.

Given a pattern P=(I_1,I_2,…,I_l) and an information gain threshold min-gain, the task is to find all patterns whose information gain in the sequence S exceed the min-gain value. There are some heuristics and methods that user can set the value of this threshold.

Information gain of pattern P is defined as follows:

$$\text{Info-gain}(P)=\text{Info}(P)*\text{Support}(P) \quad (1)$$

$$\text{Info}(P)=\text{Info}(I_1)+\text{Info}(I_2)+\ldots+\text{Info}(I_l) \quad (2)$$

$$\text{Info}(I_k)=-\log |I|^{\text{prob}(I_k)} \quad (3)$$

Where prob(I_k) is probability that symbol I_k occurs and |I| is number of events in S.

The main strategy to tackle the problem of mining statistically significant patterns is a recursive method which at the ith level of recursion examines the patterns with i events. In some applications, users may want to find the k most surprising patterns. On the other words, users may want to find top k patterns whose information gain is greater than a threshold. However, the major limitation of information gain value is that it does not recognize location of the occurrences of the patterns.

C. Approximate Patterns

Noisy data are common properties of large real world databases. A random error or variance in a measured variable is noise. The presence of noise can prevent the occurrence of a pattern and may not be recognized. moreover large patterns are more vulnerable to distortion caused by noise so it is necessary to allow some flexibility in pattern matching. The previous models only consider exact match of the pattern in the data. An approximate pattern is defined as a sequence of symbols which appears more than a threshold under certain approximation types in a data sequence. To resolve the approximate pattern mining problem, the concept of compatibility matrix is introduced [4]. This matrix provides a probabilistic connection from observed values to the true values. Based on the compatibility matrix, real support of a pattern can be computed.

A new metric, namely match is defined to quantify the significance of a pattern. The combined effect of support and match may need to scan the entire sequence database many times. Similar to other data mining methods, to tackle this problem sampling based algorithms can be used. Consequently, the number of scans through the entire database is minimized.

Given a pattern P=(I_1,I_2,…,I_{lp}) and a symbol sequence S=(I'_1,I'_2,…,I'_{ls}) where ls≥lp, match of P in S (denoted by M(P,S)) is defined as the maximal conditional probability P in every distinct subsequence of length lp in S. Eq. (4) show how compute match provided that each observed symbol is generated independently.

$$M(P,S)=\max_{s\in S} M(P,s) \quad \text{if } ls>lp$$

$$M(P,s)=\text{prob}(P|s)=\Pi_{1\le i\le lp} C(I_i,I'_i) \quad \text{if } ls=lp \quad (4)$$

III. APPROXIMATE PATTERN MINING TECHNIQUES

A. Mining of Closed Sequential Patterns

Ramin Afshar proposed a frequent closed subsequence mining approach CloSpan [5] that mines large sequences efficiently. This algorithm produces number of efficiently search pruning techniques. The algorithm makes use of hash technique that has two steps to carry out efficient optimization of the search space: 1) it create a superset of joint common sequences known as the LS set, and keeps the set in prefix order and 2) then it performs post-pruning to eliminate non-closed sequences. It works in following manner:

- Classification is performed on each set of item and carries out the elimination of not frequent items and sequences that are empty.

- The Clospan method is recursively applied on the prefix search tree in depth first search manner and builds the prefix sequence corresponding to it. Lastly, it removes the free sequences.
- Then is uses a hash index on the projected database size and only the sequences whose projected database size is same as that of current sequence are tested.

This algorithm performs well for exact pattern matching problems that is not suitable for approximate pattern matching problems. As the dataset size increases, execution time of the algorithm also increases rapidly.

B. Sequence Mining in Domain Categories

Mohammed J. Zaki proposed cSPADE [6] algorithm for mining frequent sequences. It is an efficient algorithm based on a number of syntactical limitations. They are size of the sequences, limiting the min or max gap on consecutive sequence elements, to put a time slot on acceptable sequences and searching sequences that are predictive of one or multiple classes, even rare ones. This algorithm methodically searches the sequence grid formed by the subsequence relation, from the simplest items to the nearly particular frequent sequences in a depth-first (or breadth-first) manner. It requires preprocessing of data in a special format, as it is based on syntactical constraints. For large database more time is needed for pre-processing the data as a result the performance degrades.

C. Motifs Mining using Random Projections

J. Buhler has proposed an algorithm based on random projections [7] of input substrings. It carries out a number of tests on a basic iterant. The Projection algorithm has two parts:

- A random projection is selected and hashed with each l-mer x in the input sequences to its hash bucket with every test.
- The required pattern is searched in a hash bucket that has adequate entries by applying an order of improving steps

This algorithm is more productive in searching required pattern in simulated data, but it needs improvement in time, in space and maybe for future more complex biological data and real time.

IV. CONCLUSION AND FUTURE SCOPE

In this work, we provide a brief overview of models of sequential patterns in [4]. The paper theoretically shows the three types of sequential pattern model and some properties of it. These models fall into three categories called periodic pattern, statistically pattern, and approximate pattern. The first model is rigid but provides full periodicity and partial periodicity. In former, every time point contributes to the cyclic behavior of a time series. In contrast, in partial periodicity, some time points contribute to the cyclic behavior of a time series. Use of information gain as new metric helps to find surprising patterns which comparing with the frequent patterns demonstrates the superiority of surprising patterns. The third model, approximate sequential pattern, provides a means to verify noise. But still being and active research area in the data mining field. The ColSpan[5], cSPADE[6], Random Projections[7] are approximate pattern mining techniques which allows to find the approximate patterns.

REFERENCES

[1] Dong G., and Pei J., Sequence Data Mining, Springer, 2007.
[2] Agrawal R., and Srikant R., "Fast Algorithms for Mining Association Rules", 20th Int. Conf. Very Large Data Bases, VLDB, Morgan Kaufmann, 1994, 1994, pp. 487-499.
[3] Agrawal R., and Srikant R., "Mining Sequential Patterns", 11th Int. Conf. on Data Engineering, IEEE Computer Society Press, Taiwan, 1995, pp. 3-14.
[4] Wang W., and Yang J., Mining Sequential Patterns from Large Data Sets, Springer, 2005.
[5] X. Yan, J. Han, and R. Afshar, "CloSpan: Mining Closed Sequential Patterns in Large Datasets," Proc. SIAM Int'l Conf. Data Mining (SDM), 2003.
[6] M.J. Zaki, "Sequence Mining in Categorical Domains: Incorporating Constrains," Proc. Ninth Int'l Conf. Information and Knowledge Management (CIKM), pp. 442-429, 2000.
[7] J. Buhler and M. Tompa, "Finding Motifs Using Random Projections," J. Computational Biology, vol. 9, no. 2, pp. 225-242, 2002.
[8] Zhao Q., and Bhowmick S. S., Sequential Pattern Mining: A Survey, Technical Report, CAIS, Nanyang Technological University, No. 2003118, Singapore, 2003.
[9] Wang W., and Yang J., Mining Sequential Patterns from Large Data Sets, Springer, 2005.
[10] Yang J., Wang W., Yu P. S., and Han J., "Infominet: Mining Surprising periodic Patterns", In Proceeding of the 7th ACM Int. Conf. on Knowledge Discover and Data Mining (KDD), 2001, pp. 395-400.