International Journal For Technological Research In Engineering
Volume 6, Issue 3, November-2018 *ISSN (Online): 2347 - 4718*

A SURVEY ON SEQUENTIAL RULE MINING TECHNIQUES

Girivar Modi[1], Dr. Sanjay Bansal[2], Mr. Anil Patidar[3]

Abstract: Data mining is also renowned as knowledge discovery in databases, has been recognized as the process of extracting non-trivial, inherent, previously unknown, and potentially useful information from data in databases. The exposed knowledge can be employed in numerous ways in analogous applications. The most vital tasks in data mining are the procedure of determining association rules and frequent item-sets. There is a very vital role of frequent item-sets mining in association rules mining. In the last few years, a range of approaches for uncovering frequent item-sets in especially huge databases have been emerged. Although there have been a large number of algorithms designed for frequent pattern mining, investigating efficient and scalable algorithms is still very challenging. In this paper, we have provided a survey of various sequential rule mining approaches.

Index Terms: Data Mining, Association Rule, Sequential Patterns, Knowledge Discovery, Frequent Item-sets.

I. INTRODUCTION

As computers are applied extensively in lots of areas, great amounts of data and information have been gathered and stored in the database constantly. This variety of data comprises the transaction accounts data in supermarkets, stock markets, banks, and telephone organizations. With the growing size of the stored data, one key concern is to obtain the valuable information from the huge data set. Data mining, as well famous as knowledge discovery in databases, is an area of study which mines inherent, comprehensible, earlier unidentified and potentially valuable information from data [1]. If we attempt to imprison this concept into a formal definition, then we can define data mining as "the analysis of (often large) observational data sets to find unsuspected relationships and to sum up the data in fresh ways that are mutually clear and valuable to the data vendor". Clearly, novelty which remains an open research problem must be measured relative to the user's preceding knowledge. Unluckily, certain data mining algorithms take a user's prior knowledge into account. While newness is a significant property of the relationships we seek, it is not sufficient to qualify a relationship as being important result. Data mining involves an integration of techniques from multiple disciplines such as statistics, database technology, high-performance computing, machine learning, neural network, pattern recognition, information retrieval, data visualization, spatial/temporal data analysis and image & signal processing. By means of performing data mining, regularities, interesting knowledge or high- level information can be extracted from databases and viewed or browsed from distinct angles. The exposed knowledge can be applied to decision making and information management. Therefore, data mining is considered one of the most promising interdisciplinary developments in the information industry [2].

The paper is further organized as follows- in the next section we categorized the data mining, section 3 shows a short survey of sequential rule mining algorithms and at the end we concluded the research work.

II. CATEGORIES OF DATA MINING

There are six categories of data mining [3] as shown in figure- 1. These are explained in next sections.

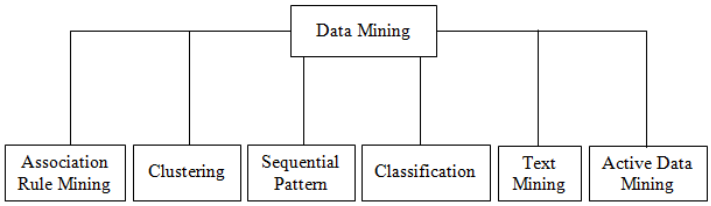

Figure- 1: Categories of Data Mining

2.1 Association Rule Mining:

Association rule mining [4] is a well-liked facts detection method for determining relations among items from a transaction records / database. Officially, a transaction database D is described like a set of transactions T = {t1, t2, … , tn} and a set of items I = {i1, i2, … , in}, where, t1, t2, …, tn ⊆ I. The support of an item-set X ⊆ I for a database is represented by sup(X) and is computed as the amount of transactions that comprises X. The difficulty of drawing out association rules from a given transaction database is to discover all association rules X→Y, where X, Y ⊆ I, X∩Y=⊆, and that the rules pay attention on certain nominal interestingness norms [5].

2.2 Clustering:

The objective of clustering is to recognize homogeneous sets of objects founded on the values of their attributes. For a specified collection of objects, the job of clustering is to collect some objects mutually in such a manner that the objects in the one cluster are close (based on some feature) to each other than the objects in another cluster. The methods of solving clustering difficulties go down into 2 categories: hierarchical and partitional. K-means approach is extensively employed in case of partitional clustering. In this method primarily decides K cluster heads, and then allocates each object to the cluster with its head closest to the object in such a manner that the summation of the square of distance among the objects and their head is diminished. Hierarchical clustering approach, it generally begins with putting each object in its personal cluster and then joins these tiny clusters into bigger and bigger clusters while all objects are in a particular one cluster [6] [7].

2.3 Classification:

For a database say D, having various records and each record comprises of various attributes. One of the attribute from

 Copyright 2018.All rights reserved.

these all attributes, perform the job of class label. This type of database D is known as a training set. An example of training data-set that have 3 attributes (age, risk and car type) is shown in Table- 1. Here, the attribute named risk (it has two values- high or low) is the class label [8].

Table- 1 : Example of Training Data-Set

Age	Car Type	Risk
23	Family	High
17	Sports	High
43	Sports	High
68	Family	High
32	Truck	Low
20	Family	High

2.4 Sequential Patterns:

If a set of data sequences is given, in which each sequence is a list of transactions ordered by the transaction time, the problem of mining sequential patterns is to discover all sequences with a user specified minimum support. Each one transaction includes a set of items. An ordered sequence or list of item-sets is known as sequential pattern. The item-sets that are contained in the sequence are called elements of the sequence. For a specified database say D, which comprises of customer transactions, furthermore every of the transaction comprises- customer-ID, items and the time stamp fields. An item-set is a nonempty set of items and as well a sequence is an ordered list of item-sets. Then we say that a sequence A <a1, a2, a3, …, an> is enclosed in another sequence B <b1, b2, b3, …, bn> if there exist integers i1<i2<i3<…<in, such that a1⊆bi1, a2⊆bi2, …, an⊆bin. For example, the sequence < (3) (4,5) (8)> is contained in <(7) (3,8) (9) (4,5,6) (8)>, because (3) ⊆ (3,8), (4,5) ⊆ (4,5,6), and (8) ⊆ (8). A customer sequence is a sequence of item-sets for each customer-ID. The support of a sequence s is defined by the following equation [3].

$$\text{Support} = \frac{\text{The no. of sequence that contains this sequence}}{\text{The total number of sequences}}$$

For the example data set in Table- 2, we can find all the sequences that have support > 25%, which are < (30) (90) >, and < (30) (40, 70) >.

Table- 2 : Sequences of an Example Database

Customer	Customer Sequence
1	< (30) (90) >
2	<(10, 20) (30) (40, 60, 70)>
3	< (30, 50, 70) >
4	< (30) (40, 70) (90) >
5	< (90) >

The objective of sequential patterns is to get the sequences that have larger than or equal to a definite user specific support. Usually the process of finding sequential patterns consists of the following phases: sorting phase, finding the large item-set phase, transformation phase, sequence phase and maximal phase.

2.5 Text Mining:

In the actuality, it is very general to discover the hidden associations in the bigger text database. For instance, it is very helpful to discover whether an organization is shifting its attention from one field to a further. Text mining is incredibly good at managing these circumstances because the database is filled with text, in place of the numeric data. The goal of text mining is to discover the trends in the text database [9] [10]. Lent [11] developed a system to discover the trends in text database. Their basic idea is to use the existing data mining algorithms and shape query language (SDL) [12].

2.6 Active Data Mining:

Active data mining joins the expertise of active database and data mining. The fundamental thought is to split the whole data into a number of sub-datasets. The data mining techniques will be utilized to every sub-dataset and the outcome for every sub-dataset will be created. Since, typically the data-sets arrive from data warehousing and consequently the amount of data is extremely vast, isolating the data-set to a number of sub-datasets will not mislay the implication of the outcome. All the rules created in every sub-data set will be collected in the rule database with some factors for instance- confidence and support for association rule difficulties. When the fresh data arrive and the amount of fresh data achieved a definite level, the data mining technique will be utilized to the fresh data-set again and afterward confirm the rule database. If one definite rule doesn't exist, it immediately keeps the rule within the rule database [13]. If this rule be present in the rule database, it will revise the values of the parameter of the obtainable rules in the rule database. To watch the parameters, the database utilizes trigger in the rule database. When criterion is met, the definite trigger will be executed.

III. LITERATURE SURVEY

In this section we presented a literature survey of some existing work on sequence rule mining approaches.

Sequential rule mining have been utilized in many domains like stock exchange analysis (Das et al. [14], Hsieh et al. [15]), weather observation (Hamilton et al. [16]) and drought management (Harms et al. [17], Deogun et al. [18]). Agrawal et al. [8] stated the AIS (Agrawal, Imielinski, Swami) algorithm that was the precursor of entire algorithms employed to produce the confident association rules and frequent item-sets. This approach contains two phases. In primary phase of algorithm constitutes the creation of the frequent item-sets. Subsequent to this, in next phase the creation of the confidence and frequent association rules is takes place. The utilization of the monotonicity characteristic of the support of item-sets and the confidence of association rules go ahead to the improvement of the approach and it was later known as Apriori by Agrawal et al [29][30].

A most renowned technique given by Mannila et al. [19] for

 Copyright 2018.All rights reserved.

sequential rule mining and also further researchers subsequently that aspires at determining partly ordered groups of events arriving repeatedly in a time window in a series of events. For a given set of "frequent episodes", any technique can obtain sequential rules relating to a minimal confidence and minimal support. Sequential rules are of the form X⇒Y, in which X and Y are 2 sets of events, and are explained like- "if event(s) X come out, event(s) Y are expected to arise with a prearranged confidence subsequently". On the other hand, their task can solely find out rules in a particular sequence of events. Other efforts that mine sequential rules from a particular sequence of events are the methods of Hamilton et al. [16], Hsieh et al. [15] and Deogun et al. [18], which correspondingly find out rules amid numerous events and a solitary event, among two events, and among numerous events.

The PrefixSpan full form is- Prefix-projected Sequential pattern mining approach is offered by Jian Pei et al. [20]. This approach corresponds to the pattern-growth approach, which locates the frequent items after inspecting the sequence data base. In this, the data base is projected in accordance with the frequent-items, into numerous tiny sized databases. Lastly, the full set of sequential-patterns is set-up via recursively rising subsequence portions in each predictable data-base. Though the PrefixSpan approach fruitfully exposed patterns by utilizing divide & conquer policy, but it takes high memory cost owing to the formation and processing of massive number of predictable associate databases.

IV. CONCLUSION

In this paper, we have provided a survey of various sequential rule mining approaches. There are several researchers have worked in this area and also several working with the previously suggested work to improve them. The majority of the earlier suggested approaches adopted Apriori approach idea and also of others works, but mostly of them not competent with huge dataset and not provides rules as per user necessities. So, still there is a need of an approach that works efficiently with required size of dataset for mining sequential rules..

REFERENCES

[1] S. Chakravarthy and H. Zhang, "Mining And Visualization Of Association Rules Over Relational DBMSs", SAC '03 Proceedings of the 2003 ACM symposium on Applied computing, Pages 922-926, 2003.

[2] Marek Maurizio, "Data Mining Concepts and Techniques", E-commerce, (http://www.dsi.unive.it/~marek/files/06%20-%20datamining), 2011.

[3] S. Chakravarthy and H. Zhang, "Mining And Visualization Of Association Rules Over Relational DBMSs", SAC '03 Proceedings of the 2003 ACM symposium on Applied computing, Pages 922-926, 2003.

[4] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules between Sets of Items in Large Databases", ACM-SIGMOD Int. Conf. Management of Data, Washington, D.C., pp 207–216, 1993.

[5] Philippe Fournier-Viger Usef Faghihi, Roger Nkambou, and Engelbert Mephu Nguifo, "CMRULES: An Efficient Algorithm for Mining Sequential Rules Common to Several Sequences", Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS 2010), 2010.

[6] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", Proc. of the ACM SIGMOD Int' l Conference on Management of Data, Seattle, Washington, June 1998.

[7] Alessia Amelio and Andrea Tagarelli, "Data Mining: Clustering", Encyclopedia of Bioinformatics and Computational Biology, Elsevier (in press), 2017.

[8] Fabricio Voznika, and Leonardo Viana, "Data Mining Classification", Retrieved from:http://courses.cs.washington.edu/courses/csep521/07wi/prj/leonardo_fabricio.pdf.

[9] Mehdi Allahyari, "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques", KDD Bigdas, Halifax, Canada, August 2017.

[10] Shilpa Dang and Peerzada Hamid Ahma, "Text Mining: Techniques and its Application", International Journal of Engineering & Technology Innovations, Vol. 1, Issue 4, November 2014.

[11] B. Lent, R. Agrawal, and R. Srikant, "Discovering Trends in Text Databases", Proc. of the 3rd Int'l Conference on Knowledge Discovery in Databases and Data Mining, Newport Beach, California, August 1997.

[12] R. Agrawal, G. Psaila, E. Wimmers, and M. Zait, "Querying shapes of histories", Proceedings of the 21st International Conference on Very Large Databases, Zurich, Switzerland, September 1995.

[13] R. Agrawal, G. Psaila G., Active Data Mining, Proc. of the 1st Int'l Conference on Knowledge Discovery and Data Mining, Montreal, August 1995.

[14] G. Das, K. I. Lin, H. Mannila, G. Renganathan, and P. Smyth, "Rule Discovery from Time Series", Proc. 4th Int. Conf. on Knowledge Discovery and Data Mining, 1998.

[15] Y. L. Hsieh, D. L. Yang and J. Wu, "Using Data Mining to Study Upstream and Downstream Causal Relationship in Stock Market", Proc. Joint Conference on Information Sciences, 2006.

[16] H. J. Hamilton and K. Karimi, "The TIMERSII Algorithm for the Discovery of Causality", Proc. 9th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 744-750, 2005.

[17] S. K. Harms, J. Deogun and T. Tadesse, "Discovering Sequential Association Rules with

 Copyright 2018.All rights reserved.

Constraints and Time Lagsin Multiple Sequences", Proc. 13th Int. Symp. On Methodologies for Intelligent Systems, pp. 373-376, 2002.

[18] J. S. Deogun and L. Jiang, "Prediction Mining– An Approach to Mining Association Rules for Prediction", Proc. of RSFDGrC Conference, 98-108, 2005.

[19] H. Mannila, H. Toivonen and A. I. Verkano, "Discovery of Frequent Episodes in Event Sequences", Data Mining and Knowledge Discovery, Kluwer Academic Publishers. Manufactured in The Netherlands, 259-289, 1997.

[20] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, and Helen Pinto, "PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth", Proceedings of the 17th International Conference on Data Engineering, 2001.

 Copyright 2018.All rights reserved.