ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Investigation on ARM Algorithm with Incremental Mining and Proposing a New Algorithm for Streaming Dataset.

Suresh, K. and Pattabiraman, V.

Research Scholar & Associate professor, School of Computing Sciences and Engineeringr, VIT University, Chennai, India

ARTICLE INFO

Article history:
Received 12 November 2014
Received in revised form 26 December 2014
Accepted 29 January 2015
Available online 10 February 2015

Keywords:
Association rule mining, Incremental mining, frequent mining, streaming dataset.

ABSTRACT

The function of Association Rule Mining (ARM) is to discover the relationship between one itemset to another itemset based on the rate of re-occurrence principles and it is a widespread topic in data mining. It directs the innovation of motivating associations in the midst of items in large transactional databases. There are several techniques used in association rule mining like incremental temporal mining, sequential incremental mining, updating temporal association mining and so on. In this paper we consider only the incremental algorithm to sustain the sequential association rules in a business database. Also presented incremental mining perception cannot be treating through temporal association rules. Moreover this observation of data mining is a significant subdivision of association rules. This research work represents the vital conception of association rules and the essential way to find out the pattern of mining association rules by using incremental mining technique. And concluded with the various aspects of analysis in transaction data in live streaming databases.

To Cite This Article: Suresh, K. and Pattabiraman, V.., Investigation on ARM Algorithm with Incremental Mining and Proposing a New Algorithm for Streaming Dataset.. ***Aust. J. Basic & Appl. Sci.,*** *9(6): 72-76, 2015*

INTRODUCTION

Data mining is the procedure of eliminating irrelevant information from massive data. Due to the improvement of information technology, there are numerous dissimilar varieties of knowledge records. Generally the information databases consist of industrial data, health data, economic data, and business transaction data. Towards this topic analyzer has to efficiently analyze the data to discover the significant information from huge databases. Data mining procedure have been broadly conversed and repeatedly used tool in modern decades. Not only its applications getting broader, but its computational efficiency and accuracy are also improving (James Bailey, Elsa Loekito, 2010).

Data mining is distinctive information finding from repository. Which it can be added to make clear as the progression of nontrivial data and potentially purposeful data from great databases (Chowdhury Farhan Ahmed, 2012). Various categories of techniques are required to get diverse kind of information (Unil Yun, 2013).

There are several algorithms used for determining association rules in transaction databases. These algorithms are extended and broadly considered. Apriori variants of mining algorithms is authentic to afford additional possible outcomes, such as incremental renew, generalized mining and different level rules, significant rules, mining of multi structure rules, various minimum supports, uncommon stuffs, and sequential association structures.

The objective of this paper is the proficient data formation to locate the widespread relationship between the attributes at peculiar stages in a classification ranking beneath the hypothesis. The unique recurrent itemized and relationship were produced in progressive. The most important dispute of proposing a well-organized mining algorithm is to build use of the innovative common itemsets and association rules to directly produce original comprehensive association rules, to a certain extent than rescanning the database. The aim of determining the optimal solution of the itemset is to influence transaction dataset. Most of the approaches do not validate the applicability of more complex methodologies, such as mining algorithms and computational models.

Related Work:

Tarek F. Gharib, Hamed Nassar and Mohamed Taha are describes about, incremental algorithm to put away the capable activist relationship rules in a

Corresponding Author: Suresh, K., Research Scholar & Associate professor, School of Computing Sciences and Engineeringr, VIT University, Chennai, India

business record. They proposed new algorithm which it remuneration from incremental procedure. This algorithm is derived from the concept of sliding-window filtering algorithm.

Chun-Wei Lin, Guo-Cheng Lan and Tzung-Pei Hong are discussing about, Pre-large sequences are defined by two thresholds, one is a lower support threshold and another is upper support threshold. These two medium will be take action as opening to evade the activities of succession in a straight line from huge data to undersized and vice versa. The benefits of this proposed algorithm no more required to repeat scanning in original databases.

Chun-Wei Lin, Guo-Cheng Lan and Tzung-pei Hong are explains about, an incremental mining algorithm for efficiently mining high utility itemsets is proposed to handle the drawbacks of two phase algorithm. This proposed algorithm is based on the concept of the fast update approach (FUP), which was in the beginning considered for organization mining. The proposed come close to first divider itemsets into four parts according to the far above the ground transaction-weighted consumption.

Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer and Byeong-Soo Jeong, this paper they proposed two novel tree structures IWFPTWA (Incremental WFP tree based on weight ascending order) and IWFPTFD (Incremental WFP tree based on frequency descending order). The proposed methods are effective for incremental and interactive mining in tree structure enhancement. This research work contributed on major issues like high compact to save memory space in incremental tree structure data.

James Bailey and Elsa Loekito in this paper, they proposed an resourceful method for incrementally mining difference models. This algorithm particularly aims to keep away from redundant computation which might occur due to concurrent transaction in placing and removing processes.

Proposed Work:

The customer sequence is to represent all the transactions of a particular customer based on the order of transaction times. Note that each transaction in a customer sequence corresponds to an itemset. Generally a sequence is prearranged directory of itemsets.

Table 1a: Sequence of transactions of customer buying behaviours.

Customer ID	*Customer Sequence*
1	*{(A)(B)}*
2	*{(C,D)(A)(E,F,G)}*
3	*{(A,H,G)}*
4	*{(A)(E,G)(B)}*
5	*{(B)(C)}*
6	*{(A)(B,C)}*
7	*{(A)(B,C,D)}*
8	*{(E,G)}*

Table 1b: Transactions of same cust_ id 5 and 9.

Cus_id	*Cus_sequence*
5	*{(E,G)}*
9	*{(E,F,G)}*

Table 1c: Merged transaction sequences of the same cus_id 5 and 9.

Cus_id	*Cus_sequence*
5	*{(B)(C)(E,G)}*
9	*{(E,F,G)}*

Sequences with their counts in the original database, the table 1(a) and table 1 (b) can be easily switched to separately. In addition in the protection stage, the relation of recently added client series to unique customer succession is habitually very undersized. This is more apparent when the database is growing larger. A sequence in table 1 (c) cannot possibly be large for the entire updated database in real-time applications. So when the quantity of recently added customer progression is small compared to the number of customer sequences in the original database.

Algorithm for mining streaming datasets:

Input: The unique database D, high transaction weighted consumption item sets through their authentic service values beginning the unique database, a utility table, each of all items in D with a profit value.

1. A set of *m* items I = $\{i_1, i_2,\ldots i, j,\ldots i_m\}$, each ij with a profit value pj, j=1 to m.
2. A transaction database D = $\{t_1, t_2,\ldots t_n\}$, in which each transaction includes a subset of items with quantities.
3. The average-profit threshold λ.

Output:

1. A set of utility itemset handled dynamic dataset and it is implemented in real-time time applications.
2. Steps:

1. Calculate the utility value uj_k of each item set in each transaction tk as $uj_k = qj_k * p_j$, where qj_k is the quantity of i_j in t_k for j =1 to m and k = 1 to n.
2. Find the maximal utility value mu_k in each transaction t_k as mu_k = max {$u1_k$, $u2_k$,…,um_k} for k=1 to n.
3. Divide the average utility.
4. Write down the profit value pj in ascending order.
5. Find the average of all the profit (avg(p_j)) values.
6. Partition the profit value p_j, into two sets namely Ka=(P_j) ≠Tid(p_i), 1≤ j≤m

7. Calculate the average utility upper bound ubj of each item i_j as the summation of the maximal utilities of the transactions which include i_j. That is

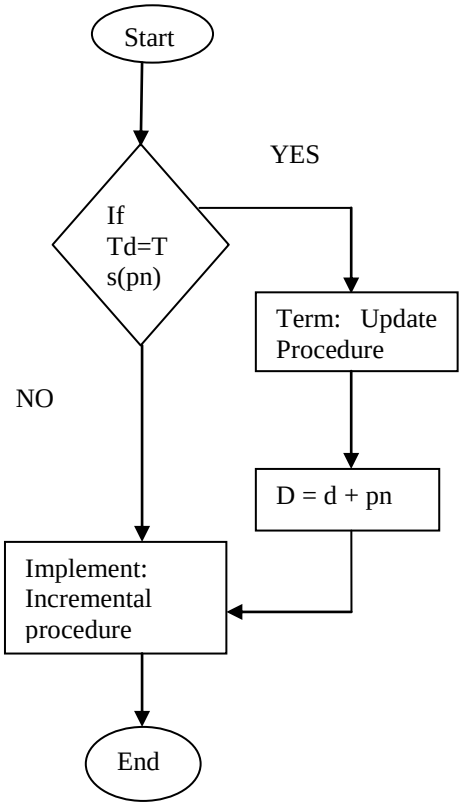
ubj =

$$ubj = \sum_{k=0}^{n} uka$$

8. Calculate the average utility lower bound ubj of each item i_j as the summation of the maximal utilities of the transactions which include i_j. That is

$$ubj = \sum_{k=0}^{n} ukb$$

Working flow of proposed technique:



Coding:

```
public static void main(String[] args) throws Exception {
 Apriori ap = new Apriori(args);
}
Private void go() throws Exception {
      //start timer
      long start = System. Current Time Millis();
      createItemsetsOfSize1();
      int itemset Number=1; //the current itemset being looked at
      int nb Frequent Sets=0;
            While (itemsets. size()>0)
      {           Calculate Frequent Itemsets();
        if (itemsets. size()!=0)
      {                 nb Frequent Sets+=itemsets. size();
            log("Found "+itemsets. size()+" frequent itemsets of size " +itemset Number + " (with support "+(min Sup*100)+"%)");;
            Create New Itemsets From Previous Ones();
            }
            Itemset Number++;
      }
       Long end = System. Current Time Millis();
      Log ("Execution time is: "+((double)(end-start)/1000) + " seconds.");
      Log ("Found "+nb Frequent Sets+ " frequents sets for support "+(min Sup*100)+"% (absolute "+Math. round(num Transactions*min Sup)+")");
```

```
      log("Done");
   }public class live stream {
  public static void main(String args[])
{
Stringurl="jdbc:msql://athens.imaginary.com:4333/db_web";
try {
    Connection con = Driver Manager. get Connection(url, "borg", "");
    Statement select = con. Create Statement();
    Result Set result = select. Execute Query
     ("SELECT key, val FROM t_test");
    System. out. Print ln("Got results:");
    while(result. next()) { // process results one row at a time
      int key = result. getInt(1);
      String val = result. get String(2);
      System. out. Print ln("key = " + key);
      System. out. Print ln("val = " + val);
    }
```

Experimental result using synthesis dataset:

In this section, the experiments for proposed work were carried out on synthetic dataset. The outcome of this work is used to estimate the accuracy of the ARM algorithm. The objective of this proposed method is to find relations involving in customers product buying in frequent visit to the retailer. The dataset contains 52 attributes and 125 transactions.

Experimental setup:

In this research work weka tool is used for analysis of large, complex, information, rich data set. This proposed work consists of various stages of data processing techniques such as data cleaning, data transformation, data reduction and discretization. The weka tool contains data pre-processing, classification, regression, clustering, association rules, and visualization.

System specifications:

This Weka tool are compatible and designed for enlarge novel machine learning methods. The proposed algorithm is applied in retail dataset and it is implemented in Java platform. It also includes the technical, performance, operational and support characteristics for the System as an entity.

Synthesis dataset used for analysis:

In this paper, retail business dataset is used for experimental results. To estimate the performance of incremental approach, this research work carry out testing on synthetic dataset. The accuracy and effective ARM rules of the proposed algorithm were found.

Chart:

In this chart the gathering temporal data are considered as a prime variable. When new points are updates to the large databases, it automatically generates a new rule to the existing dataset. The figure-1 shows the accuracy of transaction database. X-axis is considered as transaction data and Y-axis is considered as data accuracy. In this outcome the better accuracy result is found when compare to PMSE method.

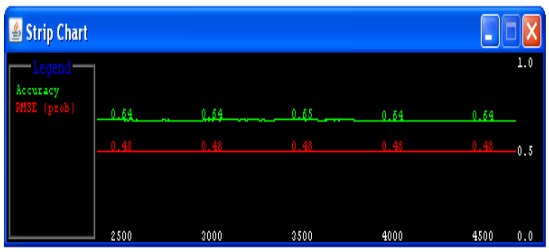


Fig. 1: Accuracy value for transaction database.

RESULT AND DISCUSSIONS

The algorithm consists of two types. One is Incremental process, which in real time it will performs the incremental mining and the second one is updating transaction table, which performs pre-processing for the Incremental procedure. The proposed algorithm initially verifies whether database is equal to the transactions of latest update in the original database. If yes, the algorithm makes some pre-processing by calling Update procedure then merging the transactions of last partition of original database with the incremental database before calling the incremental procedure. If no, the algorithm calls the incremental procedure of the ITARM algorithm directly.

Conclusion:

In this paper, a novel incremental mining algorithm accomplished of sustaining order patterns based on the concept of dynamic or streaming database is proposed. The number of rescans of original databases can be reduced. In this research work it overcomes the drawbacks of slide window filtering algorithm, two phases algorithm, re-computing large sequences method. It centre of attention on recently additional client series, which are misshapen from recently additional business, thus to a great extent shortening the amount of entrant series generation.

REFERENCES

Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Young-Koo Lee, Ho-Jin Choi, 2012. "Single-pass incremental and interactive mining for weighted frequent patterns", Expert Systems with Applications, 39: 7976-7994.

Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer, Byeong-Soo Jeong, Ho-Jin Choi, 2012. "Interactive mining of high utility patterns over data stream", Expert Systems with Applications, 39: 11979-11991.

Chun-Wei Lin, Guo-Cheng Lan, Tzung-Pei Hong, 2012. "An incremental mining algorithm for high utility itemsets", Expert Systems with Applications, 39: 7173-7180.

James Bailey, Elsa Loekito, 2010. "Efficient incremental mining of contrast patterns in changing data", Information Processing Letters, 110: 88-92.

Qiankun Zhao, Sourav S. Bhowmick, 2003. Association Rule Mining: A Survey, Technical Report, Center for Advanced Information Systems (CAIS), Nanyang Technological University, Singapore.

Tarek F. Gharib, Hamed Nassar, Mohamed Taha, Ajith Abraham, 2010. "An efficient algorithm for incremental mining of temporal association rules", Data and Knowledge Engineering, 69: 800-815.

Unil Yun, Heungmo Ryang, Keun Ho Ryu, 2013. "High Utility Itemset Mining with Techniques for reducing Overestimated Utilities and Pruning Candidates", Expert Systems with Applications.