International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 02, February 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

HEART DISEASE PREDICTION USING PSTREE

S.P. Siddique Ibrahim[1] , M.Pavithra[2], Dr. M. Sivabalakrishnan[3]

[1]Assistant Professor, Department of Computer Science and Engineering,
Kumaraguru college of Technology, Coimbatore (India)
[2]Department of Computer Science and Engineering,
Kumaraguru College of Technology, Coimbatore (India)
[3]Associate Professor School of Computing, Science and Engineering, VIT University, Chennai (India)

ABSTRACT

The data streams have modern technique to address the problems of continuous data. Mining with data streams is the process of extracting knowledge structures from continuous, rapid data records [1]. An important goal in data stream mining is mainly used to generate a compact representation of data. This algorithm useful in reducing time and space needed for further decision making process. In this paper a new scheme called Prefix Stream Tree (PST) for associative classification was proposed that helps in compact storage of data streams. This Pstree is generated based on a single scan. This Pstree discover the exact set of frequent itemsets from a single Scan.

Keywords: Data Streams, Data Stream Mining, Association, Classification.

I INTRODUCTION

Data mining is the process of examine large set of data and extracting hidden patterns from different data types in order to find previously unknown design. The discovery process can be an automatic or semi-automatic [1]. For decision making data mining is the knowledge discovery in the database and the KDD main steps are the data selection, data pre-processing, transformation, data mining, and evaluation. Data mining tasks including classification, clustering, association rule discovery, pattern recognition, regression, etc. [2]

There are two type of learning model available in data mining such as supervised and unsupervised. In supervised learning, it contain the class label .For example, in credit card scoring application, the goal is to whether the financial institution should issued a credit card to the client or not to the client. On the other hand, training data set with no class attribute is considered as unsupervised learning.

II RELATED WORK

2.1 Association

An association is an indicating expression of the form X→Y, where X and Y are disjoint item sets. Support determines how frequently a rule is applicable to given datasets, while confidence determines how frequently items in Y appear in transaction that contain X.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Example 3.1

Television→setup box [supp = 5%, confidence = 80%]

80% of customers who buy a television also buy a setup-box and 5% of customers buys all these products together.

Frequent item sets will be find from unstructured, semi structured, structured datasets. It also helps in finding the relationships between item sets.

The FP-growth mining technique [10] is one of the capable works where the performance gain is based on highly compact FP tree structure. This tree is constructed by having two database scans and on past threshold ability which restrict the usage of FP-tree on data streams. So in this paper we use a novel tree structure that constructs an FP–tree like compact prefix tree structure within a single pass.

2.2 Associative classification

Associative Classification (AC) is a natural classification learning advance in data mining that adopts association rule encounter methods and classification to build the classification models. Associative classification trust in mainly on two important thresholds called minimum support (*MinSupp*) and minimum confidence (*MinConf*). Minimum support produce the frequency of the attribute value and its associated class in the training data set from the capacity of that data set. Whereas minimum confidence produce the frequency of the attribute value.

2.3 CBA Working Model

step1: The rule will be generated based on class label in the training dada with the support value.

Step2: after the rule generation by using the minimum support count rules will be pruned

step3:classifier will be constructed based on the confidence value, if the confidence value will be equal means support count will be taken and if the support is also same means rule length will be taken for the classifier construction

Step4:this step will make a decision by comparing the rule generated with the training data

step5: Accuracy will be calculated based on the precision, recall measures.

III PROPOSED WORK

This section deals with the data stream representation, various window models, association, classification, associative classification.

3.1 Data Stream

Data Stream Mining is the process of separating knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using reserved computing and storage capabilities.

ISSN (O) 2454-9665
ISSN (P) 2454-0663

In many data stream mining applications, the goal is to predict the class or value of new instances in the data stream . Machine learning techniques can be used to learn this prediction task from labelled examples in an automated fashion. Data streams have following characteristics.

The data arrives continuously.

- No guess on data stream ordering can be made.
- During the mining process the memory usage should be limited. Knowledge must be attain as rapidly as possible.
- Each data element should be tested at most once and prepared as fast as possible because of memory limitation.
- application areas likes network traffic monitoring, web click-stream analysis, market basket data mining, fraud detection etc.,

3.2 Windowing Model

According to the data stream processing model the windows can be divided into three categories:

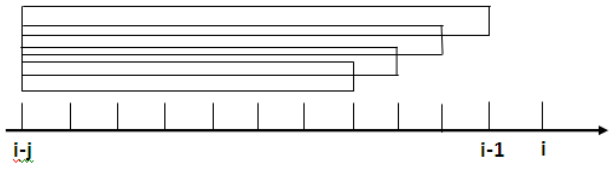
- Landmark-window based mining [5]
- Damped-window based mining[6]
- Sliding-window based mining [7]

as shown in Fig.1. A window is a subsequence between i-th and j-th arrived transactions, denoted as W[i, j]= (ti, ti + 1,...., tj), $i \leq j$.

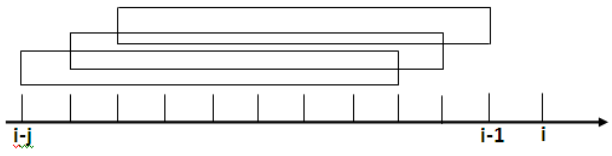
For the window based approach two naive methods will be used:

1) whenever the new transaction enter into the window or the old transaction leaves the window frequent item set would be regenerated.

2) store the frequent and non-frequent item set in the traditional data structure such as prefix tree and add its support count whenever the new transaction enter into the window or leaves the window.



(a) Landmark Window



(b) Sliding Window

Different data models are been proposed due to the nature of data streams [8], [9], [10]. This paper deals with the landmark window model. In this model, all data streams from the start time till current time are considered

ISSN (O) 2454-9665
ISSN (P) 2454-0663

for mining. As a stream arrives it is appended continuously as time grows. The second data model is based on sliding window. In this model only recent data streams which fall within a window are considered for mining. In our work used landmark window data model over data streams for prediction of heart disease.

IV PROBLEM STATEMENT

In data stream mining it is impractical to store all the data in limited memory or even in external memory. The another challenge in processing recent data to mine complete set of exact frequent item sets from data streams. Efficiently updating unused data with new data helps in reducing the memory usage which indirectly helps in increasing the performance. A compact data structure must be built with a single scan. This compact structure will help in having memory and time efficient mining. For processing recent data the compact tree constructed must be restructured. Motivated by these requirements, in this paper, proposed method called PSTree which efficiently addresses all the above challenges.

V. CONSTRUCTION OF PSTREE

The PSTree, which is based on prefix tree schema. It is an abstract and compact representation of data streams. As the window *W* slides the tree is updated. Each time the window *W* contains equal number of batches of transactions. Window slides batch by- batch.

5.1 Structure of the Prefix Stream tree

The PSTree is built based on nodes. First node of the tree is called as the root node which is introduce as "*null"*. Each subsequent node is called as ordinary node which represents the itemset and total number of passes (i.e., support) for that itemset in the path from the current window. The end nodes of the tree are leaf nodes which contains the support, class label and batch counter. two types of nodes are maintained in the tree.
The structure of non-leaf and leaf nodes.

5.2 Phases in construction of the tree

The tree will be constructed based on two phases: *Insertion* and *Restructuring*. Insertion phase catch the stream contents into tree based on arranged order *I-list*. Restructuring phase restructures the tree in descending orde from *I-list*. Restructuring is done after inserting a batch of action using Insertion phase. These two phases are dynamically executed one after the other.

5.3 Construction of tree

Consider the data streams shown in Table 1 which contain three attributes, A1 (a1, b1, c1), A2 (a2, b2, c2) and A3 (a3, b3, c3) and two classes (y1, y2). Assuming *minsupp* = 20% and *minconf* = 80%. Taking the help of landmark window for two batches where each batch contains two tuples, the PSTree is constructed using the concept of prefix tree. Initially a batch of two tuples is inserted using Insertion phase along with maintenance of

International Journal of Innovative Research in Science and Engineering
Vol. No.3, Issue 02, February 2017
www.ijirse.com
ISSN (O) 2454-9665
ISSN (P) 2454-0663

itemset list *I-list*. These insertion and restructuring phases are repeated one after the other for all consecutive batches. If all batches *Bj-1*, *Bj* in the current window *Wi* are inserted properly into the tree then the window will slide to the next batches Bj, Bj+1. While inserting the new batch *Bj+1*, the oldest batch *Bj* is deleted by changing the batch number.

TABLE I. TRAINING DATA

Tuble id	A1	A2	A3	class
1	a1	a2	b3	Y1
2	A1	A2	C3	Y2
3	A1	B2	B3	Y1
4	A1	B2	B3	Y1

I list {}

id
A1
A2
b3
C3

I sort- I list

id
A1:2
A2:2
b3:1
C3:1

from the table batch size of window is selected based on this first two dataset will be selected. first the tree will be empty and the support count will be calculated is shown in I sort list. after this insertion phase and restructuring phase will be done.

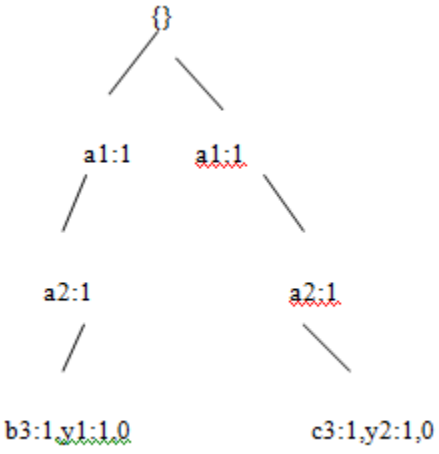


Fig 1. Insertion Phase

ISSN (O) 2454-9665
ISSN (P) 2454-0663

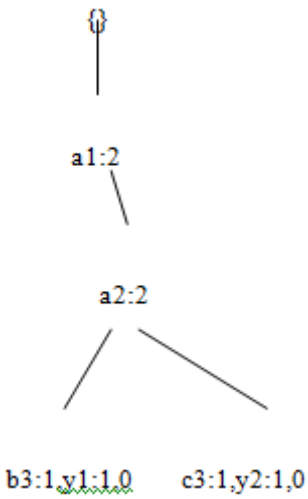


Fig 2. Restructuring Phase

The tree is refreshed all the time with the exact information about frequent itemsets along with rules is provided for the current window. In cases where a rule item is associated with multiple classes, only the class with the largest support count is considered. Restructuring of the tree can be done using either Path Adjusting method or Branch Sorting method proposed by [10] [11].

VI. CONCLUSIONS

PSTree which was composed using the concept of prefix tree and was restructured to handle the stream data. The constructed tree is a compact tree which reduces the memory consumption. It helps in finding exact set of recent frequent itemises and predicts the class label for the requested tuple and it also reduce the rule generation and to improve the performance.

REFERENCES

[1] Suzan Wedyan "review and comparison of associative classification data mining approaches", International Journal of Computer, Control, Quantum and Information Engineering Vol:8, No:1, 2014.

[2] Fayyad, U., and Irani, K. (1993) Multi—interval discretization of continues-valued attributes for classification learning. Proceedings of IJCAI, pp. 1022-1027. 1993.

[3] K.Prasanna Lakshmi, Dr.C.R.K.Reddy, "A Survey on Different Trends in Data Streams " pp.451-455, In Proc of 2010 IEEE International Conference on Networking and Information Technology, (ICNIT'10), 2010. ISBN : 978-1-4244-7577-3.

[4] Manku, G.S., & Motwani, R. (2002). Approximate frequency counts over data streams. In Proceedings of the 28th international conference on very large data bases, (pp. 346–357).

[5] J. H. Chang and W. S. Lee. Finding Recent Frequent Itemsets Adaptively over Online Data Streams. In Proc. of KDD, 2003.

[6] J. H. Chang and W. S. Lee. A Sliding Window method for Finding Recently Frequent Itemsets over Online

ISSN (O) 2454-9665
ISSN (P) 2454-0663

Data Streams. In Journal of Information Science and Engineering, Vol. 20, No. 4, July, 2004.

[7] Manku, G.S., & Motwani, R. (2002). Approximate frequency counts over data streams. In Proceedings of the 8th international conference on very large data bases, (pp. 346–357).

[8] J. H. Chang and W. S. Lee. Finding Recent Frequent Itemsets Adaptively over Online Data Streams. In Proc. of KDD, 2003.

[9] J. H. Chang and W. S. Lee. A Sliding Window method for Finding Recently Frequent Itemsets over Online Data Streams. In Journal of Information Science and Engineering, Vol. 20, No. 4, July, 2004.

[10] Tanbeer, S. K., Ahmed, C. F., Jeong, B.-S., and Lee, Y.-K. 2008. P-tree: a tree structure for single-pass frequent pattern mining. In Proc.ofPAKDD,LectNores tes Artif Int, 1022-1027.

[11] Koh, J.-L., and Shieh, S.-F. 2004. An efficient approach for maintaining association rules based on adjusting FP-tree structures. In Lee Y-J, Li J, Whang K-Y, Lee D (eds) Proc.