ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

A Comparative Study of Frequent Pattern Recognization Techniques from Stream Data

F.M.Christian[1], N.C.Chauhan[2], N.B.Prajapati[3]

PG Student[Computer], Dept. Of Computer, B.V.M Engineering College, Anand, Gujarat, India [1]

Assistant Professor, Dept. of IT, A.D.I.T Engineering College, Anand, Gujarat, India [2]

Assistant Professor, Dept of IT,BVM Engineering College, Anand, Gujarat, India[3]

ABSTRACT: Mining frequent pattern from data stream is a challenging task. Finding frequent pattern from data streams have been of importance in many application such as stock market prediction, sensor data analysis, network traffic analysis, e-business and telecommunication data analysis. Frequent Pattern Stream tree [1] is used for maintaining frequent pattern over a period of time using modified FP tree algorithm. This approach maintains tilted time window at each node which consumes larger space. Compact Pattern Stream Tree [2] assumes that only current patterns are of importance and uses sliding window protocol for maintaining it. This approach does not give importance to past frequent patterns.

Due to advancements in communication and storage technologies, large number of data streams has been generated by various applications and devices. Researchers have developed various methods to extract useful patterns from data streams. Many of the algorithms have been developed by extending the techniques that mines transaction data. Each methods work with different conditions such as offline streams, online streams, video streams, audio streams, etc. The performance and efficiency of the methods vary according to type of data streams. In this paper few recent and popular methods that extract patterns from stream data have been studied. Also a comparative analysis of different methods with reference to the conditions in which they work, and advantages/drawbacks of these methods are presented in this work.

Keywords: Data stream, frequent pattern stream tree, compact pattern stream tree, dynamic stream tree, tilted time window, sliding window

I.INTRODUCTION

Frequent pattern has been studied widely over a period of time but extending it to data stream is a challenging task. Stream data items are continuously flowing through the internet or sensor networks in applications like network monitoring and message dissemination [5]. Frequent pattern mining and its associated methods are widely used in association rule mining tasks, sequential pattern mining, sequential pattern mining, structured pattern mining, associative classification and so on[1].

Recently, the increasing prominence of data streams has led to the study of online mining of frequent Itemsets, which is an important technique that is essential to a wide range of emerging applications, such as web log and click-stream mining, network traffic analysis, trend analysis and fraud detection in telecommunications data, e-business and stock market analysis, and sensor networks. With the rapid emergence of these new application domains, it has become increasingly difficult to conduct advanced analysis and data mining over fast-arriving and large data streams in order to capture interesting trends, patterns and exceptions [5]. Mining frequent pattern from continuous data stream is more difficult than finding frequent pattern from static database. The basic properties of data stream are that they are continuous as they arrive at fast rate. Secondly, data can only be read only once because of continuity. Thirdly the total

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

amount of data that are arriving is unbounded. Mining streams require fast processing so that fast data arrival rate can be maintained and results of mining can be achieved in shorter response time.

The mining of continuously arriving data can be done through time base sliding window approach. There are three important windows models can be used based on feature of the application landmark window model, damped window model or sliding window model. Landmark window performs mining from data from a particular point called landmark to the current time. In the damped window model arriving data is assigned various weights based on their arrival, newly arriving data are assigned higher weights than older data. In sliding window fixed window is maintained and mining is applied within this window. igarettes [7].

II LITERATURE SURVEY

Various methods to find frequent pattern identification from data streams are described and discussed here.

2.1 Frequent pattern stream tree model

In this paper [1] a method for finding frequent pattern using FP stream tree is proposed. Compared to mining from a static transaction data set, the streaming case has far more information to track and far greater complexity to manage. Infrequent items can become frequent later on and hence cannot be ignored. The storage structure needs to be dynamically adjusted to reflect the evolution of Itemsets frequencies over time. In this paper, computing and maintaining all the frequent patterns (which is usually more stable and smaller than the streaming data) and dynamically updating them with the incoming data streams. They extended the framework to mine time-sensitive patterns with approximate support guarantee. They incrementally maintain -time windows for each pattern at multiple time granularities. Interesting queries can be constructed and answered under this framework. Moreover, inspired by the fact that the FP-tree provides an effective data structure for frequent pattern mining, we develop FP-stream, an effective FP-tree-based model for mining frequent patterns from data streams. An FP-stream structure consists of (a) an in-memory frequent pattern-tree to capture the frequent and sub-frequent Itemsets information, and (b) a tilted-time window table for each frequent pattern. Efficient algorithms for constructing, maintaining and updating an FP-stream structure over data streams are explored.

2.2 Compact pattern stream model

In this paper [2] a method for finding pattern from data stream using Compact pattern stream tree. An efficient technique to discover the complete set of recent frequent patterns from a high-speed data stream over a sliding window has been proposed. A Compact Pattern Stream tree (CPS-tree) to capture the recent stream data content and efficiently remove the obsolete, old stream data has been developed to provide compaction. This paper also introduces the concept of dynamic tree restructuring in our CPS-tree to produce a highly compact frequency-descending tree structure at runtime. This causes the high frequency pattern to be maintained at the higher level and low frequency components are mainly maintained at the leaf nodes. The complete set of recent frequent patterns is obtained from the CPS-tree of the current window using an FP-growth mining technique.

2.3 Classification and clustering model

In this paper [3] a method for stream classification is proposed. A new ensemble model which combines both classifiers and clusters together for mining data streams is proposed. The main challenges of this new ensemble model include (1) clusters formulated from data streams only carry cluster IDs, with no genuine class label information, and (2) concept drifting underlying data streams makes it even harder to combine clusters and classifiers into one ensemble framework. To handle challenge (1) a label propagation method to infer each cluster's class label by making full use of both class label information from classifiers, and internal structure information from clusters is proposed. To handle challenge (2), a new weighting schema to weight all base models according to their consistencies with the up-to-date base model. As a result, all classifiers and clusters can be combined together, through a weighted average mechanism, for prediction. Due to its inherent merits in handling drifting concepts and large data volumes, ensemble learning has traditionally attracted many attentions in stream data mining research. Nevertheless, as labeling training samples is a labor intensive and expensive process, in a practical stream data mining scenario, it is often the case that we may have a very few labeled training samples (to build classifiers), but a large number of unlabeled samples are available to build clusters. It would be a waste to only consider classifiers in an ensemble model, like most existing solutions do. Accordingly, in this paper, a new ensemble learning method which relaxes the original ensemble models to accommodate both classifiers and clusters through a weighted average mechanism. In order to handle concept drifting problem, we also propose a new

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

consistency-based weighting schema to weight all base models, according to their consistencies with respect to the up-to-date model.

2.4 Sliding window model

In this paper [4] they developed a novel approach for mining frequent Itemsets from data streams based on a time-sensitive sliding window model. The approach consists of a storage structure that captures all possible frequent Itemsets and a table providing approximate counts of the expired data items, whose size can be adjusted by the available storage space. . In addition, the approach guarantees no false alarm or no false dismissal to the results yielded. Self adjusting discounting table (SDT) for the limited memory is proposed. The SDT performs well when the available memory is limited.

2.5 Posterior probability based classification model

In this paper [5] a method is proposed method for classification of stream data In this paper a new approach to mine data streams by estimating reliable posterior probabilities using an ensemble of models to match the distribution over under-samples of negatives and repeated samples of positives has been proposed. They formally show some interesting and important properties of the proposed framework, e.g., reliability of estimated probabilities on skewed positive class, accuracy of estimated probabilities, efficiency and scalability.

2.6 Frequent pattern identification model based on dynamic load balancing

In this paper [6] a method is proposed for frequent pattern discovery from data streams In this paper, the practical problem of frequent-Itemsets discovery in data-stream environments which may suffer from data overload. The main issues include frequent-pattern mining and data-overload handling. Therefore, a mining algorithm together with two dedicated overload-handling mechanisms is proposed. The algorithm extracts basic information from streaming data and keeps the information in its data structure. The mining task is accomplished when requested by calculating the approximate counts of Itemsets and then returning the frequent ones. When there exists data overload, one of the two mechanisms is executed to settle the overload by either improving system throughput or shedding data load.

2.7 Dynamic stream tree model

Dynamic Stream Tree [11], which is a prefix-tree that is built based on a canonical item order, has recently been proposed for mining an exact set of recent frequent patterns over a data stream using a sliding window technique. In DSTree, the sliding window consists of several batches of transactions, and the transaction information for each batch is explicitly maintained at each node in the tree structure. To discover the complete set of recent frequent patterns, the FP-growth mining technique is applied to the DSTree of the current window. Even though the construction of the DSTree requires only one scan of the data stream, it cannot guarantee that a high level of prefix sharing (like in FP-tree) will be achieved in the tree structure because the items are inserted into the tree in a frequency-independent canonical order. Moreover, upon sliding of window the tree update mechanism of DSTree may leave some ‘garbage’ nodes in the tree structure when the mining request is delayed.

III COMPARATIVE ANALYSIS OF STREAM MINING ALGORITHMS

The frequent pattern identification model thus far considered in this review paper are frequent pattern stream tree model, compact pattern stream tree model, sliding window model, ensemble based classification and clustering model and posterior probability based classification model. As compared to frequent pattern stream tree the space requirement of compact pattern stream tree is less but compact pattern does not maintain the frequency of the pattern of the past. The structure of sliding window model based on self adjusting discounting table is compact as compared to the frequent pattern stream tree. The time required to find frequent pattern in sliding window model is less as compared to the frequent pattern stream tree. The space requirement of the compact pattern stream tree is less than sliding window model. The frequent pattern is found quickly in compact pattern tree as compared to sliding window model. The space requirement of dynamic stream tree is less than FP-stream tree. The DST-tree finds the frequent pattern at the given duration of time without considering the frequency of the pattern in the past. The frequent pattern is found more quickly by dynamic stream tree as compared to frequent pattern stream tree. The time and space requirement for finding frequent pattern is less in compact pattern stream tree as compared to Dynamic stream tree. A brief summary of important data steam mining algorithms along with advantages and drawbacks is given in Table 1.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

TABLE 1
COMPARISON OF DIFFERENT FREQUENT PATTERN RECOGNIZATION FROM DATA STREAM TECHNIQUE

Technique used	Special feature	Merits	Demerits
Frequent Pattern Stream Tree(FP-Stream tree) [1]	It maintains pattern tree and tilted time window.	It maintains history of the frequency of pattern over a period of time and provides the information of the frequent pattern of the pattern in the current window	The data structure is complex because tilted time window is to be maintained at each node
Compact Pattern Stream Tree(CP-stream tree) [2]	It constructs Frequent Pattern tree for the current window and also maintains the I-node table for pattern ordering.	It provides the frequent pattern in the current window. The data structure is not complex.	It does not keep history of the frequency of pattern. Therefore it only gives information of frequent pattern in current window.
Ensemble based classifiers and clusters[3]	The label propagation method is used.	Label propagation method is used for classification. It provides efficient classification capability.	Transferring the class label from one cluster to another cluster is tedious work when huge amount of data is present.
Time sensitive sliding window model[4].	In this the posterior probability method is used.	The approach consists of a storage structure that captures all possible frequent Itemsets and a table giving approximate counts of the expired data items, whose size can be adjusted by the available storage space	The approach does not maintain history of frequent items and only maintains information about the current frequent items.
Dynamic stream tree model[11]	The FP growth method is applied to current window. Canonical ordering is maintained.	This approach finds the frequent pattern from stream using compact data structure. The time requirement for finding frequent pattern in dynamic stream tree is less.	This does not maintain history of frequent pattern. Less compact than Compact pattern stream tree.

.

IV CONCLUSION

A comparative study of different techniques for frequent pattern identification from data streams have been discussed in this work. Frequent pattern identification from data streams are mainly used in stock market analysis, sensor network analysis, etc. The frequent pattern stream tree finds frequent pattern over a period of time. This system also gives the information frequency of the item in each window. The system has to maintain the tilted time window at each node which makes the structure of tree complex. The compact stream tree maintains the information of the pane at the tail node but gives information of the frequent pattern in current window only and does not maintain past history. The

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 1, January 2014

Frequent pattern stream is best for finding frequent pattern when there is requirement to maintain history of information. The compact pattern stream tree is best when current pattern is of importance. The dynamic stream tree is similar to compact pattern stream tree but it requires canonical ordering of the pattern to be maintained. The DST tree finds frequent pattern in current window only.

REFERENCES

[1] C. Giannella; J. Han, J. Pei; X. Yan; P.S. Yu, Mining frequent patterns in data streams at multiple time granularities, in: Data Mining: Next Generation Challenges and Future Directions, AAAI/MIT Press, 2004 (Chapter 6).
[2] S.K Tanbeer; C.F. Ahmed; B. Jeong, Y. Lee; "Sliding window-based frequent pattern mining over data streams", Information Sciences Elsevier, Vol. 179, Issue 22, pp. 3843 - 3865, 2009.
[3] P. Zhang; X. Zhu; J. Tan; L. Guo, "Classifier and Cluster Ensembles for Mining Concept Drifting Data Streams", in: proc ICDM, 2010, pp. 1175-1180.
[4] C. Lin; D.Y Chiu; Y.H. Wu; A. Chen, "Mining Frequent Itemsets from Data Streams with a Time-Sensitive Sliding Window", in: Proc. Society of industrial and applied mathematics(SIAM) ICDM ,2005.
[5] J. Gao; W. Fan; J. Han; P. S. Yu, "A General Framework for Mining Concept-Drifting Data Streams with Skewed Distributions", Society for industrial and applied mathematics(SIAM) International Conference on Data Mining, Minneapolis, April 2007.
[6] C. Li; K.F. Jea; R.P. Lin; S.F. Yen; C.W. Hsu, "Mining frequent patterns from dynamic data streams with data load Management" , Journal of Systems and Software ,vol 85,Issue 6, pp 1346-1362, 2012.
[7] M. kholghi; M. keyvanpour,"An analytical framework for Data stream mining techniques Based on challenges and Requirements", International journal of engineering science and technology (IJEST), vol. 3, no. 3, pp. 2507-2513,march 2011 .
[8] G.K Gupta, "Introduction to Data Mining with case studies", PHI publication, 2nd edition, 2006.
[9] J. Han; M. kamber; J. Pei, "Data mining concepts and techniques" -The Morgan Kaufmann series in Data Management Systems, 3rd edition, 2012.
[10] V. pudi; P. R. Krishna, "Data Mining", Oxford University Press, 1st edition, 2009.

[11] C.K.-S. Leung; Q.I. Khan, "DSTree: a tree structure for the mining of frequent sets from data streams", in: Proc. ICDM, 2006, pp. 928–932.[12] F.-Y. Ye, J.-D. Wang, B.-L. Shao, New algorithm for mining frequent itemsets in sparse database, in: Proc. the Fourth International Conference on Machine Learning and Cybernetics, 2005, pp. 1554–1558.
[13] C.K.-S Leung; Q.I. Khan; " Efficient mining of constrained frequent patterns from streams", in: Proc. 10th International Database Engineering and Applications Symposium, 2006.
[14] H.-F. Li; S.-Y. Lee; M.-K. Shan, "An efficient algorithm for mining frequent itemsets over the entire history of data streams", in: Proc. International Workshop on Knowledge Discovery in Data Streams, 2004.[15]X. Zhi-Jun, C. Hong, C. Li, "An efficient algorithm for frequent itemset mining on data streams", in: Proc. ICDM, 2006, pp. 474–491.
[16] J.X. Yu, Z. Chong, H. Lu, A. Zhou," False positive or false negative: mining frequent itemsets from high speed transactional data streams", in: Proc. VLDB, 2004, pp. 204–215.
[17] G.S. Manku, R. Motwani, "Approximate frequency counts over data streams", in: Proc. VLDB, 2002, pp. 346–357.
[18] J.H. Chang and W.S. Lee, "Finding Recent Frequent Itemsets Adaptively over Online Data Streams," in:Proc. of ACM SIGKDD Conf.,2003, pp. 487-492.
[19] S.K. Tanbeer, C.F. Ahmed, B.-S. Jeong, Y.-K. Lee, CP-tree: a tree structure for single-pass frequent pattern mining, in: T. Washio et al. (Eds.), Proc. PAKDD, 2008, pp. 1022–1027.
[20]P. Zhang, X. Zhu, Y. Shi, L. Guo, and X. Wu, "Robust Ensemble Learning for Mining Noisy Data Streams",,*Decision Support Systems*, Vol. 50, Issue 2, pp: 469-479,2011.