International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-7 December-2012

Towards a New Approach for Mining Frequent Itemsets on Data Stream

Shailendra Jain[1], Sonal Patil[2]
[1]Assistant Professor, TIT, Bhopal
[2]M.Tech Second Year (Software Systems), TIT, Bhopal

Abstract

From the advent of association rule mining, it has become one of the most researched areas of data exploration schemes. In recent years, implementing association rule mining methods in extracting rules from a continuous flow of voluminous data, known as Data Stream has generated immense interest due to its emerging applications such as network-traffic analysis, sensor-network data analysis. For such typical kinds of application domains, the facility to process such enormous amount of stream data in a single pass is critical. Nowadays, many organizations generate and utilize vast data streams (Huang, 2002). Employing data mining schemes on such massive data streams can unearth real-time trends and patterns which can be utilized for dynamic and timely decisions. Mining in such a high speed, enormous data streams significantly differs from traditional data mining in several ways. Firstly, the response time of the mining algorithm should be as small as possible due to the online nature of the data and limited resources dedicated to mining activities (Charikar, 2004). Second, the underlying data is highly volatile and subject to change over period of time (Chang, 2003). Moreover, since there is no time for preprocessing the data in order to remove noise, the streamed data can have noise inherent in it. Due to all aforementioned problems, data stream mining is receiving increasing attention and current research is now focused on the efficient resolution to the problem cited above. Although, the field of data stream mining is being heavily investigated, there is still a lack of a holistic and generic approach for mining association rules from data streams. Thus, this research attempts to fill this gap by integrating ideas from previous work in data stream mining. This investigation focuses on the degree of effectiveness of using a probabilistic approach of sampling in the data stream together with an incremental approach to maintenance of frequent itemsets in a data stream environment. The following paper describes the design and experimentation conducted with a novel association rule mining algorithm that can be deployed on a high speed data stream.

Keywords

Data stream, Association Rule mining, Frequent Patten data stream.

1. Introduction

Association rules, as its name suggests, expresses relationships between items in (generally large) datasets. This powerful data mining technique has a wide range of applications including Market-Basket analysis, text mining, web mining and pattern recognition between genes in a dataset. Such rules enable users to uncover hidden relationships and patterns in large datasets. For example, customer's buying patterns or relationships between different genes in organisms. The seminal work of (Agrawal, 1994) and the proposed Apriori algorithm spurred a plethora of research in the area of association rule mining. In a wide range of emerging applications, data is in the form of an enormous, continuous stream where the speed at which the data is produced outstrips the rate at which it can be mined (Charikar, 2004). This is in direct contrast to traditional static databases; thus data stream mining therefore is substantially deviant from conventional data mining in numerous aspects. Firstly, the absolute volume of data embedded in a data stream over its lifespan can be overwhelmingly huge (Gaber, 2005). Secondly, due to resource bottlenecks, generating timely responses by keeping response time to queries on such data streams is necessary (Jiang, 2006).Because of the issues stated above, data stream mining has become the subject of intense research and the problem of obtaining timely and accurate association rules is a contemporary research topic. There is a critical need to switch from traditional data mining schemes to those methods that are able to operate on an open-ended, high speed stream of data (Manku, 2002).Due to the inherent nature of a data stream, any mining scheme faces the following challenges (Gaber, 2005). Firstly, due to the continuous nature of stream data, the traditional approach of scanning the database multiple times for model creation is no longer feasible A Hybrid Association Rule Mining Algorithm (Zahir Tari and Wensheng2006). Most of the approaches for association rule mining focus on

the performance of the discovery of the frequent itemsets. They are based on the algorithms that require the transformation of data from one representation to another, and therefore excessively use resources and incur heavy CPU overhead. This chapter proposes a hybrid algorithm that is resource efficient and provides better performance. It characterizes the trade-offs among data representation, computation, I/O and heuristics. The proposed algorithm uses array-based item storage for the candidate and frequent itemsets. In addition, we propose a comparison algorithm (CmpApr) that compares candidate itemsets with a transaction, a filtering algorithm (FilterApr) that reduces the number of comparison operations required to find frequent itemsets. The hybrid algorithm (ARM++) integrates filtering methods within the Partition algorithm.

1. Frequent pattern mining: current status and future Directions (Jiawei Han • Hong Cheng • Dong Xin •Xifeng Yan(2007)

Frequent pattern mining has been a focused theme in data mining research for over a decade. Abundant literature has been dedicated to this research and tremendous progress has been made, ranging from efficient and scalable algorithms for frequent itemset mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification and frequent pattern-based clustering, as well as their broad applications. In this article, we provide a brief overview of the current status of frequent pattern mining and discuss few promising research directions. We believe that frequent pattern mining research has substantially broadened the scope of data analysis and will have deep impact on data mining methodologies and applications the long run. However, there are still some challenging research issues that need to be solved before frequent pattern mining can claim a cornerstone approach in data mining applications.

2. Catch the moment: maintaining closed frequent itemsets over a data stream sliding window (Yun Chi, Haixun Wang, Philip li 2006)

This paper considers the problem of mining closed frequent itemsets over a data stream sliding window using limited memory space. We design a synopsis data structure to monitor transactions in the sliding window so that we can output the current closed frequent itemsets at any time. Due to time and memory constraints, the synopsis data structure cannot monitor all possible itemsets. However, monitoring only frequent itemsets will make it impossible to detect itemsets when they become frequent. In paper, we introduce a compact data structure, the closed enumeration tree (CET), to maintain a dynamically selected set of itemsets over a sliding window. The selected itemsets contain a boundary between closed frequent itemsets and the rest of the itemsets. Concept drifts in a data stream are reflected by boundary movements in the CET. In other words, a status change of any itemset (e.g., from non- frequent to frequent) must occur through the boundary. Because the boundary is relatively stable, the cost of mining closed frequent itemsets over a sliding window is in dramatically reduced to that of mining transactions that can possibly cause boundary movements in the CET. Our experiments show that our algorithm performs much better than representative algorithms for the sate-of-the-art approaches

3. A fast online mining frequent closed itemsets (Junbo Chen and ShanPing li 2008)

Frequent closed itemsets is a complete and condensed representaion for all the frequent itemsets, and it's important to generate non- redundant association rules. It has been studied extensively in data mining research, but most of them are done based on traditional transaction database environment and thus have performance issue under data stream environment. In this paper, a novel approach is proposed to mining closed frequent itemsets over data streams. It is an online algorithm which updates frequent new closed itemsets incrementally, and can this output the current closed frequent itemsets cin real time based on users specified thresholds. The experimental evaluation shows that our proposed method is both time and space efficient, compared with the state of art online frequent closed itemsets algorithm FCI-Stream [3].

2. Itemset Pruning

The necessary step for computing itemsets from the stream of items in a transaction is the computation of a powerset that would take into account all the possible combinations of itemsets of size 1 (that is, individual items appearing in the transaction without combination with other items) seen in a transaction. The problem such an approach is that in a data stream where the incoming transaction can very well hold hundreds of items, the CPU time would be prohibitive as the time complexity for the powerset computation is O(2n) where n is th items in the transaction. As a solution to this problem, our

approach undertakes 1 itemset pruning. In this approach, the data stream is segmented into a number of equal-sized blocks which we call frames. The size of a frame is determined by the Chernoff and is given by,

$N_0 = 2+2\ln(2/\delta)/8$

(For derivation of above formula from generic Chernoff bounds formula, please refer (Yu, 2004)). As per above formula, when, the minimum support threshold's'=0.001 and delta 'σ '= 0.1, the memory bound n0 = 7991, that is every frame is marked by 7991 transactions.

'δ'= 0.1, the memory bound n0 = 7991, that is every frame is marked by 7991 transactions.

Closed Itemset Mining

Frequent itemset mining has a crucial role to play in association rule mining (Agrawal, 1 9 9 4). Nevertheless, the task of mining association rules can result in a huge number of candidate itemsets that need to be assessed against their frequency status, thus impacting on efficiency. The classical Apriori (Agrawal, 1994) approach generates candidate k-itemsets for every pair of (k-1) frequent itemsets. When the minimum support threshold is et a low enough value, this can result in a combinational explosion the number of candidates to be processed (Chi, 2004). Closed Frequent itemset mining presents itself as an attractive alternative to classical frequent itemset mining, particularly in a data stream environment. A closed itemset I is any itemset that does not contain a superset with the same frequency as itself. A closed frequent itemset is one that has support above the minimum support threshold.

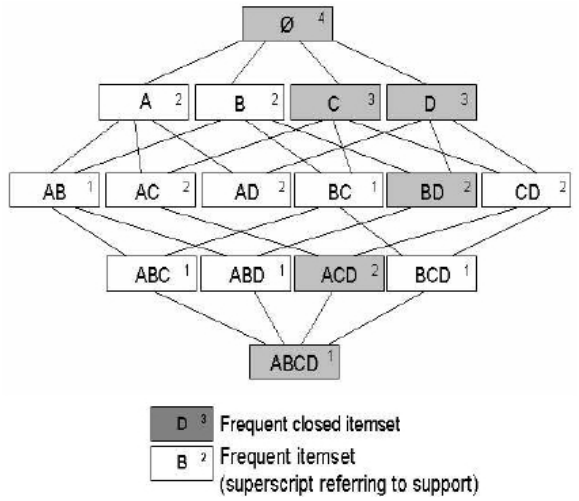

Figure 1: Closed Itemset Mining

3. Proposed Work

This section will focus on describing the describing the experiments designed to evaluate the performance of our proposed Data Stream Mining (DSM) algorithm. DSM is compared against a contemporary frequent itemset mining algorithm called the Frequent Pattern Data Miner or FPDM2 (Yu, 2004).Although DSM was designed to mine closed frequent itemsets efficiently, we have extended its capabilities to mine all frequent itemsets by incorporating a routine that determines the frequent itemsets from the closed frequent itemsets (as explained previously in Chapter 3).The following sections will explain the experimental design along with metrics that we used to compare the performance of DSM against that of FPDM2.

Datasets

The experimentation is carried out with the help of synthetic datasets that are generated through the use of a dataset generator that is publically available (Agrawal, 1994). Acquiring a real life dataset is quite difficult due to the fact that many organizations refuse to part with their data because of the sensitivity and confidentiality of the data. Therefore, artificial synthetic data generators such as IBM are very commonly used by researchers to evaluate and benchmark their algorithms' Performances.

Performance Metrics

The DSM and FPDM2 algorithms are evaluated against certain commonly used performance metrics such as Accuracy (in terms of Recall and Precision), Computational performance (in terms of time taken to process the dataset), and Memory consumption in terms of number of nodes maintained. Recall and precision can be defined.

Recall = (Frequent Itemsets in data)Π(Retrieve Frequent Itemsets) / Frequent Itemsets in data

Precision = (Frequents Itemsets in data) Π (Retrieve Frequent Itemsets)/ Retrieve Frequent Itemsets

Where "Frequent Itemsets in data" corresponds to all the frequent itemsets that are actually present in the dataset (identified using Apriori implementation), and "Retrieved frequent itemsets" are the itemsets reported by algorithms (DSM and FPDM2) as frequent itemsets. Support and Reliability. Minimum support is the threshold that determines whether an itemset is of any interest to the end user. The reliability, on the other hand is the probability that the estimated support of an itemset as measured over the frame structure imposed by the Chernoff bound .This experiment was mainly designed to compare

DSM and FPDM2 with respect to the previously explained performance metrics. For this experiment, we have used the T10I4 and T15I6 datasets which are relatively dense when compared to T5I4 used in the previous experiment. The following steps were executed in this experiment:
a) For the T10I4 dataset, vary the minimum support parameter while keeping delta constant, b) For theT10I4 dataset, vary the delta parameter while keeping minimum support constant,
c) Repeat steps a and b for the T15I4 dataset
d)Measure theaccuracy,computational performance and memory consumption for each of the steps a), b) and c) above. Data stream mining is one of the most intensely investigated and challenging research domains in contemporary research in the data mining discipline as a whole. The peculiarities of data streams render conventional mining schemes inappropriate.

4. Conclusion

We presented an overview of a novel approach for mining the closed itemsets from a data stream. We have implemented an efficient closed prefix tree to store the intermediate support information of frequent itemsets. Moreover; we have employed the Apriori principle to reduce unnecessary power set creation along with transaction pruning to further enhance transaction processing. We developed an incremental closed itemset mining algorithm based on the probabilistic guarantees of Chernoff bounds. The Chernoff bound helps in purging unnecessary itemsets from the data stream and keeps the memory requirements within reasonable bounds. We compared our approach with the FPDM2 algorithm proposed by Yu (2004).Although FPDM2 is a frequent itemset mining algorithm and DSM a closed frequent itemset mining algorithm, a basic routine computing frequent itemsets from closed frequent itemsets enabled us to assess the performance of both the approaches on the uniform grounds.

References

[1] Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules.Paper presented at the 20th International Conference on Very Large Data Bases (VLDB'94), Santiago, Chile.
[2] Ao, F., Yan, Y., Huang, J., Huang, K. (2007). streams based on FP trees. Springer Verlag Berlin Heidelberg, 479-489.
[3] Ben-David, S., Gehrke, J., Kifer, D. (2004). Detecting change in data streams Paper presented at the 30th VLDB Conference, Toronto, Canada.
[4] Burrel, G., Morgan, G (1979). Sociological paradigms and organizational analysis. London: Heinemann.
[5] Celgar, A., Roddick, J. (2006). Association mining. ACM Computing Surveys, 38(2), 1-42.
[6] Chang, J., Lee, W. (2003). Finding recent frequent itemsets adaptively over online data streams. Paper presented at the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA.
[7] Charikar, M., Chen, K., Colton, M. (2004). Finding frequent items in data streams. Theoretical Computer Science, 1-11.
[8] Cheung, D., Han, J., Vincent, T., Wong, C. (1996).Maintenance of discoveredassociation rules in large database: An incremental updating technique.Paperpresented at the IEEE International Conference on Data Mining, New York, USA.
[9] Chi, Y., Wang, H., Yu, P., Muntz, R. (2004). Moment: Maintaining closed frequent itemsets over a stream sliding window. Paper presented at the IEEE International Conference on Data Mining maximal frequent itemsets in data.
[10] Collis, J., Hussey, R. (2003). Business Research.Basingstoke, UK: Palgrave Macmillan.
[11] Cormode, G., Garofalakis, M. (2007). Sketching probabilistic data streams. Paper presented at the SIGMOD'07.
[12] Dash, N. (2005). Selection of the Research Paradigm and Methodology. Online Research Methods Resource. Retrieved from presented http://www.celt.mmu.ac.uk/researchmethods/Modules/Selection_of_methodology/index.php.
[13] Gaber, M., Zaslavsky, A., Krishnaswamy, S.(2005). Mining data streams: A review. ACM SIGMOD Record, 34(2), 18-26.
[14] Giannella, C., Han, J., Pei, J., Yan, X., Yu, P. (2003). Mining frequent patterns in data stream at multiple time granularities. In Next Generation Data Mining (pp. 105-124).
[15] Gouda, K., Zaki, M. (2001). Efficiently mining maximal frequent itemsets. Paper presented at the 2001 IEEE International Conference on Data Mining. Huang, H., Wu, X., Relue, R. (2002). Association analysis with one scan of databases. Paper presented at the IEEE International Conference on Data Mining, Maebashi City, Japan.
[16] IBM. IBM data generator,from http://www.almaden.ibm.com/cs/projects/iis/hdb/Projects/data_mining/datasets/syndata.html.
[17] Jiang, N., Gruenwald, L. (2006). CFI-stream: Mining closed frequent itemsets indata streams. Paper presented at the KDD, Philadelphia, USA.
[18] Jiang, N., Gruenwald, L. (2006). Research issues in data stream association rule mining. ACM SIGMOD Record, 35(1), 14-19.

[19] Laur, P., Nock, R., Symphor, J., Poncelet, P. (2005). On the estimation of frequent itemsets for data streams: Theory and experiments. Paper presented at the CIKM.
[20] LeeS.,Cheung,D.,Shan,M.(1997). Maintenance of discovered association rules: When to update. Research Issues on Data Mining and Knowledge Discovery, 1-8.
[21] Li, L. (2003). A graph-based algorithm for frequent closed itemset mining. Paper presented at the 2003 IEEE Systems and Information Engineering Design Symposium, Charlottesville, VA.

Prof. Shailendra Jain was born at Lalitpur (UP) on 30/08/1978. M.Tech. (Computer Technology and Application) completed in year 2006 from School of IT, UTD, Rajiv Gandhi Technical University, Bhopal (MP).Present Designation: Assistant Professor, Department of Computer Science And Engineering, Technocrats Institute of Technology, Bhopal (MP).Working since January, 2007.

Ms Sonal Patil was born at Bhusaval on 6/6/1986. She has completed her B.E. with distinction and currently studying in Mtech Second Year in Software Systems at TIT, Bhopal.