Fast Rule-Based Prediction of Data Streams using Associative Classification Mining

K.Prasanna Lakshmi
Department of Information Technology
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India
prasannakompalli@gmail.com, prasanna@griet.ac.in

Dr. C.R.K.Reddy
Department of Computer Science
Chaitanya Bharathi Institute of Technology
Hyderabad, India
crkreddy@gmail.com

Abstract— **Associative Classification is a recent and rewarding approach which combines associative rule mining and classification. This technique has attracted many researchers as it derives accurate classifier with effective rules. Associative classifiers are useful for application where maximum predictive accuracy is desired. Increasing access to huge datasets and corresponding demands to analyze these data has led to the development of new online algorithms for performing machine learning on data streams. There is a need to develop a decision support system for predicting the huge datasets generated by various applications in IT industry. In this paper we proposed two efficient techniques called PSTMiner and PSToSWMine for prediction of data streams. This research uses associative classification which builds a classifier with prediction rules of high interestingness values. Experimental results show that the performance of these two algorithms is highly competitive in terms of accuracy and performance time.**

Keywords— Data Streams, Associative classification, Prediction

I. INTRODUCTION

Data Stream Mining has attracted information industry as well as healthcare organizations in recent years as it turns the huge amounts of data into useful knowledge. Knowledge gained by mining data streams can also be used in applications related to business, production control and market analysis etc., [2]. It is an arena of knowledge discovery in the field of data mining. Mining streaming data poses many new challenges [2]. In mining data streams it is unrealistic to store the entire data in main memory or in secondary storage devices for discovering the knowledge. This huge data can be used for answering questions like "predict the probability of a transaction to be fraud". Data Stream Mining has the potential to generate knowledge which helps in improving the quality of decision making in IT industry. According to the stream data processing, the research of mining data streams can be done using landmark window or sliding window [11].

Unlike classification from static databases, classification from data streams poses challenges like producing results in real time, processing data in single pass by using limited memory. There is growing evidence that associative classification produces efficient and accurate results when compared with classification technique [3], [4], [5]. The objective of associative classification is to build a model called classifier which maps objects into predefined classes based on rules present in classifier.

Associative classification involves two phases.

1) Generation of classifier using rules generated from training dataset. Rules generated are of the form
   X→ C where C is class label.
2) Prediction of class label of test data using classifier.

The main objective of this work is to develop an archetype for prediction using both landmark window and sliding window. In this paper we propose two models which extracts hidden knowledge associated with flowing data streams.
The rest of paper is organized as follows. In section 2 we introduce data streams along with windowing concept. In section 3 we describe the associative classification technique. Our methodology PSTMiner using Landmark Window is in section 4. Section 5 presents PSToSWMine using Sliding Window. Experimental results are shown in section 6 and finally we conclude in section 7.

II. DATA STREAM MINING OVERVIEW

In recent years, data streams have become ubiquitous due to presence of large number of applications which generate huge volumes of data in an automated way [6], [2]. This streaming data has following computational characteristics.

- The data arrives continuously
- No assumptions can be made on the order of arriving stream
- Memory utilized during mining should be limited.

978-1-4673-6537-6/15/$31.00 ©2015 IEEE

- Knowledge must be gained as quickly as possible as each data element in the stream must be examined only once.

Therefore, traditional machine learning algorithms are not directly applicable to data streams. Efficiently capturing knowledge from streams has become very crucial. Existing algorithms used in data mining can be modified for handling data streams. Data stream mining combines statistical analysis and machine learning techniques of data mining for extracting hidden patterns and for finding relationships in the arriving streams. Data stream mining works using either supervised learning or unsupervised learning. Main objectives of stream mining are classification and prediction.

Different data models are been proposed due to the nature of data streams [7], [8], [9]. The first data model is called landmark window model. In this model, all data streams from the start time till current time are considered for mining. As a stream arrives it is appended continuously as time grows. PSTMiner uses this model for building classifier for prediction. 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. Working with sliding window is depicted in PSToSWMine algorithm.

III. ASSOCIATIVE CLASSIFICATION

Many researchers are focusing on designing classification algorithm to build efficient classifiers for large data sets [10], [4], [3], [11]. Recent studies in classification have exploited the use of association rule mining for generating classifiers for classification [11]. This new approach is called as associative classification. Classifying a data stream with associative classifier is a new area of research. Associative classification over data streams achieves accurate results as it uses frequent item sets mining which captures dominant relationships between data items in a data stream. The classifiers formed by rules generated from associative classification mining contain only statistically prominent associations, so this type of mining is robust in nature. In associative classification mining the class attribute is considered as consequent of rules found in classifier.

Definition 1. Let $CL=\{R_1 \to C_1, R_2 \to C_2, ...R_n \to C_n\}$ be a classifier generated by associative classification mining then in the rule $R_1 \to C_1$ C_1 is called class label and R_1 is set of frequent item sets.

IV. PSTMiner METHODOLOGY

In the proposed PSTMiner, we used associative classification mining over landmark window of data streams. Our work is organized in two phases.

1. Generating rules from associative classification mining.
2. Pruning the rules using chi-square testing and arranging the rules in an order to form a classifier.

A novel data structure called Prefix Streaming Tree (PSTree) [11] was used for holding the data stream. This compact structure was built with a single pass of data stream. Working with PSTree is clearly explained in [11]. For flexibility of work we considered the arriving stream of data as batches. These batches of streams are captured using landmark window which grows as time passes. Frequent item set mining is done over stream stored in PSTree. All the frequent items satisfying the minimum support are found. Rules of type $R \to C$ satisfying minimum confidence are extracted. All redundant rules are pruned using chi-square statistical measure. The set of rules so found are placed in a classifier by arranging them in an order based on their confidence and support. Methodology used by PSTMiner is shown in Fig.1.

The classifier so build will contain all dominant rules produced by incremental learning over landmark window. When a stream with unlabeled class arrives, prediction of class label happens with the build classifier.

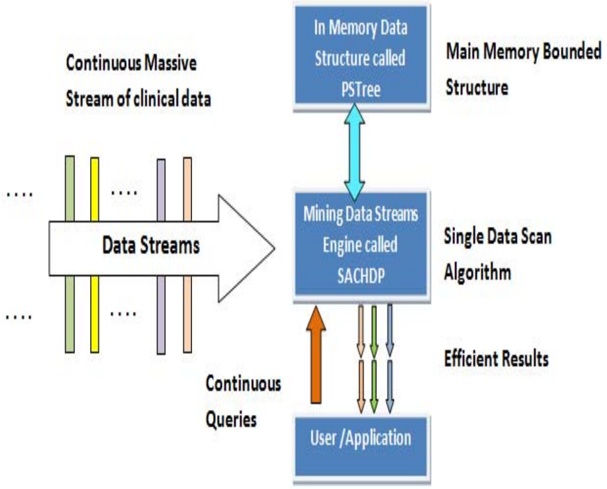


Fig. 1.Methodology of PSTMiner

V. PSToSWMine Framework

PSToSWMine is a learning classification model based on frequent item set mining. The framework consists of three stages

1. Compact stream representation using PSTree [11].
2. Mining frequent item sets and frequent rules.
3. Model a learning classifier using rules found in second stage.

The methodology is shown in Fig.2.

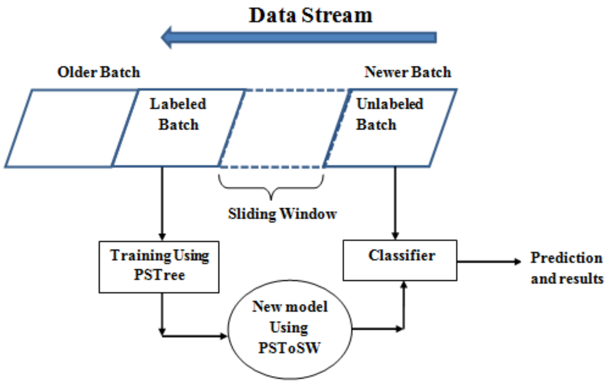


Fig. 2. PSToMine using sliding window

PSTree, the structure used to store data, is based on prefix tree concept [16]. As mentioned earlier the working of PSTree is explained in [11]. A batch of the stream data from the sliding window is inserted into PSTree. After completion of insertions the tree is restructured for maintaining compactness. Once the entire data from the window is processed using PSTree the windows sliding for the next batch of instances present in the stream. This algorithm maintains recent information from the flowing stream. As recent data is not sufficient for prediction of a new tuple the algorithm was designed as an incremental algorithm. Minimum support and minimum confidence are the two inputs. Support is used as a measure of significance of the rule and confidence is used as measure of strength of rule.

The experimental results presented in next section show that our two models helps in better prediction of unseen data.

VI. EXPERIMENTAL EVALUATION

We have evaluated the accuracy of on real datasets and synthetic datasets. Real datasets were taken from UCI repository [12]. Synthetic datasets are collected using MOA framework. The important features of these datasets are listed in Table I. Experiments were conducted on dense and sparse datasets and evaluation of methodologies gave consistent results when compared with existing techniques. Statistical measures like precision, recall and F-measure were used for evaluating the performance of both the techniques. Our PSTMiner is compared with existing Associative classification techniques C4.5 [13], CBA [14], CMAR [4] and L^3[15] and PSToSWMine is compared with STREAMGEN [17]. All the experiments were conducted on 2.53 GHz Intel PC with 1.0 GB RAM, running on windows XP machine.

TABLE I. DATASETS CHARACTERISTICS

Dataset	Number of Transactions	Number of Attributes	Number of Classes	Number of Items
Australia	690	14	2	49
Breast-w	699	10	2	29
crx	690	15	2	53
Diabetics	768	8	2	15
digit	10992	16	10	165
heart	270	13	2	18
iris	150	4	3	12
mushroom	8124	22	2	116
nursery	12960	8	5	27
pima	768	8	2	15
tictactoe	958	9	2	27
vehicle	846	18	4	71
wine	178	13	3	37
zoo	101	16	7	34

Table II shows the comparison of accuracies obtained by C4.5, CBA, CMAR and L^3 over our PSTMine on real datasets. Fig.4. shows the comparison of average ErrorRates.

TABLE II. ACCURACY COMPARISON ON REAL DATSETS

Dataset	C4.5	CBA	CMAR	L^3	PSTMiner
Australia	84.7	84.9	86.1	**98.1**	92.8
Breast-w	95.0	95.3	96.4	96.4	**97.8**
Crx	84.9	85.9	84.9	86.7	**92.4**
Diabetics	74.2	72.9	74.5	**78.7**	69.7
Heart	80.8	81.9	82.2	84.4	**96.6**
Iris	95.3	92.9	94.0	93.3	**97.33**
Mushroom	--	97.7	--	**100**	**100**
Nursery	--	80.1	--	83.3	91.6
Pima	75.5	73.1	75.1	78.5	**80.31**
Tictactoe	99.4	100	99.2	**100**	99.1
Vehicle	72.6	68.8	68.8	**73.4**	66.1
Vote	--	93.5	--	95.2	**99.3**
Wine	92.7	91.6	95.0	98.3	**98.8**
Zoo	92.2	94.6	97.1	94.1	**100**
Average Accuracy	84.3	84.6	85.7	89.2	**91.2**

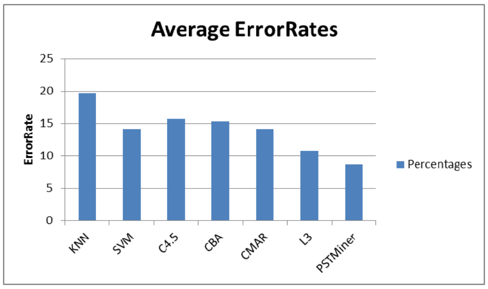
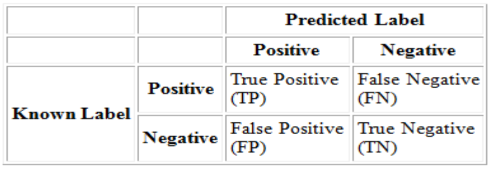


Fig. 4. Error Rate Comparison

We used precision, recall and F-measure as basic evaluation measures for checking the performance of our build classifiers.



		Predicted Label	
		Positive	Negative
Known Label	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Fig. 5. Scenerio

Fig.5. shows a scenario for which these are computed using the following equations:

$$\Pr ecision = \frac{TP}{TP+FP}, \quad \mathrm{Re}\, call = \frac{TP}{TP+FN} \quad \text{and}$$

$$FMeasure = \frac{2 * \Pr ecision * \mathrm{Re}\, call}{\Pr esision + \mathrm{Re}\, call}$$

Precision is a measure of exactness, recall is the measure of completeness and F-Measure is a compromise between recall and precision. Fig.6. depicts the graph between precision and minimum support over nursery dataset.

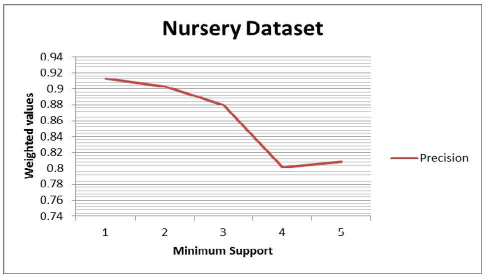


Fig. 6. Precision verses minimum support for nursery dataset

Table III and Table IV. shows the detailed report of execution of PSToSWMine algorithm.

TABLE III. PSTOSWMINE COMPUTATION DETAILS USING PRECISION, RECALL AND F-MEASURE OVER REAL DATASETS

Dataset	Accuracy	Classes	Precision	Recall	F-Measure
Australia	92.89	+	0.9387	0.8990	0.918
		-	0.9217	0.9530	0.937
Breast-w	97.85	Benign	0.9846	0.9825	0.983
		Malignant	0.9669	0.9709	0.968
crx	92.48	+	0.9322	0.8957	0.913
		-	0.9193	0.9530	0.935
Diabetics	69.14	Positive	0.8604	0.1380	0.237
		Negative	0.6813	0.988	0.806
heart	96.68	<50	0.9638	0.9756	0.969
		>50	0.9705	0.9565	0.963
iris	97.33	setosa	1	1	1
		versicolor	0.96	0.96	0.96
		virginica	0.96	0.96	0.96
mushroom	100	poisonous	1	0.8637	0.926
		edible	0.867	1	0.928
nursery	91.09	priority	0.8743	0.8518	0.862
		not_recom	1	1	1
		spec_recom	0.8590	0.9525	0.903
pima	80.31	42	0.8633	0.5186	0.647
		41	0.7871	0.9559	0.863
tictactoe	99.16	positive	0.9873	1	0.993
		negative	1	0.9759	0.987
vehicle	66.19	van	0.7070	0.9095	0.795
		saab	0.6808	0.4485	0.540
		bus	0.6386	0.885	0.741
		ope1	0.6818	0.4390	0.534

TABLE IV. PSTOSWMINE COMPUTATION DETAILS FOR SYNTHETIC DATASETS

Dataset	Number of Transactions	Accuracy	Classes	Precision	Recall	F-Measure
Stagger	1Lakh	100	true	1	1	1
			false	1	1	1
Stagger	5Lakh	100	true	1	1	1
			false	1	1	1
Random Tree Generator	1Lakh	69.19	class1	0.700	0.704	0.702
			class2	0.717	0.713	0.715
Agarwal Generator	10000	94.64	group B	0.965	0.865	0.913
			group A	0.938	0.985	0.961

Fig. 7 illustrates the comparison of F-Measure with accuracy for nursery dataset. Fig.8. depicts the comparisons of PSToSWMine with SREAMGEN and DDPMine.

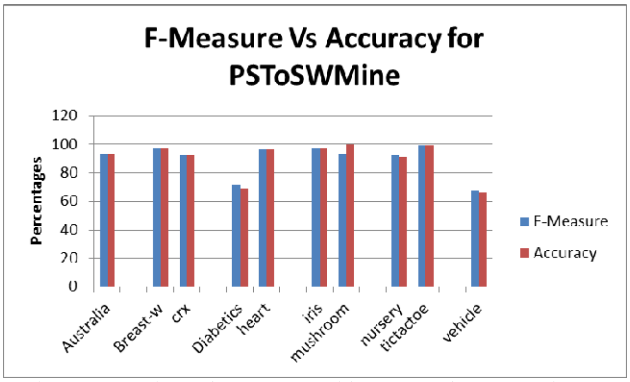


Fig. 7. Comparison of F-Measure with accuracy for nursery dataset

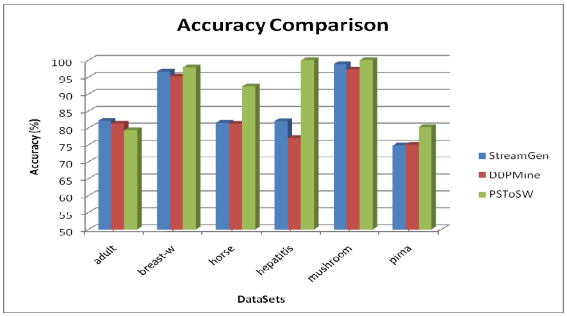


Fig. 8. Comparison of PSToSWMine with STREAMGEN and DDPMine

In our experimentation we have even compared our PSTMiner with PSToSWMine. Fig. 9. shows the precision minimum support comparisons for both the methodologies for mushroom dataset.

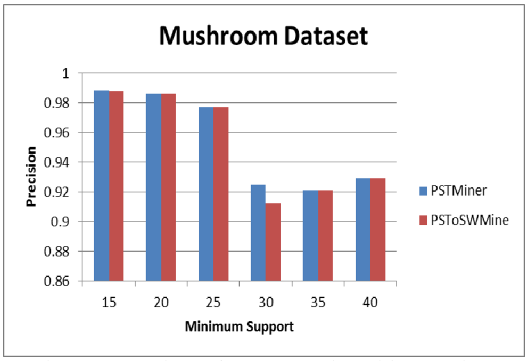


Fig. 9. Comparison of PSToSWMine with PSTMiner

The performances of our proposed methods are evaluated on 14 real data sets by comparing it with traditional associative classification approaches. Our approaches reached more than 10% improvement when compared with other approaches.

VII. CONCLUSION AND FUTURE WORK

This work presented our experiences mining stream association rules from real data and synthetic data for prediction. We used a novel dynamic tree to handle streaming data. Both our approaches are apt algorithms for mining data streams. Experimental results show that PSTMiner and PSToSWMine techniques increase the classification accuracy due to availability of large rule sets. As a future work, we plan to improve the performance of PSTMiner and PSToSWMine by reducing the number of rules generated without affecting accuracy of mining.

REFERENCES

[1] Obenshain, M.K: “Application of Data Mining Techniques to Healthcare Data”, Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[2] 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.

[3] L. Bing, H. Wynne, M. Yimming, "Intregrating Classification and Association Rule Mining," In Proc. Of the Fourth International Conference on Knowledge Discovery and Data Mining, New York,NY,pp.80-86,1998..

[4] W. Li, J. Han, J. Pei,"CMAR: accurate and efficient classification based on multiple class-association rules," in Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, pp. 369-376,2001.

[5] L. Su, H. Liu and Z. Song, “A New Classification Algorithm for data stream”. I.J.Modern Education and Computer Science, 4, 32-39, 2011.

[6] Hua-Fu Li, Suh-Yin Lee, “Mining frequent itemsets over data streams using efficient window sliding techniques”, Expert System with Applications, vol.36, no.2, pp. 1466-1477, 2009.

[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] X. Yin and J. Han, “CPAR: Classification Based on Predictive Association Rules,” Proc. Third SIAM Int’l Conf. Data Mining (SDM ’03), (May 2003).

[11] K.Prasanna Lakshmi, Dr.C.R.K.Reddy, “Compact Tree for Associative Classification of Data Stream Mining”, (IJCSI International Journal of Computer Science Issues, Vol 9, Issue 2, No 2, March 2012), ISSN(online) : 1694-0814

[12] D. J. Newman, S. Hettich, C. Blake, and C. Merz, “UCI Repository of Machine Learning Databases”. Berleley, CA: Dept. Information Comput. Sci., University of California,1998.

[13] J. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, (1993).

[14] B. Liu, W. Hsu, and Y. Ma, “Integrating Classification and Association Rule Mining,” Proc. Fourth Int’l Conf. Knowledge Discovery and Data Mining (KDD ’98), (Aug. 1998).

[15] Elena Baralis, Silvia Chiusano, Paolo Garza “A Lazy Approach to Associative Classification”, (TKDE-0302-0805), Identifier number: 10.1109/TKDE.2007.190677

[16] Tanbeer, S. K., Ahmed, C. F., Jeong, B.-S., and Lee, 2008. CP-tree: a tree structure for single-pass frequent pattern mining. In Proc. of PAKDD, Lect Notes Artif Int, 1022-1027.

[17] Chuancong Gao, Jianyong Wang, “Efficient item set generator discovery over a stream sliding window” in C IKM’09, November 2009, Hong Kong, China, ACM 978-1-60558-512-3/09/11