Ensemble Classification for Drifting Concept

E.Padmalatha
Research Scholar
JNTUH
Hyderabad

C.R.K.Reddy
Head of Department
CBIT, O. U
Hyderabad

B.Padmaja Rani
Head of Department
JNTUH
Hyderabad

ABSTRACT
Traditional data mining classifiers are used for mining the static data, in which incremental learning assumed data streams come under stationary distribution where data concepts remain unchanged. The concept of data can be changed at any time in real world application this refers to change in the class definitions over time. Classifier ensembles are rapidly gaining popularity in data mining Community, because they are comparatively more accurate, easy and react better to concept drift than single classifiers. They are general way of boosting classification accuracy. Their modularity provides natural path of absorbing changes by modifying ensemble member.The proposed approach uses ensemble classifiers to improve the accuracy of the classification in data streams .The performance of the classifiers tested with benchmark datasets from UCI machine learning repository.the experimental results prove that this approach great accuracy when comparing to the single classifier.

Keywords
Data stream ,concept drift ,boosting,static data.

1. INTRODUCTION
A data stream is a sequence of continuously arriving data items at a high speed which are real time, implicitly or explicitly ordered by timestamp, evolving and uncertain in nature. Data mining has recently emerged as a growing field of multidisciplinary research. It combines various research areas such as databases, machine learning, artificial intelligence, statistics, automated scientific discovery, data visualization, decision science, high performance computing etc. In recent years mining data streams in large real time environments has become a challenging job due to wide range of applications that generate boundless streams of data such as log records, mobile applications, sensors, emails, blogging ,credit card fraud detection, medical imaging, intrusion detection, weather monitoring, stock trading, planetary remote sensing etc. A vast amount of data related to all human activity is gathered for storage and processing purposes. A data stream mining is an approach to extract meaningful information (knowledge) from raw data streams and find changes and evolution of stream overtime. Various data mining tasks like clustering, classification etc can be performed on data streams in search of interesting useful patterns. Many issues can be envisioned with handling of data streams.

2. RELATED WORK
2.1. Classification
Classification is a two step process in which it initially learns from training data to form a classifier which is then used to classify unknown samples from testing data. The stream classifier must evolve to effectively indicate current class distribution in case of evolving data streams [13]. There are two widely used classification approaches: train the classifiers.

- **Single classifier**

It uses single classifier and works only for stationary data. Hence by using this we cannot process non stationary data. Some of the approaches used are as follows. VFDT approach [1] which used decision tree learning and hoeffding bounds to guarantee approximately correct output. Since it assumes data is stationary hence it can process in stable time and space. It is an anytime algorithm that does not requires to store any examples in memory and can learn by seeing each example only once.

The weakness of VFDT was improvised in CVFDT [2] focused to handle concept drifts. It utilizes window of examples to maintain up to date decision tree. It uses alternate sub tree, whenever old appears out of date replaces it with recent one which seems more accurate.

- **Ensemble classifier**

They are comparatively more accurate, easy and react better to concept drift than single classifiers. Classifier ensembles are rapidly gaining popularity in data mining community. They are general way of boosting classification accuracy. Their modularity provides natural path of absorbing changes by modifying ensemble member. SEA approach [3] which uses independent classifier for each chunk. This approach was designed to read blocks of data and ensembles are built incrementally. When ensemble is full, it discards old classifier to add more recent classifier and thus maintains stable size of ensemble. The method used to determine which existing tree should be replaced for inserting new tree affects the overall performance of this approach. It is easy to implement and quickly adapts to changes in concept.

Single partition and single chunk [4] used chunks of continuously flowing data stream in weighted ensemble classifier to train ensemble of classification model for classifying data streams. It fuses multiple classifiers weighted by their expected prediction accuracy on recent evaluating data. Small chunks drive up the error rates if number of classifiers in ensemble is not large.

In order to overcome the drawback of weighted ensemble classifier method, multi chunk partition multi ensemble method [5] was devised. It reduces error rate over single partition single chunk which uses simple majority voting. It keep optimally best k*v classifiers, where k is ensemble size and v is number of partitions. It uses labeled chunks to first train the classifiers.

3. LEARNING WITH CONCEPT DRIFT

Concept-drifting in data streams can be handled in three ways via: a) window-based approaches, b) weight-based approaches,and c) ensemble classifiers [5]. A window-based approach builds a classification model by selecting the instances within a fixed or dynamic stream sliding window, and adjusts window sizes based on the classification accuracy rate [7]. It combines all new and old instances together to generate a new training dataset, but only performs better for concept-drift in small datasets. In a weight-based approach, each training instance is assigned a weight. Based on the weights, some outdated training instances will be opportunistically discarded from the training dataset. The most popular evolving technique for handling concept-drift in data streams is to use an ensemble classifier [14] [15](combination of classifiers), which is shown in Figure 1 C1,C2,C3 are three classifiers used. The outputs of several classifiers are combined to determine a final classification, which is often called fusion rules. Also the weights are assigned to the individual classifier's outputs at each point in time. The weight is ususally a function of the historical performance in the past or estimated perfomance using 10-fold cross-validation. The best classifier among a number of classifiers (for either classification or prediction) can also be determined by performing 10-fold cross-validation. For example, in ensemble models, a mean error rate for each classifier is calculated and the best classifier with the lowest mean error rate will be selected.Research showed the re-use of traditional data mining algorithms in nonstationary environments is a difficult and challenging task [8]. Data mining algorithms should be adaptive so that it can be continuously updated with the novel class instances as time passes. Most of the existing data mining algorithms are trained on datasets with a fixed number of class labels.

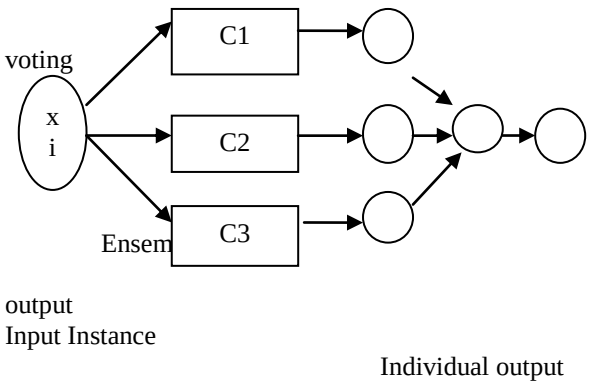

Fig1: Ensemble classifier

4. EXPERIMENTAL RESULTs

A real time data set German Credit Data from UCI repository[7] is used to demonstrate our approach.The following ensemble approaches i.e. Bagging,Boosting,Random space method and Random forest method has been considered.

Bagging

Bagging [11] goes away towards making a silk purse out of sow's ear, especially if the sow's ear is twitchy .It is relatively easy way to improve an existing method,since all that needs adding is a loop in front that selects the bootstrap sample and sends it to the procedure and back end that does the aggregation what one loses ,with the tree is a simple and interpretable structure.what one gains is increased accuracy.

AdaBoost

Boosting works by repeatedly running a given weak learning algorithm on various distributions over the training data, and then combining the classifiers produced by the weak learner into a single composite classifier. The first provably effective boosting algorithms were presented by Schapire[7] and Freund [8]AdaBoost is a boosting algorithm that has certain properties which make it more practical and easier to implement than its predecessors.[9]

JRip

JRip implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP. It is based in association rules with reduced error pruning (REP), a very common and effective technique found in decision tree algorithms. In REP for rules algorithms, the training data is split into a growing set and a pruning set. First, an initial rule set is formed that over ts the growing set, using some heuristic method. This overlarge rule set is then repeatedly simplified by applying one of a set of pruning operators typical pruning operators would be to delete any single condition or any single rule. At each stage of simplification, the pruning operator chosen is the one that yields the greatest reduction of error on the pruning set. Simplification ends when applying any pruning operator would increase error on the pruning set.

Random Forest

Breiman's ideas were decisively influenced by the early work of Amit and Geman (1997) on geometric feature selection, the random subspace method of Ho (1998) and the random split selection approach of Dietterich (2000). As highlighted by various empirical studies (see for instance Breiman, 2001; Svetnik et al., 2003; Diaz-Uriarte and de Andres, 2006; Genuer et al., 2008, 2010), random forests have emerged as serious competitors to state-of-the-art methods such as boosting (Freund and Shapire, 1996) and support vector machines **(Shawe-**Taylor and Cristianini, 2004). They are fast and easy to implement, produce highly accurate predictions and can handle a very large number of input variables without overfitting. In fact, they are considered to be one of the most accurate general-purpose learning techniques available.[10].

Tabel 1:showing the details of learning algorithm

Learning algorithm	RMSE	MAE	Kappa Statistics
Bagging	**0.41**	**0.33**	**0.34**
Adaboost	**0.43**	**0.36**	**0.1**
JRip	**0.43**	**0.36**	**0.31**
Random Forest	**0.40**	**0.33**	**0.39**

For experiment German credit card dataset [7]is used which is having 1000 instances and 21 attributes . According to Tabel1 Bagging and Random forest algorithms are having less Root Means squared error(RMSE),and high Kappa statistics.The performance of above mentioned algorithms are pictorially shown with help of the ROC curves based on the confusion matrix. Based on the experimental results for classification of data streams Bagging and Random forest algorithms can be considered for learning,JRip is not a "Weak "classifier ,but is some how damps the effect of ensemble learning.

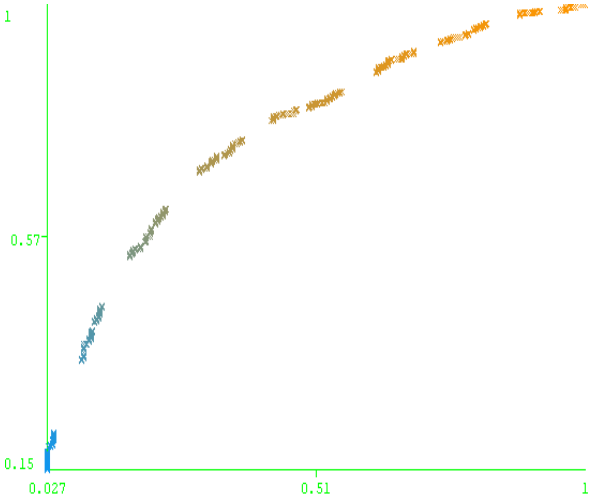

Fig2: ROC curve of Bagging

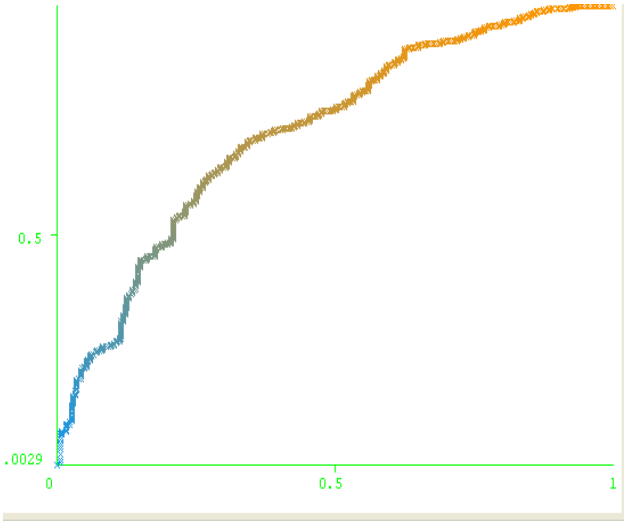

Fig3: ROC curve of Adaboost

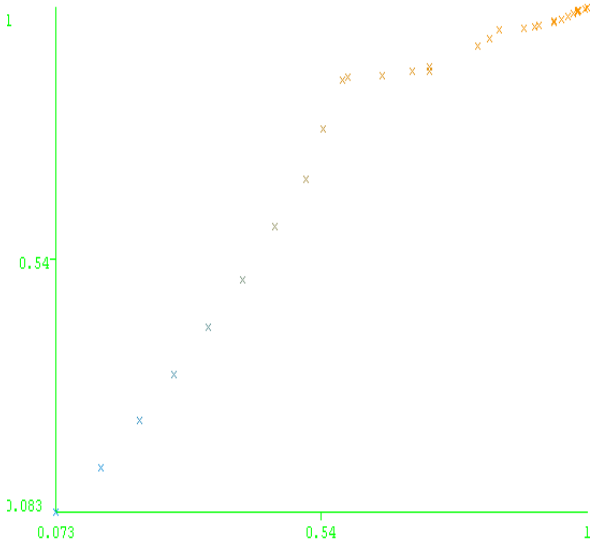

Fig4: ROC curve of JRip

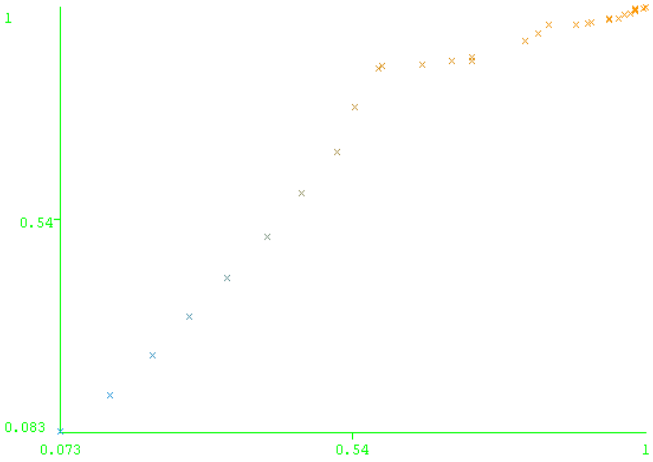

Fig5: ROC curve of RandomForest

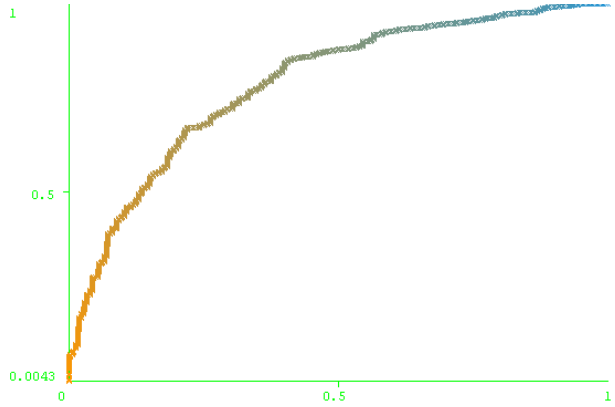

Fig6:ROC curve for bagging with random forest

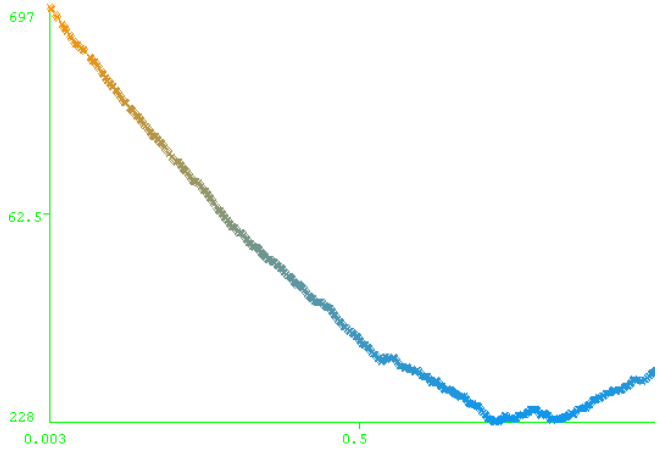

Fig 7: Cost benefit curve for Bagging with Random Forest

5. CONCLUSION

Ensemble Classifiers are used to get high accuracy in the classification of the data streams.In the proposed method the combination of the bagging and the Random forest are giving the high accuracy . Bagging and Random Forest are two methods which transform the "weak "individual models in a "strong" ensemble of models.Ensemble models can be used in novel class detection in concept drifting data streams with this misclassification error can be minimized in the concept drifting classification.The future work will be focusing on the concept drifting c;assification using ensemble models.

6. REFERENCES

[1] W .Nick Street, Yong Seog Kim, "A streaming ensemble algorithm (sea) for large-scale classification", In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, 2001, pp 377-382.

[2] Pedro Domingos, Geoff Hulten, " Mining High Speed Data Streams",KDD-00 in proceeding of sixth ACM SIGKDD international conference on knowledge discovery and data mining, USA, 2000, pp 71-80.

[3] Geoff Hulten, Laurie Spencer, Pedro Domingos, "Mining time changing data streams", ACM, USA, 2001, 97-106.

[5] Haixun Wang, Wei Fan, Philip S. Yu, Jiawei Han, "Mining ConceptDrifting Data Streams Using Ensemble Classifiers", SIGKDD '03,ACM,USA,2003,pp 226-235.

[6] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, and Bhavani Thuraisingham, "A Multi-partition Multi-chunk Ensemble Technique to Classify Concept-Drifting Data Streams" Springer-Verlag, Berlin Heidelberg, 2009, pp 363–375.

[7] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. UCI repository of machine learning databases, 1998.

[8] Robert E. Schapire. The strength of weak learnability. Machine Learning, 5(2):197–227,1990.

[9] Yoav Freund. Boosting a weak learning algorithm by majority. Information andComputation,121(2):256–285,1995.

[10] Leo Breiman (2001). Random Forests. Machine Learning. 45(1):5-32.

[11] Leo Breiman (1996). *Bagging predictors*. Machine Learning. 24(2):123-140.

[12] Yoav Freund, Robert E. Schapire: Experiments with a new boosting algorithm. In: Thirteenth International Conference on Machine Learning, San Francisco, 148-156, 1996.

[13] Tin Kam Ho (1998). *The Random Subspace Method for Constructing Decision Forests.* IEEE Transactions on Pattern Analysis and Machine Intelligence. 20(8):832-844.

[14]H. Wang, W. Fan, P. S. Yu, and J. Han. Mining conceptdrifting data streams using ensemble classifiers. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages226–235, New York, NY, USA, 2003. ACM Press.

[15] W. N. Street and Y. Kim. A streaming ensemble algorithm (sea) for large-scale classification. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 377–382, New York, NY, USA, 2001. ACM Press.