Learning Framework for Non-stationary and Imbalanced Data Stream

Meenakshi A. Thalor[1], Dr. S. T. Patil[2]

[1]Research Scholar, Department of Computer Engineering, Vishwakarma Institute of Technology, Savitribai Phule Pune University, Pune,India

[2]Professor, Department of Computer Engineering, Vishwakarma Institute of Technology, Savitribai Phule Pune University, Pune,India

[1] thalor.meenakshi@gmail.com

[2]stpatil77@gmail.com

Abstract—Although learning on non-stationary data and imbalanced data have been extensively studied in the literature separately, however little work has been done to tackle the imbalanced issue on non-stationary data stream as the joint probability distribution between the data and classes changes with time and may results skewed class distribution. Especially in airlines delay detection, data sources are dynamic generated at high speed in real time, type of delay activity changes with time and in each chunk of stream, delay detection instances are less so concept drift and class imbalanced issues arises simultaneously. Through this research work we propose an ensemble based incremental learning approach towards non-stationary imbalanced data stream.

Keyword-Concept Drift, Ensemble, Imbalanced Data, Incremental Learning,Non-stationary Data

I. INTRODUCTION

Learning in non-stationary imbalanced environment is most difficult and challenging tasks in data mining and machine learning. Conventionally, the learning carries out on static data i.e. complete data is available at the time of learning but nowadays most of the online applications provide data in form of streaming hence data arrives after some time(batch) or available continuously(instance) and it's not possible to process such growing volume of the data in multiple passes rather we need to process a data item at most once. Extra challenge arises when in streaming data, definition of class is changing with time so learning in such environment faces the problem of concept drift[1,2] which sometimes results in class unbalancing in data stream.Concept drifts change the classifier results over time. The classifier is most likely to be outdated after a time due to the continuous change of the streaming information on a temporal basis.So it is very important to train classifiers incrementally over the time so that they can learn different concepts of non-stationary imbalanced data streams. For providing training to classifiers incrementally over the time so that they can learn different concepts of non-stationary imbalanced data streams we are employing ensemble based approach [3] where a set of classifiers whose individual predictions are combined in some way to classify unseen data. The approach in ensemble systems [4] is to generate many classifiers, and combine their outputs in such a way that this combination will improve the performance as compared to single classifier.

In literature, batch incremental learning(gather example and then train model) and instance incremental learning (learning at the time of arrival) approaches are used to handle streaming data. This paper concentrates only on batch incremental learning and state-of-the-art methods for handling concept drift and unbalanced data streams generally used ensemble approach with sampling.Goa et al. proposed uncorrelated bagging strategy [5] which propagates all previously collected minority examples into current training chunk to balance data. Chen et al. proposed SERA[6], MuSeRA[7] and REA[8] which selectively propagates minority examples from previous chunks into current training chunk using Mahalanobis distance and a k-nearest neighbors algorithm. Learn^{++}.CDS[9] uses Learn^{++}.NSE and SMOTE algorithm where Learn^{++}.NIE [9], uses Learn^{++}.NSE [10] and BaggingPropagation where for each batch they create different balanced subsets using undersampling and then created combine models learnt on each balanced subset. Lichtenwalter and Chawla [11] suggest propagation ofallpositives instances and previous chunks's negative class observations to increase the boundary definition between the two classes. Hoens and Chawla in [12] used Naïve Bayes to select previous positive instances which are relevant to the current minority class context.

This work introduces a classification approach on non-stationary imbalanced data using an ensemble based approach which will generate a new classifier sequentially for each bag of data arriving at time t; only weights will be assigned to classifiers based on its error rate on current training chunk. The performance of proposed approach will be improved by propagating the misclassified instances to next subsequent classifier with its training chunk.

II. PROPOSED APPROACH

Fig. 1 depicts the architectural view of the ensemble for non-stationary imbalanced data stream (ENSIDS) where input data stream is non-stationary and imbalanced. State-of-the-art techniques have addressed non-stationary and imbalanced data learning by assigning weights to instances or propagating minority instances between batches. Our approach is, all instances are equally important while they are employing for training so uniform weight is considered and secondly, we are propagating the false positive and false negative observations(represented by solid arrow) of a classifier to subsequent classifier for improving the performance.

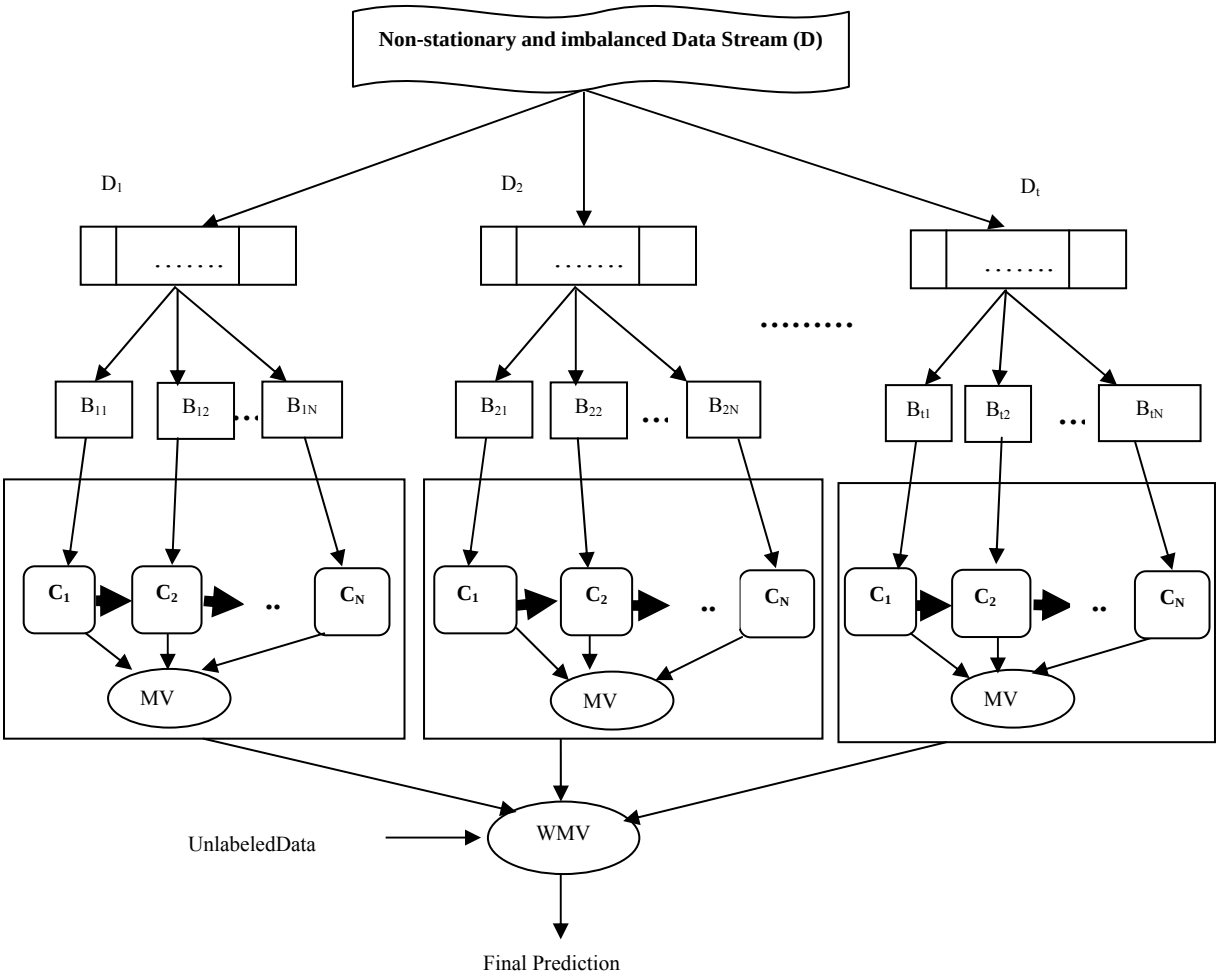

Fig. 1.Ensemble for non-stationary imbalanced data stream

Initially, for each batch of data at time t, a sub ensemble is generated for each batch. In each sub ensemble, a classifier is generated for each bag of batch data then the performance of classifier is evaluated with current batch of instances. For each bag of batch data, the incorrect classified instances of previous classifier would be propagated to current batch and then apply classification scheme. This process is continued till we get classifiers for all bags of current batch. Then all these classifiers are combined and apply weighted majority voting(WMV) scheme and finally we get labels for unseen data.

The time complexity of proposed approach is O(t*k*O(x*m)+k*t*m) where O(x*m) is the time complexity of Naïve Bayes classifier, x is number of features and m is number of instances in training set, k indicates number of classifiers, t indicates number of data chunks to be predicted.

III. EXPERIMENTAL RESULTS

For experiment analysis, all tested algorithms are implemented in Java using MOA and WEKA framework. For doing the comparison of proposed approach and existing approaches we are considering Naïve Bayes as base classifiers, different batch sizes and no pruning strategy. For evaluation,we have considered different evaluation metrics like Precision Recall, Accuracy, F-Measure, G-Mean[13].The proposed approach is tested over following datasets.

1. Diabetes Dataset: This is Pima Indians diabetes imbalanced dataset contains 768 examples taken from UCI machine learning repository where we are considering eight attributes to find class labels Tested_Positive and Tested_Negative.

2. Airlines Dataset: Thisis non-stationary and imbalanced dataset consists of 5000 examples to represent the detail of flight arrival departure for all commercial flights within the USA for year 1989 where we are considering 29 attributes to find class label Delay and NoDelay.

Fig. 2 shows that on diabetes dataset the false negative observations are less which play a significant role in medical and diagnostic applications hence the recall, accuracy, f-measure and g-mean of proposed approach is higher compare to existing algorithms.

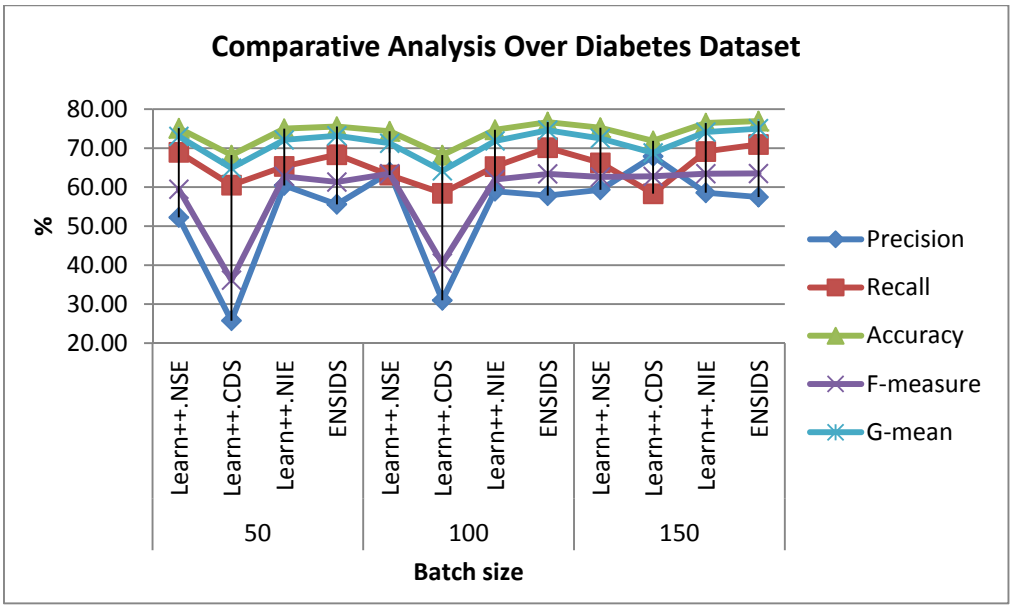

Fig. 2.Performance evaluation over diabetes dataset

In medical and diagnostic domain positive and negative predictive values are useful when considering the value of a test and proposed approach gives higher values for sensitivity and specificity evaluation measure. The values of evaluation measures proved the validity of proposed work.

Generally, there always remains a tradeoff between precision and recall i.e. if the precision rate is high, then the recall rate will be low. In imbalanced application domain, the recall is more important than precision but we can observer the balance between precision and recall using f-measure. Fig. 3 depicts that on airlines non-stationary and imbalanced dataset the accuracy, recall, f-measure and g-mean of proposed work gives better result as compare to existing algorithms.This presents that one can use proposed work on non-stationary and imbalanced dataset and can achieve a good balance between majority and minority instances.

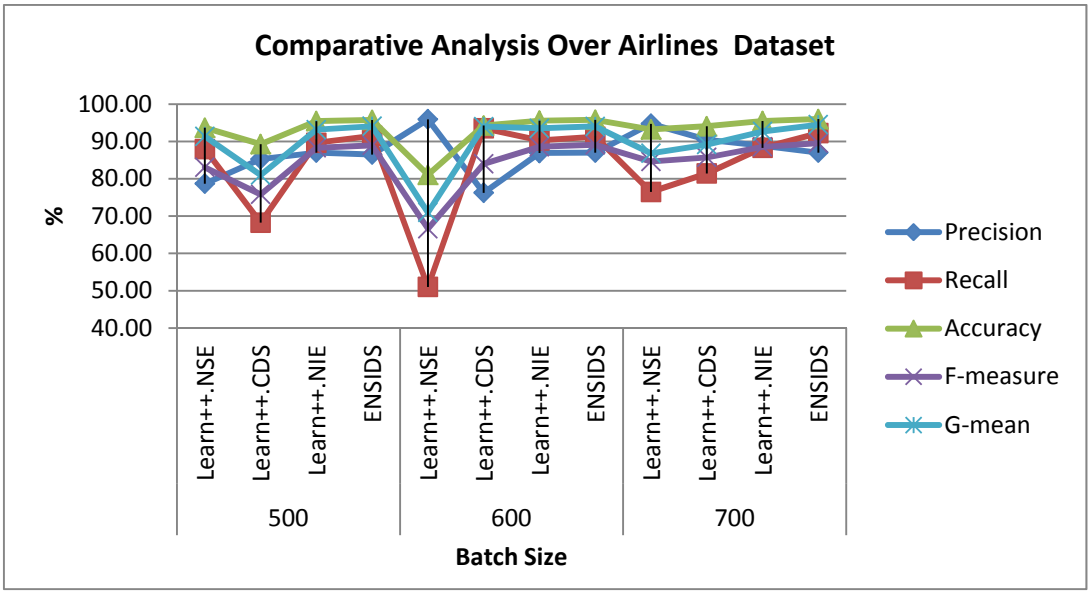

Fig. 3.Performance evaluation over airlines dataset

After testing the proposed work on different datasets, evaluation measures confirm the validity and excellence of the proposed approach. The non-stationary data can have class imbalanced problem so result can be biased toward the majority class; thus the classifier tends to misclassify the minority class instances. In imbalanced application area, the proposed system can be used and can provide a balance between majority and minority instances.

IV. CONCLUSION

From the implementation and analysis of proposed system we can conclude that proposed work gives better results as compared to Learn^{++}.NSE,Learn^{++}.CDS and Learn^{++}.NIE on different datasets. The selection of optimal batch size varies from dataset to datasets. Experiments have been conducted on real datasets which gives a remark that the proposed model is very generic and can be adapted to many other situations. As no measure is taken to detect the point of drift so drift detection mechanism can also be added in the proposed approach as future work.

REFERENCES

[1] Moreno-Torres, J., Raeder, T., Alaiz-Rodríguez, R., Chawla, N.V.,Herrera, F.,A unifying view on dataset shift in classification, Pattern Recognition. 45, (2011) ,521–530 .
[2] Meenakshi A.Thalor and S.T. Patil, Learning on High Frequency Stock Market Data Using Misclassified Instances in Ensemble,International Journal of Advanced Computer Science and Applications, Vol. 7 No.5, 2016.
[3] Meenakshi A.Thalor ,S.T.Patil, Review of Ensemble Based Classification Algorithms for Nonstationary and Imbalanced Data ,IOSR Journal of Computer Engineering, Vol. 16, pp. 103-107, 2014.
[4] R. Polikar ,Ensemble Based Systems in Decision Making, IEEE Circuits and Systems Magazine, Vol. 6, No. 3. ,pp. 21-45, 2006
[5] J. Gao, B. Ding, W. Fan, J. Han, and P. S. Yu. Classifying data streams with skewed class distributions and concept drifts. Internet Computing,12(6):37–49, 2008.
[6] S. Chen and H. He, SERA: Selectively recursive approach towards nonstationary imbalanced stream data mining, International Joint Conference on Neural Networks , pp. 522-529, 2009.
[7] Sheng Chen, Haibo He, Kang Li, and Sachi Desai ,MuSeRA: Multiple Selectively Recursive Approach towards Imbalanced Stream Data Mining, IEEE World Congress on Computational Intelligence, 18-23, 2010.
[8] S. Chen and H. He, Towards incremental learning of nonstationary imbalanced data stream: a multiple selectively recursive approach,Evolving Systems,vol. 2, no. 1, pp. 35-50, 2011.
[9] Gregory Ditzler, Robi Polikar, Incremental Learning of Concept Drift from Streaming Imbalanced Data, IEEE Transactions on Knowledge & Data Engineering, vol.25, no. 10, pp. 2283-2301, 2013.
[10] Elwell R. and Polikar R., Incremental Learning of Concept Drift in Non-stationary Environments, IEEE Trans. on Neural Networks, vol. 22, pp. 1517-1531, 2011.
[11] R. N. Lichtenwalter and N. V. Chawla, Adaptive methods for classification in arbitrarily imbalanced and drifting data streams. In New Frontiers in Applied Data Mining, pages 53–75. Springer, 2010.
[12] T. R. Hoens, N. V. Chawla, and R. Polikar, Heuristic updatable weighted random subspaces for non-stationary environments, IEEE 11th International Conference on Data Mining , pages 241–250, 2011.
[13] Meenakshi A.Thalor and S.T. Patil,Incremental Learning on Non-stationary Data Stream using Ensemble Approach,International Journal of Electrical and Computer Engineering,Vol. 6 No. 4,2016

AUTHOR PROFILE

Meenakshi A. Thalor received the BE degree in Information Technology from Kurukshetra University in 2004. She completed M.E. in Information Technology from Pune University in 2011. Her Research interest includes Data Mining, Machine learning and Pattern Recognition. Currently she is working with All India Shri Shivaji Memorial Society's Polytechnic, Pune and pursuing Ph.D. Computer Engineering from Savitribai Phule Pune University, Pune.

Shrishailappa T. Patil received degree of M.Tech. Computer Science and Engineering from Vishveshwaraya Technological University, Belgum in July 2003. Ph.D. Computer Engineering and Technology from BharatiVidyapeeth Deemed University, Pune. He is having 27 years of experience in teaching as a lecturer, Training and Placement Officer, Head of the department, Assistant Professor, Professor and Principal. He is presently working as a full tenure Professor in Computer Engineering and Information Technology Department in Vishwakarma Institute of Technology, Pune (India). Recipient of best teacher award two times and fetched research grants from AICTE, New Delhi three times. He published more than 14 papers international journals and conference proceedings, more than 65 papers in national journals and conferences.