IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. IX (Feb. 2014), PP 103-107
www.iosrjournals.org

Review of Ensemble Based Classification Algorithms for Nonstationary and Imbalanced Data

Meenakshi A.Thalor[1], Dr.S.T.Patil[2]
[1](Research Scholar, Department of Computer Engineering, Pune University, Pune, India)
[2](Professor, Department of Computer Engineering, Pune University, Pune, India)

Abstract *: Learning data samples from a non-stationary distribution has been shown to be a very challenging problem in machine learning, because the joint probability distribution between the data and classes changes over time. Most real time problems as they change with time can suffer concept drift. For example, a recommender or advertising system, in which customer's behavior may change depending on the time of the year, on the inflation and on new products made available. An additional challenge arises when the classes to be learned are not represented equally in the training data i.e. classes are imbalanced, as most machine learning algorithms work well only when the class distributions are balanced. The objective of this paper is to review the ensemble classification algorithms on the framework of non-stationary and imbalanced dataset, with focus on two-class problems. In addition, we develop a thorough comparison of these algorithms by the consideration of the most significant published approaches.*

Keywords: *Concept Drift, Ensemble, Imbalanced Data, Incremental Learning, Non-stationary Data*

I. INTRODUCTION

Concept drift [1] and class imbalance are traditionally addressed separately in machine learning, yet data streams can experience both phenomena. Due to the complexity of each of these issues, the combination of class imbalance and concept drift is understudied. So this work introduces classification algorithm on non-stationary and imbalanced data using ensemble based approach which will explicitly and simultaneously address the aforementioned phenomena.

1. Ensemble Based Classification

In ensemble based classification [2] a set of classifiers whose individual predictions are combined in some way to classify new examples. The strategy in ensemble systems[3] is to create many classifiers, and combine their outputs in such a way that this combination will improves the performance over a single classifier.Fig.1 is showing the basic steps of ensemble based classification.

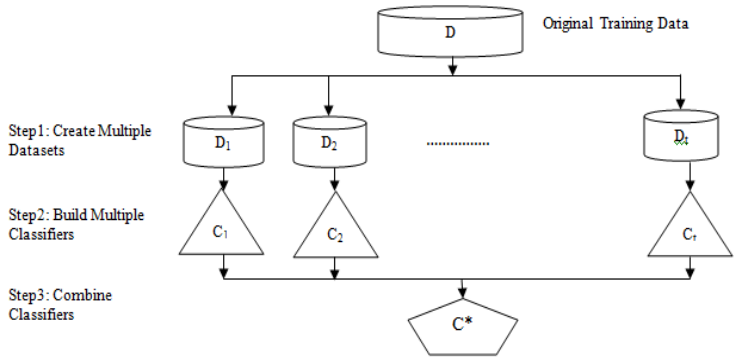

Fig.1: Ensemble based Learning

2. Nonstationary and Imbalanced Data

Nonstationary data is one type of time series data where data at time t is not equal to data at time t+1 The time series Y_t is nonstationary if for all values, and every time period, it is true that:

$E(Y_t) \neq \mu$ (not having constant mean)

$Var(Y_t) \neq \sigma^2$ (not having constant variance)

Fig. 2 is showing the difference between stationary data and nonstationary data.

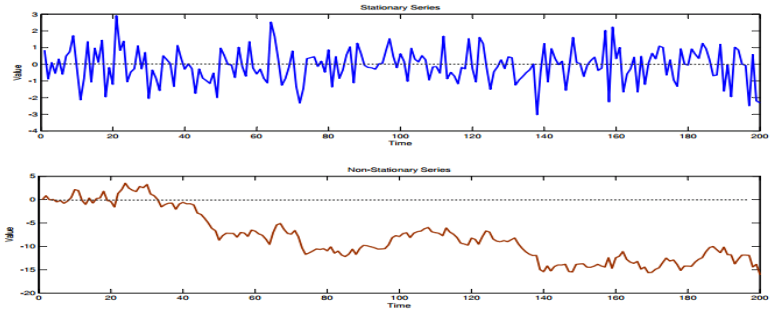

Fig. 2: Stationary series and Non-stationary series

Traditional data mining assumes that each dataset is produced from a single, static and hidden function. That is, the function (model) generating data at training time is the same as that of testing time. Whereas in data stream, data is continuously coming and the function which generating instances at time *t* need not be the same function at time *t+1*. This difference in the underlying function is called as concept drift [4, 5]. Thus, past data may become irrelevant for the current context, Concept drift is defined as [6]

$$p_{tr}(y|x) \neq P_{tst}(y|x) \; and \; P_{tr}(x) = P_{tst}(x) \; in \; X \rightarrow Y \; Problems$$

Following is the categories of concept drift algorithms:

1. Online or batch approaches depends on the number of training instances that is one instance or a batch of instances used at training time.
2. Single classifier or ensemble-based approaches depends on the number of classifiers used in decision making that is one classifier or multiple classifiers.
3. Incremental or non-incremental approaches based on whether previous data is reused to refine classifiers or not. Incremental learning is a useful and practical technique of learning new data over time. Learning incrementally is not only useful because it allows us to refine our models/classifiers over time without using previous data.
4. Active or passive approaches depending on drift detection mechanism is used or not. In active drift detection algorithm first determines the point where a change /drift have occurred then take any corrective action whereas passive drift detection algorithm assumes that in streaming data drift may occur at any instance of time hence updates a model every time whenever new data arrive.

In this paper we have reviewed only batch, ensemble, incremental and passive approaches to handle concept drift.

Unbalanced data, or imbalanced data [7, 8], refers to an unequal representation of classes that is one class is overrepresented by other class. In imbalanced pattern recognition problem there are two types on class-majority (negative) and minority (positive). The majority (negative) class is those class or set of classes that appear more frequently in a dataset. The minority (positive) class is rarely appears in the training data. The minority (positive) class is of great interest than the majority (negative) class in pattern recognition. Fig. 3 is showing the difference between balanced and imbalanced two class data.

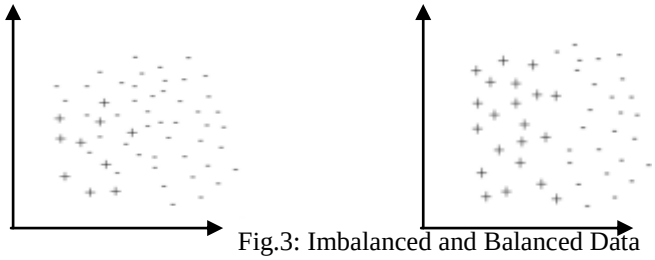

Fig.3: Imbalanced and Balanced Data

Following is the categories of class imbalances:

1. Preprocessing methods/Sampling Methods:
   The use of sampling techniques in imbalanced learning is to modify imbalanced data set by some mean and obtain a balanced distribution. Oversampling appends more minority data to the original dataset while under-sampling removes some of majority data from the original data set.
2. Cost sensitive Methods
   Cost-sensitive learning methods take the misclassification costs (which are computed from the cost associated with the four outcomes of two class confusion matrix) in to consideration. The purpose of this type of learning is to decrease the total cost. In cost-sensitive learning we not assign any cost to

correct classifications. Since the positive (minority) class is of more interest than the negative (majority) class. Especially in medical diagnosis the cost of false negative (positive example was misclassified) is more important.

3. Ensemble Based Methods
   In this ensemble learning algorithm are combined with one of the techniques mentioned above. The integration of data level approach and the ensemble learning algorithm will produce a new hybrid method which manipulates the data before training each classifier.

II. STATE OF ART ON ENSEMBLE BASED CLASSIFICATION ON NONSTATIONARY AND IMBALANCED DATA

In Literature there are very few algorithms are available to overwhelm the class imbalance and nonstationary problems simultaneously, some of these are studied as given below.

W. Nick Street [9] proposes SEA (Streaming Ensemble Algorithm) which builds separate classifier on sequential chunk of training points. Component classifiers added into ensemble with a fixed size. If ensemble is full, Old classifiers are replaced by new classifier only if they satisfy quality criterion. Performance estimates are done by testing the new tree (and the existing ensemble) on the next chunk of data points.

Gao et. al. [10] proposed general framework for building accurate classification models on skewed data streams. Each incoming chunk S is split into two parts P (positive examples) and Q (negative examples). In imbalanced data the size of P is much smaller than that of Q. K samples are generated to train C_1 to C_k classifiers. Positive examples are propagated to each sample and each negative example in the training set is randomly propagated to these samples, taking in to consideration that the negative examples in the samples are completely disjoint. To make distribution balanced, a parameter r as input, which is the ratio of positive examples over negative examples in each sample is also considered. By collecting all minority (positive) class instances, this approach implicitly assumes that no drift in the minority instances.

Chen and He propose SERA[11] algorithm selects the best minority class instances which are computed by the mahalanobis distance instead of propagating all minority class instances to each sample. So each previous minority class set computes its mahalanobis distance with the current minority class examples, depending on that SERA add best minority class instances into the current training data chunk. SERA build an ensemble of classifiers using bagging. Thus SERA algorithm possibly not able to track drift in minority instances as it assumes no separate sub-concepts within the minority class concept exists.

Chen and He propose REA[12] framework to learn on nonstationary imbalanced data streams which solves the flaw of SERA by adopting the k-nearest neighbors approach to estimate the similarity degree where each previous minority class example computes the number of minority class examples which are within its k-nearest neighbors of current training data chunk. REA retains all hypotheses built on each training data chunks over time and weighs them based on their classification performance on the current training data chunk. The prediction on the current testing data set is made by weighted combination of all the hypotheses.

Michael D. Muhlbaier and Robi Polikar propose Learn^{++}.NSE [13, 14, 15, 16], algorithms to learn in nonstationary environments. For each incoming batch of data Learn^{++}.NSE trains one new classifier and combines them using dynamically weighted majority (DWM) [17, 18] voting .In DWM each classifier's weight is determined by its error, age, and performance on current and all previous environments.. The algorithm learns in incremental way i.e. it does not use prior data. Learn^{++}.NSE does not discard any of the classifiers, but rather temporarily lowers their voting weights or even suspends them in proportion of their performance on the current environment. Learn++.NSE can accommodate a wide variety of drift scenarios like gradual, abrupt, slow, fast , cyclical, and variable rate drift.

Finally, Ditzler and Polikar [19,10,21] propose Learn^{++}.NIE which is extension of Learn^{++}.NSE algorithm with the case of class imbalance. The authors propose Learn^{++}.CDS (combination of two algorithms i.e. Learn^{++}.NSE and SMOTE [22]) and Learn^{++}.NIE (learning in nonstationary and imbalanced environments). The authors of Learn^{++}.NIE used Learn^{++}.NSE algorithm with bagging instead of a single base classifier and claim reduction in error.

Table1: Comparison of Existing algorithms to handle nonstationary and imbalanced data on Electricity Pricing Dataset

Parameter	**SEA**	**UCB**	**SERA**	**Learn^{++}.NSE**	**Learn^{++}.CDS**	**Learn^{++}.NIE**
Approach	Ensemble based algo. to handle concept drift.	Uses all previous minority instances into sample to train classifier	Collects selective minority examples from previous training data chunks into the current training data chunk	Algorithm for non-stationary data only. Does not provide any mechanism for imbalanced data	Combination of SMOTE and Learn^{++} .NSE	Building bagging based sub ensembles at each time stamp and uses Learn $^{++}$.NSE
Incremental	Not	Not	Not	Yes	Yes	Yes
Ensemble size	Fixed	Fixed	Fixed	Fixed	Not forced a fixed size	Not forced a fixed size
RCA	92.15	68.23	76.42	90.75	88.48	82.6
F-measure	9.37	18.68	19.91	15.4	18.09	20.79
AUC	58.48	69.74	62.42	59.66	60.58	72.45
Recall	10.53	58.87	46.46	16.87	22.91	38.72

From Table 1 it is clear that UCB and SERA are not truly incremental approaches to handle concept drift with imbalance data as these requires access to previous data. Learn^{++}.CDS and Learn^{++}.NIE are extension of Learn^{++}.NSE algorithm where all classifiers are retained and believe that such classifiers provide relevant information for the current environment because of this the computation cost as well as storage requirement is high. Learn^{++}.NIE first divides the training data of current chunk into the majority and minority classes then a sub-ensemble of classifiers is generated by using the minority class instances and majority class instances are randomly sampled. Hence employs a variation of bagging to generate sub-ensembles of classifiers. The problem with Learn^{++}.NIE is that misclassifies instances are not considered here.

Table 2 is showing the rank of above mentioned algorithms. Performance of Learn^{++}.NIE is on rank one on electricity pricing dataset as it score more on AUC and F-measure as compare to others. Performance of UCB and SERA is equal and holds rank 2.Perfromance of Learn^{++}.NSE and SEA is less as they are not designed with considering the issue of imbalance data.

Table 2: Ranks of all the algorithms on Evaluation Measures for Electricity Pricing dataset

	RCA	**F-measure**	**AUC**	**Recall**	**Average**	**Rank**
SEA	1	6	6	6	4.75	5
UCB	6	3	2	1	3	2
SERA	5	2	3	2	3	2
Learn^{++}.NSE	2	5	5	5	4.25	4
Learn^{++}.CDS	3	4	4	4	3.75	3
Learn^{++}.NIE	4	1	1	3	2.25	1

Fig. 4 is showing the comparison of existing algorithms based on their evaluation measures. It is clear from the fig. 4 that only the RCA (Raw Classification Accuracy) is not an effective evaluation measure when there is imbalance data.

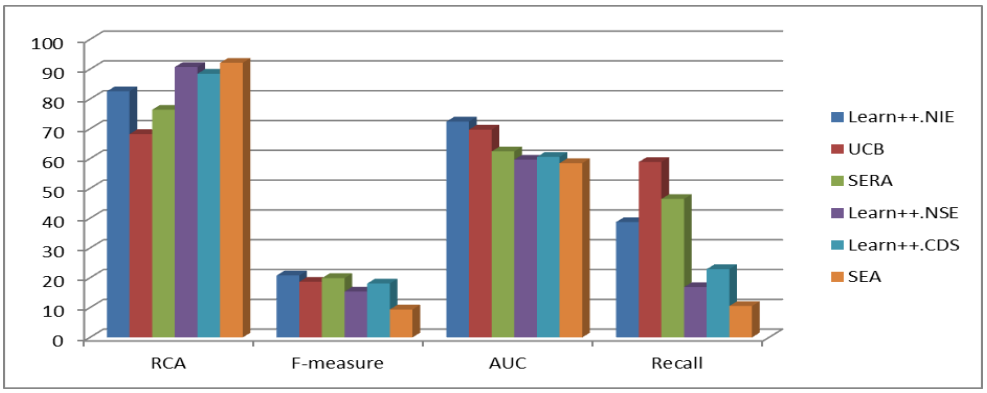

Fig. 4: Comparison of existing algorithms based on evaluation measures

Other evaluation measures [23, 24, and 25] like recall, f-measure and AUC (Area under Curve) should be considered for non-stationary and imbalanced data.

III. CONCLUSION

In this paper, the state of the art on ensemble methodologies to deal with non-stationary and class imbalance problem has been reviewed. Our survey concludes that batch learning, incremental learning, ensemble learning, passive learning and imbalanced learning is widely used in combination as a research issue. Finally, we have concluded that ensemble-based algorithms are worthwhile, improving the results that are obtained by the usage of data preprocessing techniques and training a single classifier. The use of more classifiers makes them more complex, but this growth is justified by the better results that can be assessed.

.

REFERENCES

[1] Tsymbal A.,The problem of concept drift: definitions and related work, Technical Report TCD-CS-2004-15, Department of Computer Science, Trinity College,2004.

[2] L. Rokach, Ensemble-based Classifiers, Artif. Intell. Rev.,vol. 33, pp. 1–39, 2010

[3] R. Polikar ,Ensemble Based Systems in Decision Making, IEEE Circuits and Systems Magazine, Vol. 6, No. 3, pp. 21-45, 2006

[4] G. Widmer and M. Kubat, Learning in the presence of concept drift and hidden contexts, Machine Learning, vol. 23, no. 1, pp. 69-101, 1996.

[5] Martin Scholz, Ralf Klinkenberg, An Ensemble Classifier for Drifting Concepts, Second International Workshop on Knowledge Discovery in Data Streams, pp. 53-64 , Vol. 11 ,2005.

[6] Moreno-Torres J., Raeder T., Alaiz- Rodríguez R., Chawla N.V., Herrera F.,A unifying view on dataset shift in classification, Pattern Recognition 45, pp. 521–530 , 2011.

[7] Haibo He and E. A. Garcia, Learning from Imbalanced Data, IEEE Tranansactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept.2009.

[8] D. Williams, V. Myers, and M. Silvious, Mine classification with imbalanced data, IEEE Geosci. Remote Sens. Lett., vol. 6, no. 3, pp. 528–532, Jul. 2009.

[9] W. N. Street and Y. Kim, A streaming ensemble algorithm (SEA) for large-scale classification, Int'l Conf. on Knowledge Discovery & Data Mining, pp. 377-382, 2001.

[10] J. Gao, W. Fan, J. Han, and P. S. Yu, A general framework for mining concept-drifting data streams with skewed distributions, SIAM International Conference on Data Mining, vol. 7, 2007.

[11] S. Chen and H. He, SERA: Selectively recursive approach towards nonstationary imbalanced stream data mining, International Joint Conference on Neural Networks (IJCNN 2009), pp. 522-529, Atlanta, GA, 2009.

[12] 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.

[13] Elwell R. and Polikar R., Incremental Learning of Concept Drift in Non-stationary Environments, IEEE Transactions on Neural Networks, vol. 22, no. 10, pp. 1517-1531, October 2011.

[14] R. Elwell and R. Polikar, Incremental Learning of Variable Rate Concept Drift, International Workshop on Multiple Classifier Systems (MCS 2009) in Lecture Notes in Computer Science, vol. 5519, pp. 142-151, Reykjavik, Iceland, 2009.

[15] M. Muhlbaier and R. Polikar, An Ensemble Approach for Incremental Learning in Nonstationary Environments, Multiple Classifier Systems, pp. 490-500, 2007.

[16] M. Muhlbaier and RobiPolikar, Multiple Classifiers Based Incremental Learning Algorithm For Learning In Nonstationary Environments, Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, vol. 6, pp. 3618–3623. 2007.

[17] J. Z. Kolter and M. A. Maloof, Dynamic weighted majority: an ensemble method for drifting concepts,Journal of Machine Learning Research, vol. 8, pp. 2755-2790, 2007.

[18] J. Z. Kolter and M. A. Maloof, Dynamic Weighted Majority: A New Ensemble Method for Tracking Concept Drift, Proceedings of the Third International IEEE Conference on Data Mining, pp. 123–130,2003.

[19] G. Ditzler and R. Polikar, Incremental Learning of Concept Drift from Streaming Imbalanced Data, IEEE Transactions on Knowledge and Data Engineering, 2012.

[20] G. Ditzler and R. Polikar, An ensemble based incremental learning framework for concept drift and class imbalance,World Congress on Computational Intelligence - International Joint Conference on Neural Networks , pp. 1-8, Barcelona, Spain, 2011

[21] G. Ditzler, R. Polikar, and N. Chawla, An Incremental Learning Algorithm for Non-stationary Environments and Class Imbalance,20th International Conference on Pattern Recognition (ICPR 2010), pp. 2997-3000, Istanbul, Turkey, 2010.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, SMOTE: Synthetic Minority Oversampling Technique, Journal of Artificial Intelligence,vol. 16, pp. 321-357, 2002.

[23] Jesse Davis, Mark Goadrich ,The Relationship Between Precision-Recall and ROC Curves, In Proceedings of the 23rd international conference on Machine learning , pp. 233-240,2006

[24] Flach, Peter A. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. Machine Learning-International Workshop Then Conference-.Vol. 20.No. 1. 2003.

[25] Bradley A. P., The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recog., vol. 30, no. 7, pp. 1145–1159, 1997.