Volume 4, Issue 4, April 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Comparison of Ensemble Based Classification Algorithms

Meenakshi A.Thalor*
Research Scholar,
Computer Engineering,
University of Pune, Pune, India

Dr. S.T.Patil
Professor,
Computer Engineering,
Vishwakarma Institute of Technology,Pune , India

Abstract— Learning data sampled from a nonstationary 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 existing ensemble classification algorithms on the framework of nonstationary and imbalanced datasets, 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, Nonstationary 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 presents thorough comparison of existing classification algorithm on nonstationary data and classification algorithm on nonstationary and imbalanced data using ensemble based approach [2].

A. *Nonstaionary 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)

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[3,4]. Thus, past data may become irrelevant for the current context, e.g. in Consumer ad relevance, spam detection, weather prediction .Concept drift is defined as [5]

$$p_{tr}(y|x) \neq P_{tst}(y|x) \, and \, P_{tr}(x) = P_{tst}(x)$$

Following is the categories of concept drift approaches:

1. *Online or batch approach*:
   It 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 approach*:
   It depends on the number of classifiers used in decision making that is one classifier or multiple classifiers.
3. *Incremental or non-incremental approach:*
   It depends 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 useful because it allows us to refine our models/classifiers over time without using previous data.
4. *Active or passive approach* :
   It depends on whether drift detector 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.

B. *Imbalanced Data:*

Unbalanced data, or imbalanced data [6,7] 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 for example in Credit card fraud detection, cancer detection, and financial data. Following is the categories of class imbalance approaches:

1. *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. Especially in medical diagnosis,the positive (minority) class is of more interest than the negative (majority) class.
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.

Our survey depicts that batch learning, incremental learning, ensemble learning, passive learning and imbalanced learning is widely used in combination as a research issue.

II. ENSEMBLE BASED CLASSIFICATION

In ensemble based classification [8] a set of classifiers whose individual predictions are combined in some way to classify new examples. The strategy in ensemble based systems [9] is to create many classifiers, and combine their outputs in such a way that this combination will improve the performance over a single classifier. Studies show that to obtain high accuracy, it is necessary to diversify ensemble members from each other. Components can differ from each other by the data they have been trained on, the attributes they use, or the base learner they have been created. Fig.1 is showing the basic steps of ensemble based classification.

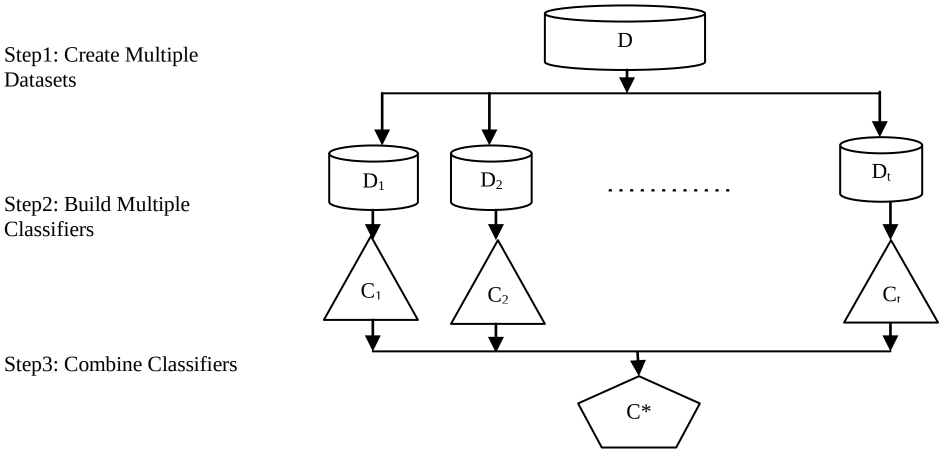

Fig.1. Ensemble based classification

III. REVIEW OF EXISTING ENSEMBLE SYSTEMS

Table 1 provides a brief summary about the sixteen existing ensemble based systems. Table 2 presents the description of each algorithm with their shortcomings.

TABLE I
SUMMARY OF EXISTING ENSEMBLE BASED ALGORITHMS

Sr. No.	Algorithms	Abbreviation	Authors	Year of Publishing	Type of Learner	Purpose
1	SEA[10]	Streaming Ensemble Algorithm	Street and Kim	2001	Batch classifier ensemble systems	To handle concept drift
2	DWM[11,12]	Dynamic Weighted Majority	Kolter and Maloof	2003	Online classifier ensemble systems.	To handle concept drift
3	AddExp [13]	Additive Expert	Kolter and Maloof	2005	Online classifier ensemble systems.	To handle concept drift
4	AWE[14]	Accuracy Weighted	H. Wang et al.	2003	Batch classifier	To handle concept

		Ensemble			ensemble system.	drift
5	AUE1[15]	Accuracy Updated Ensemble	D. Brzezinski, J. Stefanowski	2011	Batch and Incremental Ensemble System.	To handle concept drift
6	AUE2[16]	Accuracy Updated Ensemble2	Dariusz Brzezinski, Jerzy Stefanowski	2014	Batch and Incremental Ensemble System.	To handle concept drift
7	ONSBOOST [17]	On-line Nonstationary Boosting	Pocock et al.	2010	Online and Incremental Ensemble System.	To handle concept drift
8	ACE[18]	Adaptive Classifiers Ensemble	Kyosuke Nishida	2005	Online, Incremental and Drift detector Ensemble System.	To detect and handle concept drift
9	UCB[19]	Uncorrelated Bagging	Jing Gao et al.	2007	Batch and Incremental Ensemble System.	Handles concept drift and imbalanced data
10	SERA[20]	Selectively Recursive Approach	Sheng Chen , Haibo He	2009	Batch and Incremental Ensemble System.	Handles concept drift and imbalanced data
11	MuSERA[21]	Multiple Selectively Recursive Approach	Sheng Chen ,Haibo He	2010	Batch and Incremental Ensemble System.	Handles concept drift and imbalanced data
12	REA[22]	Recursive Ensemble Approach	Sheng Chen ,Haibo He	2011	Batch and Incremental Ensemble System.	Handles concept drift and imbalanced data
13	Learn++. NSE[23-27]	Learn++ with Nonstationary Environments	M. Muhlbaier and R. Polikar	2007	Batch and Incremental Ensemble System.	Handles concept drift and imbalanced data
14	Learn++. CDS[28-30]	Learn++ with Concept Drift and SMOTE[31]	Gregory Ditzler,Robi Polikar	2010	Batch and Incremental Ensemble System.	Handles concept drift and imbalanced data
15	Learn++. NIE[28-30]	Learn++. With Nonstationary & Imbalanced Environments	Gregory Ditzler,Robi Polikar	2010	Batch and Incremental Ensemble System.	Handles concept drift and imbalanced data

TABLE II
COMPARISON OF EXISTING ENSEMBLE BASED ALGORITHMS

Sr. No.	Algorithms	Description	Limitations
1	SEA	First batch based ensemble algorithms for concept drift. It responds to concept drift by replacing an unnecessary classifier with a new one. SEA uses a simple majority vote	Ensemble may not be able to perform in recurring environments
2	DWM	The dynamic weighted majority (DWM) system is based on the weighted majority algorithm, but they added mechanisms to add and remove classifiers dynamically in response to changes in performance. DWM maintains an ensemble of base online classifiers and predicts a class for the current input by using a weighted majority vote from the outputs of the base classifiers.	For DWM, the inclusion/ elimination of classifiers only at every p training examples makes it not possible to handle new drifts during that period .Not consider the occurrence of recurring concepts.
3	AddExp	A new classifier is always added when the prediction of the ensemble as a whole is wrong. The weight of this classifier is equal to the sum of the weights of all the existing ensemble members multiplied by a constant ($0 \leq \gamma \leq 1$). The weight of an ensemble member is multiplied by a constant ($0 \leq \gamma < 1$) at every time step in which it gives a wrong Prediction. Authors proposed two pruning methods: oldest first and weakest first.	However, as it adds a new classifier at every example misclassified by the ensemble, the training can become prohibitively slow when there is a concept drift and the accuracy of the system drops. Not deal with recurrent concepts.

4	WCEA/AWE	The authors propose to train a new classifier on each incoming data block and use that block to evaluate all the existing classifiers in the ensemble. The weight of each classifier is the error rate of a random classifier minus the mean square error of the classifier for the current chunk. The mean square errors of old classifiers are high, and thus the weights of old classifiers are small.	Performance is dependent upon the size of the data chunks. AWE's weighting mechanism hinders classification performance. Poorer accuracy for data streams with slow and gradual drift
5	AUE1	AUE not only selects classifiers, but also updates them according to the current distribution members rather than just adjusting their weights. The combination of classifier selection and updating should make AUE better than AWE in times of stability or gradual drift..	Conditionally updates component classifiers so less accurate for sudden drift.
6	AUE2	Batch and Incremental Ensemble System .AUE2 introduces a new weighting function, does not require cross-validation of the candidate classifier, does not keep a classifier buffer, prunes its base learners, and always updates its components. Classifiers are updated after every chunk, So they can react to gradual drifts.	It can react to sudden drifts and gradually drifts but not for reoccurring concepts.
7	ONSBOOST	A floating search is integrated into on-line boosting that allows the addition of new classifiers and the removal of poorly performing classifiers. Like DWM, ONSBoost uses an update parameter to determine when the algorithm needs to be updated.	Good accuracy at the cost of speed.
8	ACE	Uses a single online classifier, many batch classifiers, a drift detection mechanism, a sliding window, and a long-term buffer. Its weighting method, which combines the outputs of base classifiers, uses the predictive accuracy of each classifier for recent training examples in order to avoid interference between the outputs of classifiers from old concepts and classifiers for new ones. ACE focuses especially on handling recurring concepts	Slowly reconstructing a classifier ensemble with large chunks of examples
9	UCB	It maintains a fixed number of classifiers. Classifiers are generated by using all positive instances (of current batch and previous batches) and smaller randomly drawn samples of the negative instances in the most recent batch of data.	Assumes, no drift in the minority instances.
10	SERA	Rather than using all accumulated minority class data, SERA judiciously selects instances from the accumulated minority data. In SERA a single hypothesis based on current training chunk is maintained for making predictions on test data chunk.	As required access to previous data so not suitable when minority data is nonstationary Not able to track drift in minority instances.
11	MuSERA	In MuSeRA balancing of training chunk is done in the similar way by using mahalanobis distance as similarity measure to accommodate minority samples accumulated from all the previous training chunks. In MuSeRA a hypothesis is built on every training chunk, thus a set of hypothesis is built over time as opposed to SERA which maintains only single hypothesis.	As required access to previous data so not suitable when minority data is nonstationary. Need to retain historical data so not purely incremental approach.
12	REA	In this when next training chunk arrives, it is balanced by adding those positive instances from previous chunks which are nearest neighbours of the positive instances in the current training chunk, then it is used to build a soft typed hypothesis. For every training chunk a new soft typed hypothesis is built.	As required access to previous data so not suitable when minority data is nonstationary. Need to retain historical data so not purely incremental approach.
13	Learn^{++}.NSE	Batch and Incremental Ensemble System .It generates a series of classifiers using batches of examples,by converting the online datastream into a series of chunks of a fixed size. At each time step one new classifier is trained on a batch of new examples, using an example weighting distribution	Handles all types of drift but not considering imbalanced class data.
14	Learn^{++}.CDS	Batch and Incremental Learner. Combination of Learn^{++}.NSE and SMOTE	High Performance at the cost of memory and time.

15	Learn++.NIE	Batch and Incremental Learner. Combination of Learn++.NSE and uses a Bagging based sub-ensemble method to rebalance data.	High Performance at the cost of memory and time.

Fig. 2 provides evaluation in terms of classification accuracy, chunk training time and testing time in centiseconds on electricity pricing dataset. The electricity pricing data is presented in the original Splice-2 paper [32] and has been used as a benchmark for concept drift problems.

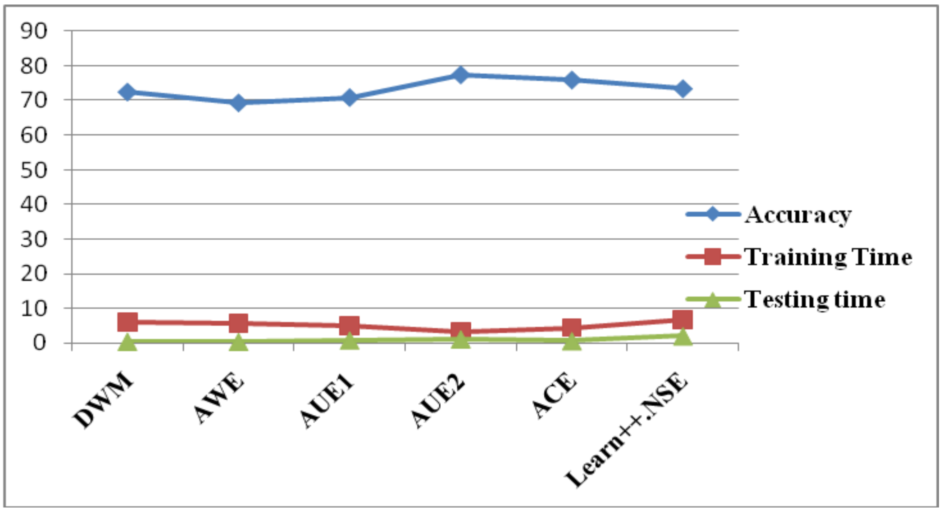

Fig. 2: Comparison of ensemble base dalgorithms to handle concept drift

Fig. 3 provides evaluation in terms of accuracy, f-measure, area under curve and recall on electricity pricing dataset.

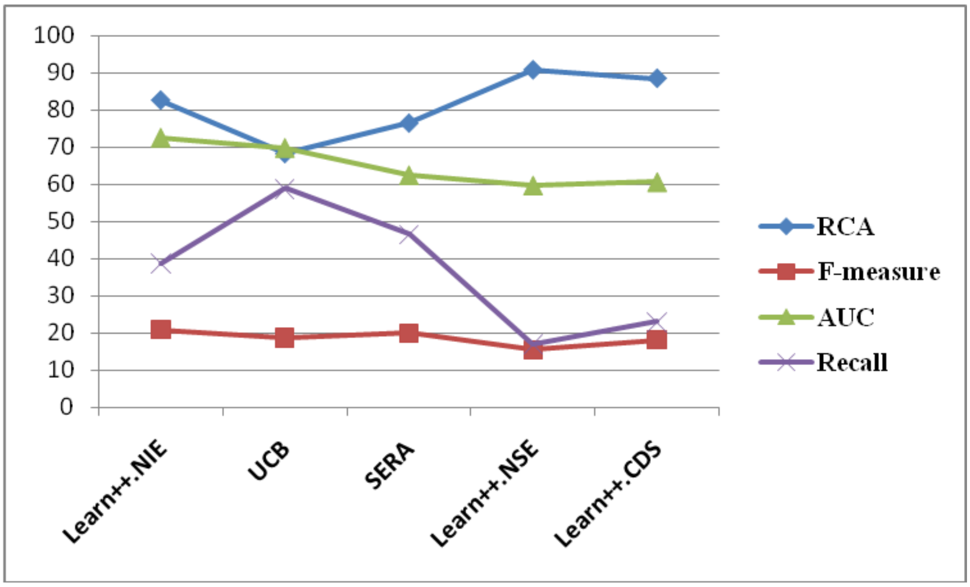

Fig. 3: Comparison of ensemble based algorithms to handle concept drift and class imbalanced problem.

Based on review, we state that Learn++.NSE and AUE2 and ACE are most competing algorithm to handle concept drift whereas Learn++.NIE and Learn++.CDS are competing algorithms for handling concept drift and class imbalance problem simultaneously.

IV. CONCLUSIONS

In this paper, the state of the art on ensemble methodologies to deal with nonstationary data 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] Meenakshi A.Thalor ,Dr.S.T.Patil, “Review of Ensemble Based Classification Algorithms for Nonstationary and Imbalanced Data” ,IOSR Journal of Computer Engineering, e-ISSN: 2278-0661, p-ISSN: 2278-8727,Volume 16, Issue 1, Ver. IX 2014, PP 103-107,Feb 2014.
[3] 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.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] L. Rokach, Ensemble-based Classifiers, Artif. Intell.Rev.,vol. 33, pp. 1–39, 2010
[9] R. Polikar ,Ensemble Based Systems in Decision Making, IEEE Circuits and Systems Magazine, Vol. 6, No. 3, pp. 21-45, 2006
[10] 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.
[11] 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.
[12] 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.
[13] J. Z. Kolter and M. A. Maloof, “Using additive expert ensembles to cope with concept drift,” in Proceedings of the 22nd International Conference on Machine Learninig, 2005, pp. 449–456.
[14] H. Wang, W. Fan, P. S. Yu, and J. Han, “Mining concept-drifting data streams using ensemble classifiers,” in Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 226–235.
[15] D. Brzezinski and J. Stefanowski, “Accuracy updated ensemble for data streams with concept drift,” in Proc. 6th HAIS Int. Conf. Hybrid Artificial Intell. Syst., Part II, 2011, pp. 155–163.
[16] Brzezinski, D.; Stefanowski, J., "Reacting to Different Types of Concept Drift: The Accuracy Updated Ensemble Algorithm," *Neural Networks and Learning Systems, IEEE Transactions on* , vol.25, no.1, pp.81,94, Jan. 2014
[17] A. Pocock, P. Yiapanis, J. Singer, M. Lujan, and G. Brown, “Online nonstationary boosting," in International Workshop on Multiple Classifier Systems, 2010.
[18] K. Nishida and K. Yamauchi, “Adaptive classifiers ensemble system for tracking concept drift," in Proceedings of the Sixth International Conference on Machine Learning and Cyberne tics, 2007, pp. 3607{3612.
[19] 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.
[20] S. Chen and H. He, SERA: Selectively recursive approach towards nonstationaryimbalanced stream data mining,International Joint Conference onNeural Networks (IJCNN 2009), pp. 522-529, Atlanta, GA, 2009.
[21] Sheng Chen; Haibo He; Kang Li; Desai, S., MuSeRA: Multiple Selectively Recursive Approach towards imbalanced stream data mining, Neural Networks (IJCNN), The 2010 International Joint Conference on , vol., no., pp.1,8, 18-23 July 2010S.
[22] 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.
[23] M. Muhlbaier and R. Polikar, An Ensemble Approach for Incremental Learning in Nonstationary Environments, Multiple Classifier Systems, pp. 490-500, 2007.
[24] M. Muhlbaier and Robi Polikar, 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.
[25] Karnick, M.; Ahiskali, M.; Muhlbaier, M.D.; Polikar, R., "Learning concept drift in nonstationary environments using an ensemble of classifiers based approach," Neural Networks, 2008. IJCNN 2008. (IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on , vol., no., pp.3455,3462, 1-8 June 2008
[26] 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.
[27] Elwell R. and Polikar R., Incremental Learning of Concept Drift in Nonstationary Environments, IEEE Transactions on Neural Networks, vol. 22, no. 10, pp. 1517-1531, October 2011.
[28] 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, 2010.
[29] G. Ditzler, R. Polikar, and N. Chawla, An Incremental Learning Algorithm for Nonstationary Environments and Class Imbalance,20th International Conference on Pattern Recognition (ICPR 2010), pp. 2997-3000, Istanbul, Turkey, 2010.

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

[31] 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.

[32] M. Harries, "Splice-2 comparative evaluation: Electricity pricing," The University of South Wales, Tech. Rep., 1999.

