International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Learning of Concept Drift and Multi Class Imbalanced Dataset using Resampling Ensemble Methods

K. Vasantha Kokilam, D. Ponmary Pushpa Latha, D. Joseph Pushpa Raj

Abstract: ***In modern days, very often usage of mobile phones paves way for advanced technologies which includes Internet-of-Things (IoT), wearable technology and big data. As the technology grows, huge volume of data with its complexities also increases rapidly. Flooding of data leads to combat in terms of online class imbalance problem and concept drift. Class imbalance problem is one of the issues in which number of class labels is not balanced and also majority classes are given more importance than the minority class. This type of situations leads to none accurate classification of data. Spam filtering, Fault detection in Engineering industry, Disease diagnosis are few applications where multiclass imbalance with concept drift makes prediction challenging. In this paper, a novel approach of Concept Drift Detector and Resampling Ensemble (CDRE) algorithm was proposed to overcome the problem of concept drift in multi-class. Misclassification occurs sometimes due to imbalance ratio and data distribution. Detailed analysis was done based on different levels of imbalance ratio and data distribution. There is decline in accuracy when multi-class problem suffers from concept drift also. When compared to normal multi-class imbalance problem, class imbalance problem with concept drift is analyzed. Concept Drift Detector and Resampling Ensemble (CDRE) algorithm was implemented to deal multi-class problem with concept drift. CDRE algorithm shows better results in recall, precision, F-measure on an average 85% when compared with algorithm without optimization.***

Keywords: ***Concept Drift, Imbalance Ratio, Multi-class, Data Distribution, Bagging.***

I. INTRODUCTION

Data Mining is a process of learning new facts in extremely large datasets. It plays a vital role in machine learning technique and artificial intelligence. Data streams flowing from these devices may not work well using the traditional approaches and contribute to the emerging paradigm of big data. IoT (Internet of Things) devices are used vastly in healthcare sector for alerting patient's health status online.

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
Mrs. K Vasantha Kokilam*, Department of Information Technology at Karunya Institute of Technology and Sciences, Coimbatore Tamil Nadu, India.
Dr. D. Ponmary Pushpa Latha, M.C.A., MPhil., M.E.,Ph.D.,, associate Professor, Department of Information Technology, Karunya Institute of Technology and Sciences, Coimbatore Tamil Nadu India.
D. Joseph Pushpa Raj , M.E. degree in Computer Science and Engineering and currently, he is working in Francis Xavier Engineering, College, Coimbatore Tamil Nadu India

Nowadays chronic illness like blood pressure, respiratory diseases, heart-related diseases and diabetes affects one in ten persons [1].
These diseases require healthcare services. There are many smart home devices available in the market like heart rate sensing wearables,
activity tracking wearable, smart clothing, blood pressure measuring device, weight detecting device etc.[2]. These devices can work as smart home tele monitoring devices when connected with IoT networks.
Data streams have the following characteristics: dynamically changing, non-stationary, continuous, huge volume of data, arriving continuously and uncountable. The resulting data has to be processed in order to avoid faulty data, erroneous values, worthless data, missing values, concept changes etc. Traditional algorithms support decision making and multidimensional analysis but new data analysis techniques must be introduced to predict the changes in data over time.

Features of online data streams:

1. The online data stream cannot have the whole dataset over time, whereas whole data can be accessed in case of static data set.
2. In online data stream, the data streams drastically change its nature of resultant data called as concept drift.

3. Data arrive continuously in an online fashion at a very high speed. Classic algorithms do not satisfy the classification accuracy of the online dataset. Static data streams can be passed at the moderate rate of speed. The accuracy of static classification is higher than the online dataset when it incorporates the classic algorithm [3].

A. Class Imbalance Learning

Class imbalance learning is one of the types of classification problems. In the University of California, Irvine (UCI) breast cancer dataset [4]; there are two classes such as benign or malignant, sometimes benign classes will be more in number than that of malignant class in the dataset. So, there is difficulty in predicting the results if various types of instance arrive representing malignant class. Few skewed distribution algorithms makes traditional machine learning techniques degrades the performance of learning, mainly when predicting less number representing class label like malignant. Iris dataset which is available in UCI repository is an example for multi-class. The characteristic of iris data set is multivariate as represented in Table.1. The iris dataset consist of three classes and fifty instances for each class. Each class in the dataset refers to different types of iris.

Table 1. Iris Data Set

Dataset Characteristics	*Attribute Characteristics*	*Number of Instances*	*Number of Attributes*	*Associated Tasks*	*Missing Values ?*
Multivariate	*Real*	*150*	*4*	*Classification*	*No*

Class imbalance problem means in a dataset if the number of instances belonging to one class is more than that of instances belonging to another class [5]. More populated class labels are represented as majority class whereas less populated class labels belongs to minority class. Only rare instances appear in case of minority sample, but it is very significant in class imbalance. The following are four types of data mining approaches.

1. Data Level Approach(External Technique)
a. Oversampling (Random or Direct)
b. Undersampling(Random or Direct)
c. Active Sampling
2. Algorithmic approach(Internal Technique)
a. Adjusting the cost
b. Adjusting the decision threshold
3. Cost sensitive approach (Both algorithmic and Data level approaches)
4. Ensemble methods(Multiple Classifier)

All the above-stated approaches have their own pros and cons. The detailed literature study is conducted in this area based on complexity and imbalance level between the two classes. Dataset can be classified as one class, two classes or multi-class problem. Let C represents the class when the value of C is one, it represents one class problem. When the value of C is two then it represents two classes or binary class problem. If the value of C crossed two then, it is called as multi-class problem [6].

Imbalanced datasets are facing the problem in classification when the class labels are not equally distributed. If the class C is equal to two, more number of instances representing class is called positive class and less number of instances representing class is called as negative class.

Several methods are proposed for solving class imbalance problem by providing solutions at data level which includes re-sampling and feature selection. Other algorithms like cost sensitive and single class learning provide solutions at the algorithm level.

B. Concept Drift:

In machine learning, the objective of this study is to predict the class represented as y. In (X,y), X represents attributes and y represents the class label. P(y|X) is denoted in equation (1) and P(X|y) is denoted in equation(2) for all classes y=1..n, where n represents the number of classes[7].

The classification prediction for class y can be represented as

$$P(y|X) = \frac{P(y)P(X|y)}{p(X)} \quad (1)$$

Where, $$P(X) = \sum_{y=1}^{c} P(y)P(X|y) \quad (2)$$

Formula 2 represent misclassification costs. Concept drift refers to changes over time from time point t0 to time point t1 and it can defined as

$$\exists X: Pt_0(X, y) \neq Pt_1(X, y) \quad (3)$$

The conditions in which concept drift occurs refers to

a) Change in P(y|X), is real drift
b) Change in P(X|y), is virtual drift

Because of the above mentioned two conditions change in concept may occur over a period of time.

Concept Drift is classified as two types as stated below.

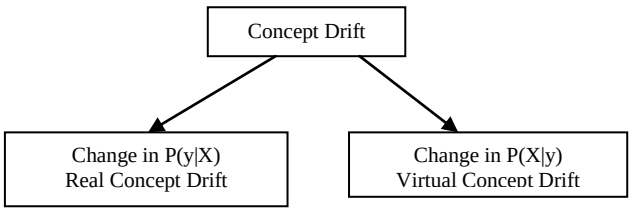


Fig. 1, Different types of Concept Drift

Fig.1. represents two types of concept drifts. Real concept drift refers to changes in the target variable for the given input features, while the attributes of the instances are not changed.

Virtual drift refers to the change in the attributes of the arriving data changes without affecting the target variable. (i.e., p(X) changes without affecting p(y|X).

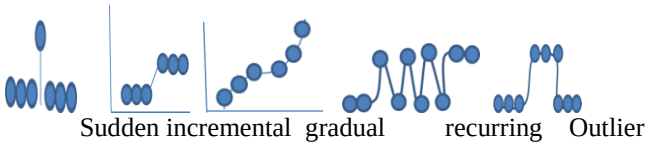


Fig. 2. Changes in Pattern Over time

A different type of concept drift occurs to the changes in the arriving data over time as represented in Fig.2. Below stated are explanations of different types of drifts.

Abrupt/Sudden Concept Drift: Changes from one concept to another.

Incremental Concept Drift: Changes happen slowly and predictions become less accurate in intermediate positions.

Gradually: Changes in concept drift occurs gradually over time.

Recurring Concept drift: Changes occurs for a period of time and then it reaches normal.

Outlier: Anomaly detection takes place in this concept over a period of time.

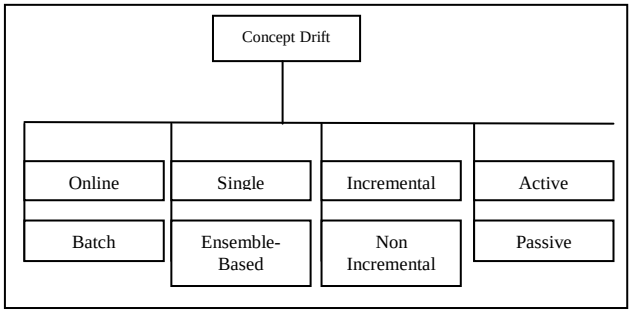


Fig. 3. Concept Drift Algorithms

Concept drift algorithms can be classified according to Fig.3 as either online based approach or batch processing approach by determining static or dynamic training set.

Concept drift solving algorithm can use either single classifier or ensemble based methods.

Single classifiers use only one classifier to solve concept drift whereas ensemble-based classifier uses more than one classifier for making the decision.
Incremental approach uses only current data, whereas in case of non-incremental approach, previously collected data are reused;
It actively checks for concept drift. Actions will be taken only based on the occurrence of drift then it is called active drift approach. There are no unnecessary updates when there is no drift detection.
A passive drift detection algorithm frequently updates a learning model whenever new data arrive.. Difficulties occur due to frequent update.
In summary, the novel contributions and objectives included in this research works are:
1. To understand the process and working of various methodologies to overcome all the permutations and combinations of multi-class imbalance and concept drift
2. To propose CDRE algorithm to tackle the joint problem of multi-class imbalance problem and concept drift.
3. To compare the results with EMUOB and EMOOB, WEOB, CDRE..

II. RELATED WORK

A. Review on Concept Drift

Concept drift algorithms work on the incoming data with sliding window concepts, which results in instances under the window are treated as stationary and a new classifier starts learning the accumulated data present in the present window. The earliest passive batch based instance selection methods [8] are as follows STAGGER [9] and FLORA [10]. Depending on the velocity of the drift present, FLORA algorithms have certain mechanisms which do active drift detection using ensemble approaches wherein, the widening or narrowing of data is based on the data flow. Evaluating based on the most recent data, the classifiers are labeled as relevant (if there is no concept change), irrelevant (if there is concept change) or potentially relevant (concept change can be treated as normal) by FLORA algorithm.
Counter is implemented which denotes classifiers predicting results correctly and A counter is maintained by such classifiers depending upon the number of correctly classified examples and trimming of classifiers done based on the significance of data in the current data window. But this approach results in causing sudden alterations in the system. Active concept drift algorithm include Cumulative Sum (CUSUM), Alippi and Roveri's Just in Time (JIT) [11], Intersection of Confidence Intervals (ICI) [12]. In Cumulative Sum (CUSUM), classification done based on control charts. For drift detection and classifier updating certain measure like entropy, hoeffding bounds are used. Various algorithms like Concept of Very Fast Decision Tree (CVFDT) [13] and Incremental OnLine Information Network (IOLIN) [14] are used for drift detection and classifier updates. While such algorithms are successful in detection of abrupt changes, they are not so successful in handling gradual drift scenarios.
The Early Drift Detection Method (EDDM) [15] address the gradual drift and it is discussed flags concept but it gives warning measures also, by calculating the distance between the classifier errors and when the particular mean values touch the threshold level. Ensemble Systems or Multiple Classifier Systems (MCS) [16] algorithm is implemented for l learning (Non-Stationary Environment) NSE. MCS works on updating of the knowledge base by using addition, removal or updating of classifiers. Ensemble-based algorithms works by adding or deleting classifiers, based on performance of the classifier.
Streaming Ensemble Algorithm (SEA) [17] uses passive drift selection and the ensemble size is fixed. Replacements of oldest member with new the one are done using passive drift detection. Hoeffding tree bagging with Dynamic Weights are used in Dynamic Weighted Majority (DWM) [18] samples.
Voting is an approach which ranks the classifier performance using proximity measure. Based on classifier error due to concept drifting weights are allocated using simple majority voting. The same information cannot be reused for future data sets.
Hybrid approaches are implemented that combine various techniques like sliding windows, active detection and classifier ensembles. ADWIN algorithm is an example for hybrid approach because it combines random forests with entropy and a filter is integrated with an adaptive sliding window.
Massive Online Analysis (MOA), is a framework which integrates ADWIN and other tools for extracting useful information from streams with concept drift. Active ensemble approach include Learn++NSE algorithm which is able to track a variety of drifts like gradual, fast, abrupt and cyclical drift. Based on voting strategy, it can classify best and worst performing classifiers.. Based on the detailed study of the literature there are various works done on concept drift. Our proposed work is to perform optimization on the concept drift algorithm to enhance the performance.

B. Reviews on Class Imbalance in Machine Learning

Dataset can be split into training and testing data set. Real-time data set consist of noise, outliers, missing values etc., Hence there is a need for us to do preprocessing before it does the classification. Preprocessing can be done using sampling method which has the ability to resolve the class imbalance problem by constructing a balance between majority and minority class.
There are two major sampling methods named as undersampling and over-sampling which is used to classify the data. Even though, the undersampling method gives the good results in terms of classification. Sometimes it may lead to remove imperative information. Oversampling may cause overfitting the data by replicating and modifying the information, which may result in the additional computational cost to those who like to use the same [19].
Though oversampling techniques are used to improve the accuracy of minority class, it is not appropriate for the skewed data stream and continuous concept drifting. AdaBoost is an ensemble based learning algorithm uses boosting method to increase the performance of the classifier. Most imbalance learning techniques are intended for the binary class problem. Accuracy can be improved on the minority class when AdaBoost is pooled with random oversampling. Skewed data stream can be solved by clustering the samples but it does not detect drifting. Using under-sampling helps to preserve the useful data but it does not help us in detecting the drift.

One class skewed data streams learning method can be solved by SVM algorithm but it does not support concept drifting.
One class skewed data streams can be solved by RUS (Random Under-Sampling).
It uses single classifier approach for predicting the data streams. It removes majority class at random fashion till an intended class distribution is achieved [20]. SMOTE Boost [21] algorithm integrates boosting concept and data sampling techniques which decrease the time required to construct a model. Tomek links uses under sampling method to reduce the majority class by measuring the distance between two instances.
Bagging Ensemble Variation (BEV) [22] has implemented bagging concept to train both majority and minority class instances. Learn++.UDNC (Unbalanced Data New Class) [23] uses voting the weights of classifier based on misclassification. Many algorithms have given better performance both in data level and algorithm level. Our proposed work is on data streams and their class imbalance combined with concept drift. For this purpose, the above stated traditional algorithm will not solve the purpose. So work has to be focused on class imbalance with concept drift on data streams.

C. Review on Concept drift and Imbalanced Dataset

Many algorithms were proposed for the non-stationary environment. Uncorrelated Bagging (UCB) algorithm trains the classifier with recent majority class instances and it accumulates all recent minority class instances union it and performs classification. Recent minority class instances are observed from recent time steps and are stored separately. But this makes minority class instances to be stationary. This algorithm keeps on storing all the minority class instances for a long time. Over a period of time these minority class instances will become majority. These minority class instances might also go irrelevant in future. So it cannot be used for streaming data. Selectively Recursive Approach (SERA)[24] algorithm stores the minority class instance same like previous algorithm but only recent minority class instances are stored. It can be implemented in two ways:
1) Single classifier is generated for each data set.
2) It uses Bagging concept
Bagging concept increases the minority class instances manually. SERA an ensemble approach uses mahalanobis distance metric to eliminate unrelated instances from the current training set. SERA and UCB both are not suitable for incremental learning as it stores the minority class instances. K-NN is an algorithm used for predicting multi-label classification that works for both concept drift and class imbalance.
The foremost algorithm to detect concept drift is Drift Detection Method for Online Class Imbalance (DDM-OCI) [25] It monitors whether any misclassification is present. If there is any drip, then drifts will be reported. This algorithm works well when there drift in minority class instance and having less performance in case of majority class instances.
Improvement of DDM-OCI was proposed in the name of Linear Four Rates (LFR) approach which uses precision, majority class recall and minority class recall as metrics to identify drifts. If there are any changes found in the four rates of confusion matrix, then it will be concluded that there is a drift. P-AUC (Prequential Area Under Curve) [26,27] is the performance metric, proposed to measure the overall performance in case of online scenarios and Page-Hinkley(PH) [28] test which is used as the indicator to represent the concept drift. One drawback of the approach is it can access only historical data.
Recursive Least Square Perceptron Model (RLSACP) [29] and Online Neural Network for Non-stationary and Imbalanced Data Stream s ONN [30] are single-model approaches used for classifying online data streams Recursive least square filter is used to identity concept drift and class imbalance RLS error model(forgetting factor) is implanted for handling class imbalance. Based on the imbalance rate, weights are updated to find out better performing classifier.
Ensemble of Subset Online Sequential Extreme Learning Machine ESOS-ELM [31] is an ensemble approach. It maintains many numbers of Online Sequential Extreme Learning Machines (OS-ELM) to identify class imbalance. Each OS-ELM [32] is trained with more or less equal number of minority and majority class instances in order to handle concept drift and resampling is used to overcome class imbalance. Voting strategy is adopted to rate the performance of the classifier. ELM- store is a module which is included that can handle recurring drifts efficiently. ELM-store also maintains all old information in WELM. If a concept drift is identified, immediately a new WELM is built and kept in ELM-store.. Voting based weighted online sequential extreme Learning Machine (VWOS-ELM) [33] is the foremost sequential classifier that can handle multi class imbalance problem. EMUOB and EMOOB algorithms [34] work well on multi-class imbalance problem. WEOB calculate average gmean which gives the better result than EMUOB and EMOOB. But when multi-class imbalance problem suffers from concept drift the values of the performance measures decreases.
The algorithms discussed above works both on class imbalance and concept drift. Our proposed work is to optimize the concept drift algorithm to increase the accurate prediction of classifier by introducing CDRE algorithm to solve concept drift and multi class imbalance problem simultaneously.

III PROBLEM FORMULATIONS OR METHODOLOGY

A. Evaluation metrics:

The performance measures which are used for comparisons of classification algorithms are accuracy(proportion of positive results), sensitivity(how well the algorithm classifies true positive instances as positive) and specificity(refers the probability of diagnosing class labels without false positive results). An illustrious confusion matrix is gained for calculating the three measures. Table.2 represents measures of the confusion matrix. A confusion matrix represented in Table.3 determines the classification results. Specificity, accuracy, sensitivity represented in Table.4 is important measures that determine the performance of the algorithm.

Table 2. Confusion Matrix Measures

S.No	True/False	Description
1.	True positive (TP)	Number of positive instances in dataset that are correctly predicted.
2.	False negative (FN)	number of positive instances in dataset that are wrongly predicted.
3.	False positive (FP)	number of negative instances that are wrongly predicted as positive instances.
4.	True negative (TN)	number of negative instances that are correctly predicted.

Table 3. Confusion Matrix

ACTUAL	PREDICTED	
	POSITIVE	NEGATIVE
POSITIVE	True Positive	False Negative
NEGATIVE	False Positive	True Negative

More matrices which are used in classification are as follows:

Time: Time required for a learning model to complete the training and predicting the result. It is calculated in seconds.

Kappa Statistic: It is a measure of agreement between the classifications and the true classes.

ROC Curves: Receiver Operating Characteristics (ROC) curves are similar to lift chart, which is used to show the compromise between true alarm rate and false alarm rate over the noisy channel. Lift charts have X-axis, which shows false positive percentages in samples rather than the sample size. Y-axis represents the percentage of true positives in the sample, rather than absolute number.

Precision = TP/ (TP+FP).

Recall = TP/(TP+FN).

Fmeasure is a metric that combines the precision and recall values.

Table 4. Evaluation Parameters

S.No	Classifier Measures	Calculation
1.	Sensitivity(True Positive Rate)	True Positive(TP)/(True Positive(TP) + False Negative(FN)
2.	Specificity(True Negative Rate)	True Negative(TN)/(True Negative(TN) +False Positive (FP)
3.	Accuracy	True Positive(TP) + True Negative(TN)/(True Positive(TP) + False Positive(FP) + True Negative(TN) + False Negative(FN)

IV THE PROPOSED METHOD

SEA generator introduced by Street Kim is used in this work. This dataset is a synthetic data stream generator which generates various range of concept drifts. It contains 60,000 instances with 3 columns representing attributes and 3 classes representing multi-classes. First and second column's numeric value ranges between 0 and 10. 10% of noise is also included in the dataset.

In the below Fig 4., Concept Drift Detector and Resampling Ensemble (CDRE) algorithm is given to handle concept drift and multi-class imbalance problem simultaneously.

Data stream $(x_1, y_1)..(x_n, y_n)$ at current training dataset is stored in $D^{(t)}$, where x_t represents instance and y_t is prediction variable. Minority samples belonging to previous $D^{(t)}$ are stored for future resampling. If the number of minority examples captured during the current timestamp is not sufficient then stored minority examples can be used to rebalance. Number of Classes in the current training set is stored in i. Number of classes observed during previous blocks is stored in m. If i is not equal to m then there is either arrival of new class or deletion of class has occurred. If there is any concept drift, then the sample will be captured in War_{Win}. If none of the examples arrived is related to the samples stored in War_{Win}, then it alerts with a warning signal. If more number of examples arrived is related to the stored examples then the training examples will be transferred to the Det_{Win}. The ensemble (¢) performs classification. The Worse performing classifier will be replaced with good one.

Input: Current Training dataset $D^{(t)}$=($x_1, y_1)..(x_n, y_n)$where x_t represents instance and y_t is prediction variable.
Current Class size ($\omega_i^t = \omega_1^t..\omega_n^t$)
Number of classifier(Ensemble Size :¢)
Ψ : Imbalance ratio for current block

Output: To Predict Y
I =Distinct(y_t) ;// *i represents current number of classes*
Divide the instances based on number of ω_i^t

//M represents previous number of classes
if m==i;
{
For i=1,2,…n do
{
1. Calculate Ψ for ω_i^t during current timestamp based on size threshold ϑ_1 and performance threshold ϑ_2 }
2. Based on Ψ, the algorithm either goes to classification or triggers EMUOB and EMOOB algorithm.
}}
Else
3. Extract the new class and place in it War_{Win}.
4. If number of samples reaches the threshold it get transferred to Det_{Win}.
5. else it alerts warning signal.
6. Build candidate classifier Γ
7compute Ω_{ht} for each classifier
8. if the performance of the classifier is below threshold
9. Replace the worst peforming classifier with new member
10. output the label y

Fig. 4 Concept Drift Detector and Resampling Ensemble (CDRE) to handle concept drift and multi-class imbalance problem.

ω_i^t represents current class size. Based on prelimnary experiments, size threshold ϑ_1=0.6 and performance threshold ϑ_2=0. Size and Performance threshold is set in order to avoid unnecessary balancing.

Recent minority class instance are stored in order to avoid imbalance. Based on majority voting strategy ensemble members are combined to form an ensemble classifier. If any ensemble member's performance reaches below threshold then that particular classifier will be eliminated in order to provide good performance.

A. Performance measures for various data Distribution

Table 5. Comparision of Precision values among EMOOB and EMUOB, WEOB, CDRE

Dataset	Imbalance Ratio	Data Distribution	EMOOB and EMUOB	WEOB	CDRE
Sea Dataset	10	Safe	0.6090	0.6355	0.8523
	20		0.5999	0.6320	0.8616
	30		0.6097	0.6645	0.8613
	40		0.6337	0.6807	0.8643
	10	Borderline	0.6081	0.6431	0.8657
	20		0.6042	0.6505	0.8659
	30		0.6116	0.6620	0.8580
	40		0.6402	0.6611	0.8631
	10	Outlier	0.6051	0.6384	0.8647
	20		0.6022	0.6384	0.8526
	30		0.6265	0.6501	0.8681
	40		0.6392	0.6589	0.8451

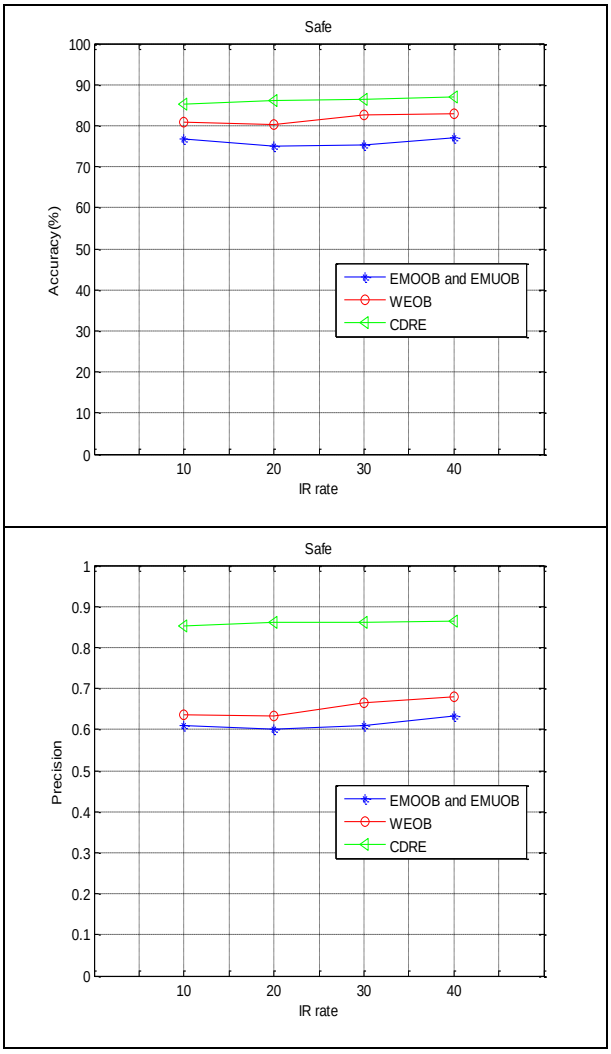
Table 6. Comparision of F-measure values among EMOOB and EMUOB, WEOB, CDRE

Dataset	Imbalance Ratio	Data Distribution	EMOOB and EMUOB	WEOB	CDRE
Sea Dataset	10	Safe	0.6823	0.7149	0.8522
	20		0.6663	0.7045	0.8619
	30		0.6727	0.7370	0.8620
	40		0.6942	0.7485	0.8691
	10	Borderline	0.6789	0.7221	0.8657
	20		0.6711	0.7267	0.8660
	30		0.6729	0.7323	0.8596
	40		0.7049	0.7272	0.8653
	10	Outlier	0.6753	0.7176	0.8645
	20		0.6698	0.7154	0.8531
	30		0.6945	0.7233	0.8699
	40		0.7035	0.7239	0.8482

Table 7. Comparison of Recall values among EMOOB and EMUOB, WEOB, CDRE

Dataset	Imbalance Ratio	Data Distribution	EMOOB and EMUOB	WEOB	CDRE
Sea Dataset	10	Safe	0.7758	0.8168	0.8524
	20		0.7494	0.7958	0.8622
	30		0.7500	0.8274	0.8628
	40		0.7685	0.8318	0.8729
	10	Borderline	0.7685	0.6432	0.8658
	20		0.7548	0.6505	0.8661
	30		0.7480	0.6620	0.8613
	40		0.7841	0.6610	0.8675
	10	Outlier	0.6051	0.6383	0.8644
	20		0.6022	0.6384	0.8536
	30		0.6265	0.6501	0.8718
	40		0.6391	0.6588	0.8511

B. Graphical representation of various performance measures



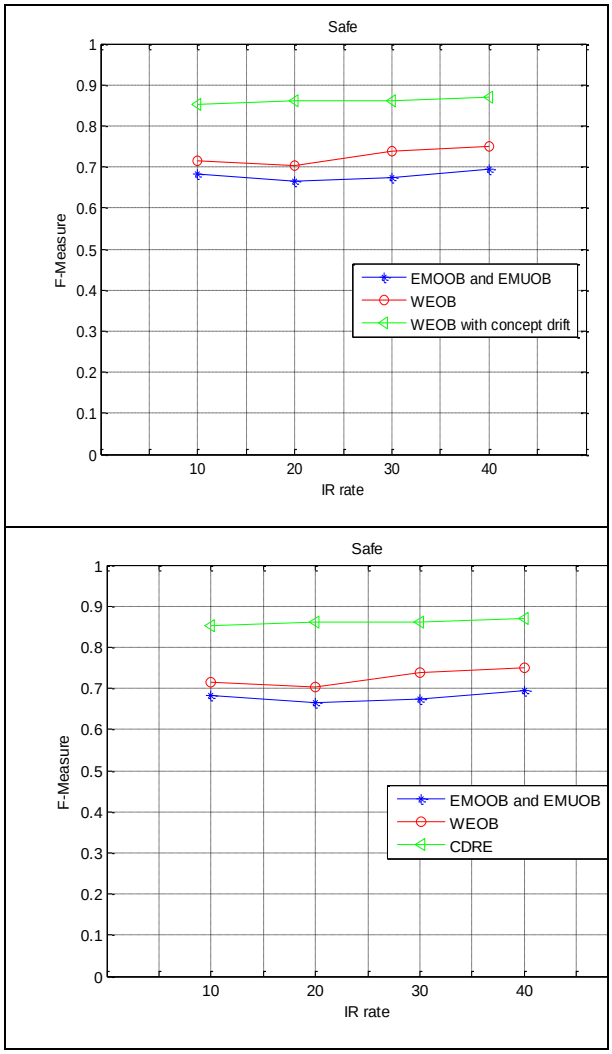
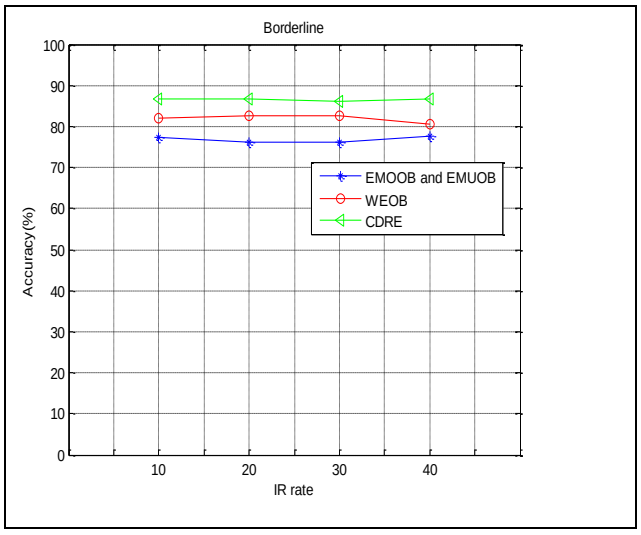


Fig. 5. Graph Representing Accuracy, Precision, F-measure, Recall in Safe Dataset



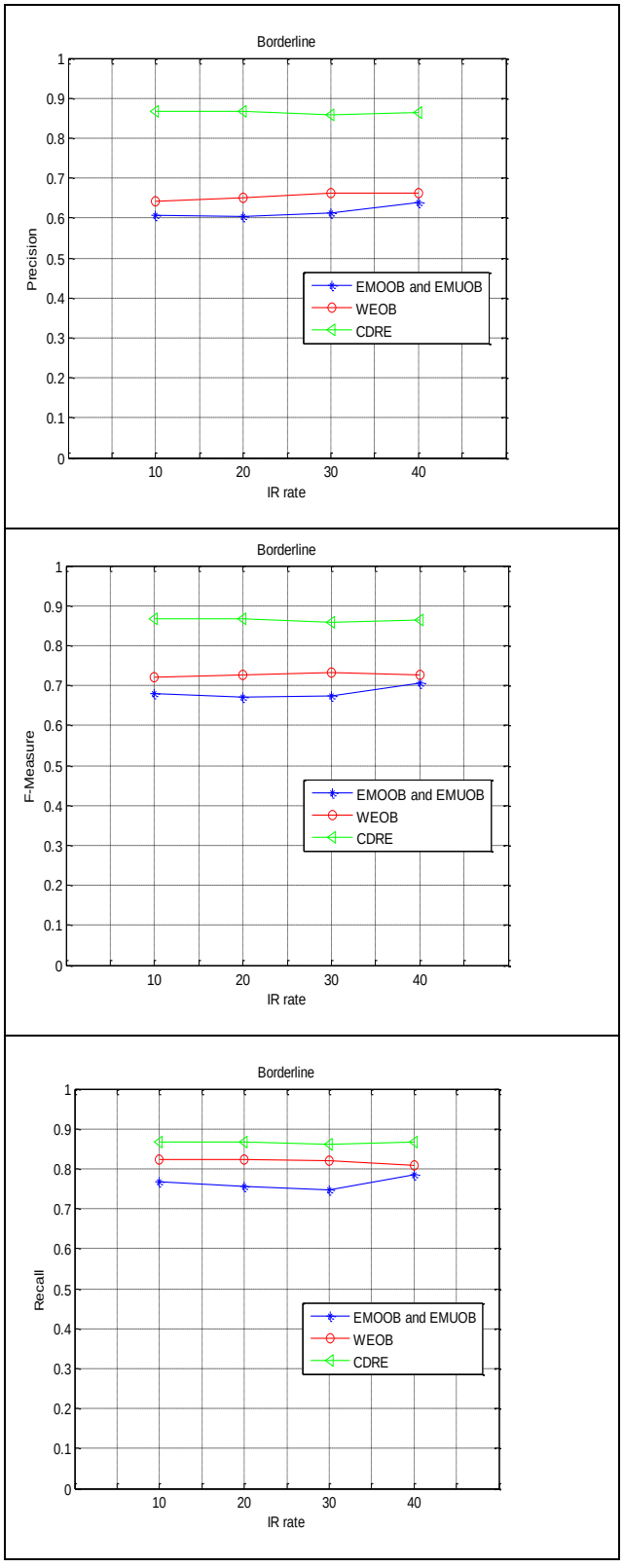


Fig. 6. Graph Representing Accuracy, Precision, F-measure, Recall in Borderline Dataset

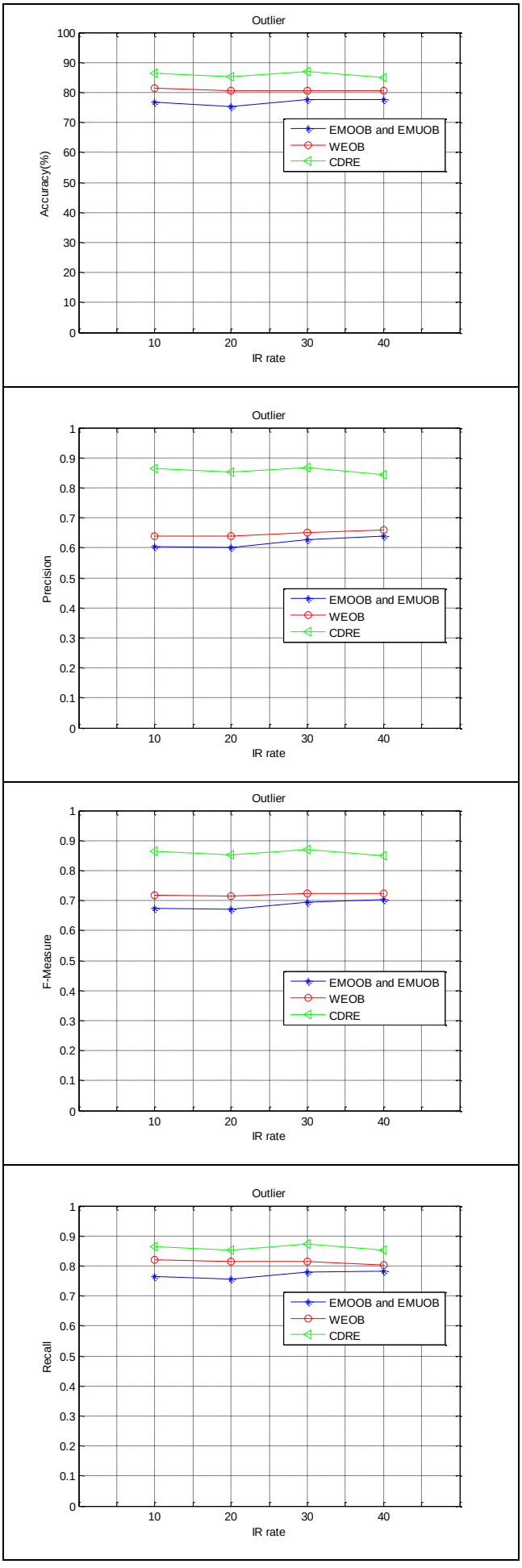


Fig.. 7. Graph Representing Accuracy, Precision, F-measure, Recall in Outlier Dataset

V RESULTS, ANALYSIS AND DISCUSSIONS

First, resampling setting strategies are tested in EMOOB and EMUOB. The performance of the algorithm is improved by calculating adaptive weights as WMOB. Further Concept drift was tested in case of artificial dataset i.e SEA dataset. According to Table 5, 6 and 7 CDRE performs well and provide accuracy on the average of above 85% when EMOOB and EMUOB provide only below that. During learning in non-stationary environment imbalance ratio makes poor recall value and decrease the performance of the classifier. In this work, imbalance ratios are simulated with different levels like 10,20,30,40. Proposed algorithm provides consistent results in terms above mentioned imbalance ratio levels. In all the cases of concept drift with optimization gives consistent results which are clearly visible in Fig 5, 6 and 7. And also based on the results it is proved that though the dataset suffers from class distribution and imbalance ratio proposed algorithm does not bring changes in the classification accuracy.

CONCLUSION

Learning Class Imbalance Problem with concept drift is a challenging in machine learning. In this research work, review of joint problem of concept drift and class imbalance is analyzed. Only few works of literature have addressed these types of problem. When multi class imbalanced dataset do not suffer from concept drift EMUOB and EMOOB or WMOB performs well, but when there is drift performance of the classifier degrades and leads to wrong predictions. Our proposed algorithm of Concept Drift Detector and Resampling Ensemble (CDRE) shows improvement and it provides the better result when there is drift in multi class imbalance problem. Though Analysis was performed on different imbalance ratio and data distribution, CDRE stood good showing more than 80% accuracy in the prediction. The future work will be extended to different types of drift and will have the deep understanding in analyzing artificial and real datasets.

ACKNOWLEDGMENT

I hereby thank and acknowledge the support and guidance received from my guide Dr. D. Ponmary Pushpa Latha., Associate Professor, Department of Information and Technology, Karunya University, Coimbatore.

REFERENCES

1. G. Paré,, K. Moqadem, G. Pineau,, C. St-Hilaire, "Clinical effects of home telemonitoring in the context of diabetes, asthma, heart failure and hypertension: a systematic review," J Med Internet Research, 2010.
2. M. Sedlmayr, H. Prokosch, U. Münch, "Towards smart environments using smart objects," Paper presented at. Studies in Health Technology and Informatics, Vol. 169, pp. 315-319.
3. C. Aggarwal, " On change diagnosis in evolving data streams," IEEE Trans. on Knowl. and Data Eng. 17, pp.587–600, 2011.

4. M. Abdel-Zaher, Ahmed, and Ayman M. Eldeib, "Breast cancer classification using deep belief networks," Expert Systems with Applications Vol. 46, pp. 139-144, 2016.
5. H. He, and E.A. Garcia, "Learning from imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, 2009.
6. Shuo Wang., "Multi-class Imbalance Problems: Analysis and Potential Solutions", IEEE Transactions on systems, MAN and cybernetics, Vol 22, 2012.
7. Gama, João, et al., "A survey on concept drift adaptation." ACM Computing Surveys (CSUR) 46.4 : 44, 2014.
8. Žliobaitė, Indrė, Marcin Budka, and Frederic Stahl., "Towards cost-sensitive adaptation: When is it worth updating your predictive model?," Neurocomputing, pp. 240-249, 2015.
9. Jeffrey, C., Schlimmer , Richard H. Granger, Jr., Incremental Learning from Noisy Data, Machine Learning, Vol no1. pp. 317-354, 1986.
10. Gerhard Widmer, Miroslav Kubat, Learning in the presence of Concept drift and hidden contexts, Machine Learning, vol.23, pp. 69-101, 1996.
11. C. Alippi, and M. Roveri, "Just-in-Time Adaptive Classifiers – Part I: Detecting Nonstationary Changes," IEEE Trans. Neural Networks, vol.19,no.7, pp. 1145-1153, 2008.
12. C. Alippi, G. Boracchi and M. Roveri, "Change Detection Tests Using the ICI Rule", Proc.World Congress Computational Intelligence, pp. 1190-1196, 2010.
13. G. Hulten, L. Spencer and P. Domingos, "Mining Time-Changing Data Streams," Proc. Conference on Knowledge Discovery in Data, pp. 97-106, 2001.
14. L. Cohen, G. Avrahami, M. Last and A. Kandel, "Inf Fuzzy Algorithms for Mining Dynamic Data Streams," Applied Soft Computing, Vol. 8, no. 4, pp. 1283-1294, 2008.
15. M. Baena-Garcia, J.del Campo-Avila, Fidalgo, R., ABifet, R. Gavald and R. Bueno-Morales, "Early Drift Detection Method," Proc. ECML PKDD Workshop Knowledge Discovery from Data Streams, pp. 77-86, 2006.
16. L. I. Kuncheva, "Classifier Ensembles for Detecting Concept Change in Streaming Data: Overview and Perspectives," Proc. European Conference on Artificial Intelligence (ECAI), (pp. 5-10), 2008.
17. W. N. Street and Y. Kim Y, "A Streaming Ensemble Algorithm (SEA) for Large-Scale Classification," Proc. International Conference on Knowledge Discovery and Data Mining, pp. 377-382, 2001.
18. J. Z. Kolter and M. A. Maloof, "Dynamic Weighted Majority: An Ensemble Method for Drifting Concepts," J. Machine Learning Research, vol. 8, pp. 2755-2790, 2007.
19. E. S. Xioufis, M. Spiliopoulou, G. Tsoumakas and I. Vlahavas, "Dealing with Concept Drift and Class Imbalance in Multi-Label Stream Classification," Proc. Int'l Joint Conf. Artificial Intelligence, 2011.
20. V. Garcia, J. S. Sanchez and R. A. Mollineda, " On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," Knowledge-Based Systems, vol. 25, pp. 13-21, 2012.
21. N. V. Chawla, A. Lazarevic, L. O. Hall and K. W. Bowyer, "SMOTEBoost: Improving Prediction of the Minority Class in Boosting," Proc. Seventh European Conerence on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 107-119, 2003.
22. C. Li, "Classifying Imbalanced Data using a Bagging Ensemble Variation (BEV)," Proc. ACM Southeast Regional Conference, pp. 203-208, 2007.
23. M. D. Muhlbaier, A. Topalis and R. Polikar, "Learn++.NC: Combining Ensemble of Classifiers With Dynamically Weighted Consult-and-Vote for Efficient Incremental Learning of New Classes," IEEE Trans. Neural Networks, vol. 20, no. 1, pp. 152-168, 2009
24. H. Wang and Z. Abraham, Z, "Concept drift detection for streaming data, in International Joint Conference of Neural Networks, pp. 1–9, 2015.
25. S. Wang, L. L. Minku, D. Ghezzi, D. Caltabiano, P. Tino, and X. Yao, "Concept drift detection for online class imbalance learning," in International Joint Conference on Neural Networks (IJCNN '13), pp. 1–8, 2013.
26. D. Brzezinski and J. Stefanowski, "Prequential auc for classifier evaluation and drift detection in evolving data streams," New Frontiers in Mining Complex Patterns, Vol. 8983, pp. 87–101, 2015.
27. D. Brzezinski J. Stefanowski, "Prequential auc: properties of the area under the roc curve for data streams with concept drift," Knowledge and Information Systems, pp. 1–32, 2017.
28. E. S. Page, "Continuous inspection schemes," Biometrika, Vol. 41, no. 1, pp. 100–115, 1954.
29. A. Ghazikhani, R. Monsefi and H. S. Yazdi, "Recursive least square perceptron model for non-stationary and imbalanced data stream classification," Evolving Systems, Vol. 4, no. 2, pp. 119–131, 2013.
30. A. Ghazikhani, R. Monsefi, and H.S. Yazdi, "Online neural network model for non-stationary and imbalanced data stream classification," International Journal of Machine Learning and Cybernetics, Vol. 5, no. 1, pp. 51–62, 2014.
31. B. Mirza, Z. Lin and N. Liu, "Ensemble of subset online sequential extreme learning machine for class imbalance and concept drift," Neurocomputing, Vol. 149, pp. 316–329, 2015.
32. N. ying Liang, G. Bin Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," IEEE Transactions on Neural Networks, Vol. 17, no. 6, pp. 1411 – 1423, 2006.
33. B. Mirza, Z. Lin, and K. A. Toh, "Weighted online sequential extreme learning machine for class imbalance learning," Neural Processing Letters, Vol. 38, no. 3, pp. 465–486, 2013.
34. K. V., Kokilam, K., & D., P., P., Latha, 2018, Improved Ensemble Methods to Solve Multi-class Imbalance Problem Using Adaptive Weights. In Proceedings of International Conference on Computational Intelligence and Data Engineering, pp. 333-343, Springer, Singapore, 2018.

AUTHORS PROFILE

Mrs. K Vasantha Kokilam, MCA., PGDBA., M.Phil., SLET qualified is a research scholar pursuing Ph.D in Department of Information Technology at Karunya Institute of Technology and Sciences. Her area of interests includes Data Mining, Big Data, Cloud Computing. She has published papers in international journals and presented papers at various seminars and wishes to contribute in computing arena. She can be reached at t.kokilam@gmail.com

Dr. D. Ponmary Pushpa Latha, M.C.A., MPhil., M.E.,Ph.D.,, has completed her M.E. degree in Anna University and Doctorate in Computer Application. She has completed her Ph.D. in Data Mining and Bioinformatics. Her area of interest is Data Mining, Big data and Bioinformatics. Presently, she is working as associate Professor, Department of Information Technology, Karunya Institute of Technology and Sciences, Coimbatore-641114. She can be reached on ponmarymca@gmail.com or 9488460107

D. Joseph Pushpa Raj , has completed his M.E. degree in Computer Science and Engineering and his area of interest is Image Processing. Currently, he is working in Francis Xavier Engineering College