Integrated Intelligent Research (IIR)
International Journal of Data Mining Techniques and Applications
Volume: 03 Issue: 02 December 2013, Page No.69-80
ISSN: 2278-2419

A Novel Class Imbalance Approach using Cluster Disjuncts

Syed Ziaur Rahman [1] G. Samuel Vara Prasad Raju [2]
[1] Research Scholar, Department of CS&SE, University College of Engineering, Andhra University, Visakhapatnam, A.P, India.
[2] Professor, Department of CS&SE, University College of Engineering, Andhra University, Visakhapatnam, A.P, India.
E-Mail- sdzrahman@gmail.com, gsvpraju2012@gmail.com

Abstract- In Data mining and Knowledge Discovery hidden and valuable knowledge from the data sources is discovered. The traditional algorithms used for knowledge discovery are bottle necked due to wide range of data sources availability. Class imbalance is a one of the problem arises due to data source which provide unequal class i.e. examples of one class in a training data set vastly outnumber examples of the other class(es). Researchers have rigorously studied several techniques to alleviate the problem of class imbalance, including resampling algorithms, and feature selection approaches to this problem. In this paper, we present a new hybrid frame work dubbed as Cluster Disjunct Minority Oversampling Technique (CDMOTE) and Naïve Bayes for Cluster Disjunct (NBCD) for learning from skewed training data. These algorithms provide a simpler and faster alternative by using cluster disjunct concept. We conduct experiments using fifteen UCI data sets from various application domains using four algorithms for comparison on six evaluation metrics. The empirical study suggests that CDMOTE and NBCD have been believed to be effective in addressing the class imbalance problem.

Keywords — Classification, class imbalance, cluster disjunct, CDMOTE, NBCD.

I. INTRODUCTION

A dataset is class imbalanced if the classification categories are not approximately equally represented. The level of imbalance (ratio of size of the majority class to minority class) can be as huge as 1:99 [1]. It is noteworthy that class imbalance is emerging as an important issue in designing classifiers [2], [3], [4]. Furthermore, the class with the lowest number of instances is usually the class of interest from the point of view of the learning task [5]. This problem is of great interest because it turns up in many real-world classification problems, such as remote-sensing [6], pollution detection [7], risk management [8], fraud detection [9], and especially medical diagnosis [10]–[13].

There exist techniques to develop better performing classifiers with imbalanced datasets, which are generally called Class Imbalance Learning (CIL) methods. These methods can be broadly divided into two categories, namely, external methods and internal methods. External methods involve preprocessing of training datasets in order to make them balanced, while internal methods deal with modifications of the learning algorithms in order to reduce their sensitiveness to class imbalance [14]. The main advantage of external methods as previously pointed out, is that they are independent of the underlying classifier. Whenever a class in a classification task is under represented (i.e., has a lower prior probability) compared to other classes, we consider the data as imbalanced [15], [16]. The main problem in imbalanced data is that the majority classes that are represented by large numbers of patterns rule the classifier decision boundaries at the expense of the minority classes that are represented by small numbers of patterns. This leads to high and low accuracies in classifying the majority and minority classes, respectively, which do not necessarily reflect the true difficulty in classifying these classes. Most common solutions to this problem balance the number of patterns in the minority or majority classes. The proposed framework which is shown in Figure 1 addresses the above said issues for class imbalance datasets.

Resampling techniques can be categorized into three groups. Undersampling methods, which create a subset of the original data-set by eliminating instances (usually majority class instances); oversampling methods, which create a superset of the original data-set by replicating some instances or creating new instances from existing ones; and finally, hybrids methods that combine both sampling methods. Among these categories, there exist several different proposals; from this point, we only center our attention in those that have been used in under sampling. Either way, balancing the data has been found to alleviate the problem of imbalanced data and enhance accuracy [15], [16], [17]. Data balancing is performed by, e.g., oversampling patterns of minority classes either randomly or from areas close to the decision boundaries. Interestingly, random oversampling is found comparable to more sophisticated oversampling methods [17]. Alternatively, undersampling is performed on majority classes either randomly or from areas far away from the decision boundaries. We note that random undersampling may remove significant patterns and random oversampling may lead to overfitting, so random sampling should be performed with care. We also note that, usually, oversampling of minority classes is more accurate than under sampling of majority classes [17]. In this paper, we are laying more stress to propose an external class imbalance learning method for solving the class imbalance problem.

This paper is organized as follows. Section II presets the problem of cluster disjucts. Section III briefly reviews the Data Balancing problems and its measures and in Section IV, we discuss the proposed method of CDMOTE (Cluster Disjunct Minority Oversampling technique) and NBCD (Naïve Bayes for Cluster Disjunct) for CIL. Section V presents the imbalanced datasets used to validate the proposed method, while In Section VI, we present the experimental setting and In Section VII discuss, in detail, the classification results obtained by the proposed method and compare them with the results obtained by different existing methods and finally, in Section VIII we conclude the paper.

II. PROBLEM OF CLUSTER DISJUNCT

In Class Imbalance learning, the numbers of instances in the majority class are outnumbered to the number of instances in the minority class. Furthermore, the minority concept may additionally contain a sub concept with limited instances, amounting to diverging degrees of classification difficulty [35-36]. This, in fact, is the result of another form of imbalance, a within-class imbalance, which concerns itself with the distribution of representative data for sub concepts within a class [37-39].

The existence of within-class imbalances is closely intertwined with the problem of small disjuncts, which has been shown to greatly depreciate classification performance [37-40]. Briefly, the problem of small disjuncts can be understood as follows: A classifier will attempt to learn a concept by creating multiple disjunct rules that describe the main concept [35-36], [40]. In the case of homogeneous concepts, the classifier will generally create large disjuncts, i.e., rules that cover a large portion (cluster) of examples pertaining to the main concept. However, in the case of heterogeneous concepts, small disjuncts, i.e., rules that cover a small cluster of examples pertaining to the main concept, arise as a direct result of underrepresented sub concepts [35-36], [40]. Moreover, since classifiers attempt to learn both majority and minority a concept, the problem of small disjuncts is not only restricted to the minority concept. On the contrary, small disjuncts of the majority class can arise from noisy misclassified minority class examples or underrepresented subconcepts. However, because of the vast representation of majority class data, this occurrence is infrequent. A more common scenario is that noise may influence disjuncts in the minority class. In this case, the validity of the clusters corresponding to the small disjuncts becomes an important issue, i.e., whether these examples represent an actual subconcept or are merely attributed to noise. To solve the above problem of cluster disjuncts we propose the method cluster disjunct minority oversampling technique for class imbalance learning.

III. LITERATURE REVIEW

The different imbalance data learning approaches are as follows [43]:

Imbalanced Data learning Approaches

❖SAMPLING METHODS
❖
- ✓ BASIC SAMPLING METHODS
- ➢ Under-Sampling
- ➢ Over-Sampling
- ✓ ADVANCED SAMPLING METHODS
- ➢ Tomek Link
- ➢ The SMOTE approach
- ➢ Borderline-SMOTE
- ➢ One-Sided Selection OSS
- ➢ Neighbourhood Cleaning Rule (NCL)
- ➢ Bootstrap-based Over-sampling (BootOS)

❖ ENSEMBLE LEARNING METHODS
- ✓ BAGGING
- ➢ Asymmetric bagging, SMOTE Bagging
- ➢ Over Bagging, Under Bagging
- ➢ Roughly balanced bagging
- ➢ Lazy Bagging
- ➢ Random features selection
- ✓ BOOSTING
- ➢ Adaboost
- ➢ SMOTEBoost
- ➢ DataBoost-IM
- ✓ RANDOM FORESTS
- ➢ Balanced Random Forest BRF
- ➢ Weighted Random Forest WRF

❖ COST-SENSITIVE LEARNING
- ✓ Direct cost-sensitive learning methods
- ✓ Methods for cost-sensitive meta-learning
- ✓ Cost-sensitive meta-learning
- ✓ Thresholding methods
- ✓ MetCost
- ✓ Cost-sensitive meta-learning sampling methods

❖ FEATURE SELECTION METHODS
- ✓ Warpper
- ✓ PREE (Prediction Risk based feature selection for Easy Ensemble)

❖ ALGORITHMS MODIFICATION
- ✓ Proposal for new splitting criteria DKM
- ✓ Adjusting the distribution reference in the tree
- ✓ Offset Entropy

In this section, we first review the major research about clustering in class imbalance learning and explain why we choose undersampling as our technique in this paper. Siti Khadijah Mohamad *et al.* [18] have conducted a review to look into how the data mining was tackled by previous scholars and the latest trends on data mining in educational research. Hongzhou Sha *et al.* [19] have proposed a method named EPLogCleaner that can filter out plenty of irrelevant items based on the common prefix of their URLs. M.S.B. PhridviRaj *et al.* [20] have proposed an algorithm for finding frequent patterns from data streams by performs only one time scan of the database initially and uses the information to find frequent patterns using frequent pattern generation tree. Chumphol Bunkhumpornpat *et al.* [21] have a new over-sampling technique called DBSMOTE is proposed. DBSMOTE technique relies on a density-based notion of clusters and is designed to oversample an arbitrarily shaped cluster discovered by DBSCAN. DBSMOTE generates synthetic instances along a shortest path from each positive instance to a pseudo centroid of a minority-class cluster. Matías Di Martino *et al.* [22] have presented a new classifier developed specially for imbalanced problems, where maximum F-measure instead of maximum accuracy guide the classifier design.

IV. Garcia *et al.* [23] have investigated the influence of both the imbalance ratio and the classifier on the performance of several resampling strategies to deal with imbalanced data sets. The study focuses on evaluating how learning is affected when different resampling algorithms transform the originally imbalanced data into artificially balanced class distributions. Table 2 presents recent algorithmic advances in class imbalance learning available in the literature. Obviously, there are many other algorithms which are not included in this table. A profound comparison of the above algorithms and many others can be gathered from the references list.

Table 1
Recent advances in Class Imbalance Learning

ALGORITHM	DESCRIPTION	REFERENECE
DCEID	Combining ensemble learning with cost-sensitive learning.	[37]
RUSBoost	A new hybrid sampling/boosting Algorithm.	[40]
CO2RBFN	A evolutionary cooperative–competitive model for the design of radial-basis function networks which uses both radial-basis function and the evolutionary cooperative-competitive technique.	[42]
Improved FRBCSs	Adapt the 2-tuples based genetic tuning approach to classification problems showing the good synergy between this method and some FRBCSs.	[45]
BSVMs	A model assessment of the interplay between various classification decisions using probability, corresponding decision costs, and quadratic program of optimal margin classifier.	[49]

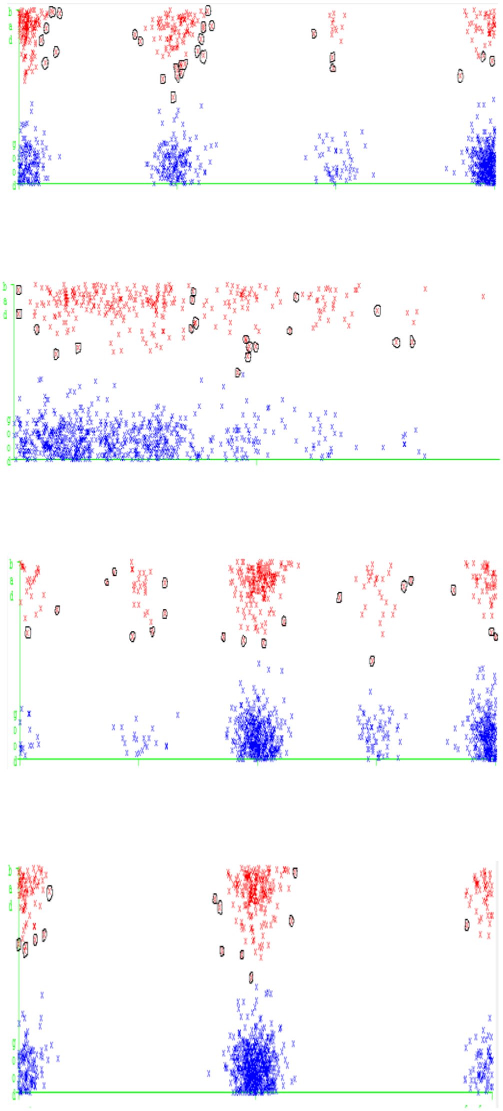
María Dolores Pérez-Godoy *et al.* [24] have proposed CO2RBFN, a evolutionary cooperative–competitive model for the design of radial-basis function networks which uses both radial-basis function and the evolutionary cooperative-competitive technique on imbalanced domains. CO2RBFN follows the evolutionary cooperative–competitive strategy, where each individual of the population represents an RBF (Gaussian function will be considered as RBF) and the entire population is responsible for the definite solution. This paradigm provides a framework where an individual of the population represents only a part of the solution, competing to survive (since it will be eliminated if its performance is poor) but at the same time cooperating in order to build the whole RBFN, which adequately represents the knowledge about the problem and achieves good generalization for new patterns.

Der-Chiang Li *et al.* [25] have suggested a strategy which over-samples the minority class and under-samples the majority one to balance the datasets. For the majority class, they build up the Gaussian type fuzzy membership function and a-cut to reduce the data size; for the minority class, they used the mega-trend diffusion membership function to generate virtual samples for the class. Furthermore, after balancing the data size of classes, they extended the data attribute dimension into a higher dimension space using classification related information to enhance the classification accuracy. Enhong Che *et al.* [26] have described a unique approach to improve text categorization under class imbalance by exploiting the semantic context in text documents. Specifically, they generate new samples of rare classes (categories with relatively small amount of training data) by using global semantic information of classes represented by probabilistic topic models. In this way, the numbers of samples in different categories can become more balanced and the performance of text categorization can be improved using this transformed data set. Indeed, this method is different from traditional re-sampling methods, which try to balance the number of documents in different classes by re-sampling the documents in rare classes. Such re-sampling methods can cause overfitting. Another benefit of this approach is the effective handling of noisy samples. Since all the new samples are generated by topic models, the impact of noisy samples is dramatically reduced. Alberto Fernández *et al.* [27] have proposed an improved version of fuzzy rule based classification systems (FRBCSs) in the framework of imbalanced data-sets by means of a tuning step. Specifically, they adapt the 2-tuples based genetic tuning approach to classification problems showing the good synergy between this method and some FRBCSs. The proposed algorithm uses two learning methods in order to generate the RB for the FRBCS. The first one is the method proposed in [28], that they have named the Chi et al.'s rule generation. The second approach is defined by Ishibuchi and Yamamoto in [29] and it consists of a Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML) algorithm.

J. Burez *et al.* [30] have investigated how they can better handle class imbalance in churn prediction. Using more appropriate evaluation metrics (AUC, lift), they investigated the increase in performance of sampling (both random and advanced under-sampling) and two specific modeling techniques (gradient boosting and weighted random forests) compared to some standard modeling techniques. They have advised weighted random forests, as a cost-sensitive learner, performs significantly better compared to random forests.Che-Chang Hsu *et al.* [31] have proposed a method with a model assessment of the interplay between various classification decisions using probability, corresponding decision costs, and quadratic program of optimal margin classifier called: Bayesian Support Vector Machines (BSVMs) learning strategy. The purpose of their learning method is to lead an attractive pragmatic expansion scheme of the Bayesian approach to assess how well it is aligned with the class imbalance problem. In the framework, they did modify in the objects and conditions of primal problem to reproduce an appropriate learning rule for an observation sample. In [32] Alberto Fernández *et al.* have proposed to work with fuzzy rule based classification systems using a preprocessing step in order to deal with the class imbalance. Their aim is to analyze the behavior of fuzzy rule based classification systems in the framework of imbalanced data-sets by means of the application of an adaptive inference system with parametric conjunction operators. Jordan M. Malof *et al.* [33] have empirically investigates how class imbalance in the available set of training cases can impact the performance of the resulting classifier as well as properties of the selected set. In this K-Nearest Neighbor (k-NN) classifier is used which is a well-known classifier and has been used in numerous case-based classification studies of imbalance datasets. The bottom line is that when studying problems with imbalanced data, using the classifiers produced by standard machine learning algorithms without adjusting the output threshold may well be a critical mistake. This skewness towards minority class (positive) generally causes the generation of a high number of false-negative predictions, which lower the model's performance on the positive class compared with the performance on the negative (majority) class.

V. METHODOLOGY

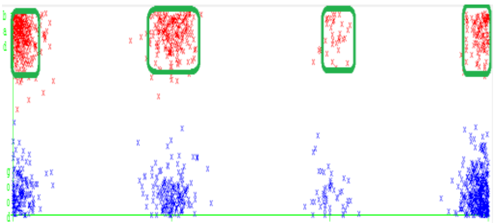
In this section, we follow a design decomposition approach to systematically analyze the different imbalanced domains. We first briefly introduce the framework design for our proposed algorithm.The working style of oversampling tries to generate synthetic minority instances. Before performing oversampling on the minority subset, the main cluster disjuncts has to be identified and the borderline and noise instances around the cluster disjuncts are to be removed. The number of instances eliminated will belong to the '*k'* cluster disjuncts selected by visualization technique. The remaining cluster disjunct instances have to be oversampled by using hybrid synthetic oversampling technique. Here, the above said routine is employed o every cluster disjunct, which removes examples suffering from missing values at first and then removes borderline examples and examples of outlier category.



(a)

(b)

(c)

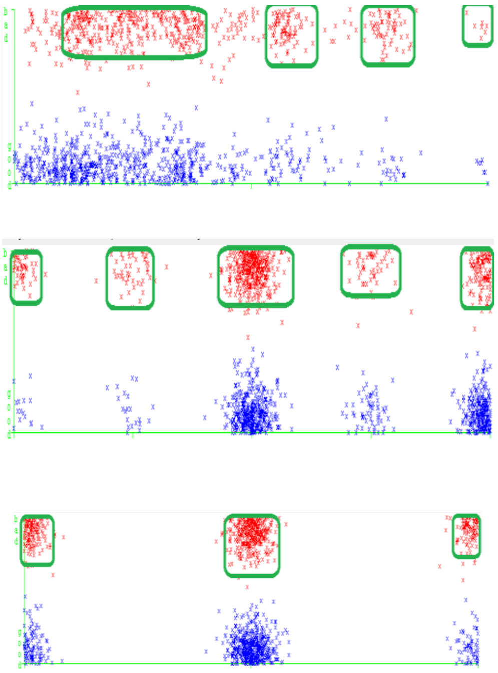


(d)

Fig 1 Before (a). Checking Status (b). Duration (c). Credit History (d). Housing

The algorithm 1: CDMOTE and NBCD can be explained as follows, the inputs to the algorithm are majority subclass "p" and minority class "n" with the number of features j. The output of the algorithm will be the average measures such as AUC, Precision, F-measure, TP rate and TN rate produced by the CDMOTE and NBCD methods. The algorithm begins with initialization of k=1 and j=1, where j is the number of cluster disjuncts identified by applying visualization technique on the subset "n" and k is the variable used for looping of j cluster disjuncts.

(a)



(b)

(c)

(d)

Fig 2 After Applying CDMOTE: (a). Checking Status (b). Duration (c). Credit History (d). Housing

The 'j' value will change from one dataset to other, and depending upon the unique properties of the dataset the value of k can be equal to one also i.e no cluster disjunct attributes can be identified after applying visualization technique on the dataset. In another case attributes related cluster disjunct oversampling can also be performed to improve the skewed dataset. In any case depending on the amount of minority examples generated, the final "strong set" can or cannot be balanced i;e number of majority instances and minority instances in the strong set will or will not be equal. The presented CDMOTE and NBCD algorithms are summarized as below.

Table 2 Algorithm 1 CDMOTE or NBCD

```
Input: A set of major subclass examples $P$, a set
       of minor subclass examples $N$, $jPj < jNj$,
       and $Fj$, the feature set, j > 0.
Output: Average Measure { AUC, Precision,
        F-Measure, TP Rate, TN Rate}
Phase I: Initial Phase:
begin
  $k \leftarrow 1, j \leftarrow 1$.
  Apply Visualization Technique on subset $N$,
  Identify cluster disjunct $Cj$ from N, j= number
    of cluster disjunct identified in visualization
  repeat
    k=k+1
    Identify and remove the borderline and outlier
      instances for the cluster disjunct $Cj$.
  Until k = j
Phase II: Over sampling Phase
  Apply Oversampling on $Cj$ cluster disjunct
    from N,
  repeat
    k=k+1
    Generate '$Cj \times s$' synthetic positive
      examples from the minority examples in each
      cluster disjunct $Cj$.
  Until k = j
Phase III: Validating Phase
  Train and Learn A Base Classifier (C4.5) or
    (Naïve Bayes) using $P$ and $N$
end
```

The different components of our new proposed framework are elaborated in the next subsections.

A. Preparation of the Majority and Minority subsets

The datasets is partitioned into majority and minority subsets. As we are concentrating over sampling, we will take minority data subset for further visualization analysis to identify cluster disjuncts.

B. Initial phase of removing noisy and cluster disjunct borderline instances

Minority subset can be further analyzed to find the noisy or borderline instances so that we can eliminate those. For finding the weak instances one of the ways is that find most influencing attributes or features and then remove ranges of the noisy or weak attributes relating to that feature. How to choose the noisy instances relating to that cluster disjunct from the dataset set? We can find a range where the number of samples are less can give you a simple hint that those instances coming in that range or very rare or noise. We will intelligently detect and remove those instances which are in narrow ranges of that particular cluster disjunct. This process can be applied on all the cluster disjuncts identified for each dataset.

C. Applying oversampling on cluster disjunct

The oversampling of the instances can be done on the improved cluster disjuncts produced in the earlier phase. The oversampling can be done as follows:

Apply resampling supervised filter on the cluster disjunct for generating synthetic instances. The synthetic minority instances generated can have a percentage of instances which can be replica of the pure instances and reaming percentage of instances are of the hybrid quality of synthetic instances generated by combing two or more instances from the pure minority sunset. Perform oversampling on cluster disjunct can help so as to form strong, efficient and more valuable rules for proper knowledge discovery.

D. Forming the strong dataset

The minority subset and majority subset is combined to form a strong and balance dataset, which is used for learning of a base algorithm. In this case we have used C4.5 or Naïve Bayes as the base algorithm.

VI. EVALUATION METRICS

To assess the classification results we count the number of true positive (TP), true negative (TN), false positive (FP) (actually negative, but classified as positive) and false negative (FN) (actually positive, but classified as negative) examples. It is now well known that error rate is not an appropriate evaluation criterion when there is class imbalance or unequal costs. In this paper, we use AUC, Precision, F-measure, TP Rate and TN Rate as performance evaluation measures.

Let us define a few well known and widely used measures:

The Area under Curve (AUC) measure is computed by equation (1),

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \qquad \text{\_\_\_\_\_\_\_\_} \quad (1)$$

The Precision measure is computed by equation (2),

$$\Pr ecision = \frac{TP}{(TP) + (FP)} \qquad \text{\_\_\_\_\_\_\_\_} \quad (2)$$

The F-measure Value is computed by equation (3),

$$F - measure = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \qquad \text{\_\_\_\_} (3)$$

The True Positive Rate measure is computed by equation (4),

$$\text{TruePositi veRate} = \frac{TP}{(TP)+(FN)} \qquad \underline{\qquad\qquad} (4)$$

The True Negative Rate measure is computed by equation (5),

$$\text{TrueNegati veRate} = \frac{TN}{(TN)+(FP)} \qquad \underline{\qquad\qquad} (5)$$

VII. EXPERIMENTAL FRAMEWORK

In this study CDMOTE and NBCD are applied to fifteen binary data sets from the UCI repository [34] with different imbalance ratio (IR). Table 3 summarizes the data selected in this study and shows, for each data set, the number of examples (#Ex.), number of attributes (#Atts.), class name of each class (minority and majority) and IR. In order to estimate different measure (AUC, precision, F-measure, TP rate and TN rate) we use a tenfold cross validation approach, that is ten partitions for training and test sets, 90% for training and 10% for testing, where the ten test partitions form the whole set. For each data set we consider the average results of the ten partitions.

To validate the proposed CDMOTE and NBCD algorithms, we compared it with the traditional Support Vector Machines (SVM), C4.5, Functional Trees (FT), and SMOTE (Synthetic Minority Oversampling TEchnique).

Table 3 Summary of benchmark imbalanced datasets

S.no Datasets	# Ex.	# Atts.	Class (_,+)	IR
1.Breast	268	9	(recurrence; no-recurrence)	2.37
2. Breast_w	699	9	(benign; malignant)	1.90
3.Colic	368	22	(yes; no)	1.71
4.Credit-g	1000	21	(good; bad)	2.33
5.Diabetes	768	8	(tested-potv; tested-negtv)	1.87
6.Heart-c	303	14	(<50,>50_1)	1.19
7.Heart-h	294	14	(<50,>50_1)	1.77
8.Heart-stat	270	14	(absent, present)	1.25
9.Hepatitis	155	19	(die; live)	3.85
10.Ionosphere	351	34	(b;g)	1.79
11. Kr-vs-kp	3196	37	(won; nowin)	1.09
12. Labor	56	16	(bad ; good)	1.85
13. Mushroom	8124	23	(e ; p)	1.08
14. Sick	3772	29	(negative ; sick)	15.32
15. Sonar	208	60	(rock ; mine)	1.15

VIII. RESULTS

For all experiments, we use existing prototype’s present in Weka [42]. We compare the following domain adaptation methods:

We compared proposed methods CDMOTE and NBCD with the SVM, C4.5 [44], FT, and SMOTE [41] state-of -the-art learning algorithms. In all the experiments we estimate AUC, Precision, F-measure, TP rate and TN rate using 10-fold cross-validation. We experimented with 15 standard datasets for UCI repository; these datasets are standard benchmarks used in the context of high-dimensional imbalance learning. Experiments on these datasets have 2 goals. First, we study the class imbalance properties of the datasets using proposed CDMOTE and NBCD learning algorithms. Second, we compare the classification performance of our proposed CDMOTE and NBCD algorithms with the traditional and class imbalance learning methods based on all datasets.

Following, we analyze the performance of the method considering the entire original algorithms, without pre-processing, data sets for SVM, C4.5 and FT. we also analyze a pre-processing method SMOTE for performance evaluation of CDMOTE and NBCD. The complete table of results for all the algorithms used in this study is shown in Table 4 to 9, where the reader can observe the full test results, of performance of each approach with their associated standard deviation. We must emphasize the good results achieved by CDMOTE and NBCD, as it obtains the highest value among all algorithms.

Table 4 Summary of tenfold cross validation performance for Accuracy on all the datasets

Datasets	SVM	C4.5	FT	SMOTE	CDMOTE	NBCD
Breast	67.21±7.28	**74.28±6.05**	68.58±7.52	69.83±7.77	70.23±5.91	**73.356±6.603**
Breast_w	**96.75±2.00**	95.01±2.73	95.45±2.52	96.16±2.06	96.58±1.79	**97.971±1.503**
Colic	79.78±6.57	85.16±5.91	79.11± 6.51	**88.53±4.10**	87.92±4.70	**90.641±4.640**
Credit-g	68.91±4.46	71.25±3.17	71.88±3.68	**76.50±3.38**	75.06±3.89	**76.844±4.494**
Diabetes	76.55±4.67	74.49±5.27	70.62± 4.67	76.08±4.04	**81.75±4.08**	**79.333±4.137**
Heart-c	81.02±7.25	76.94±6.59	76.06±6.84	**82.99±4.98**	80.57±6.55	**83.052±6.371**
Heart-h	81.81±6.20	80.22±7.95	78.33±7.54	**85.65±5.46**	83.56±5.81	**85.178±5.143**
Heart-stat	**82.07±6.88**	78.15±7.42	76.15±8.46	**83.89±5.05**	80.31±7.75	81.872±7.342
Hepatitis	81.90±8.38	79.22±9.57	81.40±8.55	78.35±9.09	83.59±9.65	**89.529±8.001**

Ionosphere	90.26±4.97	89.74±4.38	87.10±5.12	90.28±4.73	**94.64±3.74**	**94.411±3.590**
Kv-rs-kp	99.02±0.54	99.44±0.37	90.61±1.65	**99.66±0.27**	**99.45±0.42**	98.103±1.636
Labor	**92.40±11.07**	78.60±16.58	84.30±16.24	80.27±11.94	88.33±11.09	**95.905±7.259**
Mushroom	100.0±0.00	100.0±0.00	100.0±0.000	100.0±0.00	100.0±0.00	100.0±0.00
Sick	**99.26±0.04**	98.72±0.55	96.10±0.92	97.61±0.68	**99.07±0.50**	98.379±0.691
Sonar	75.46±9.92	73.61±9.34	86.17±8.45	82.42±7.25	**86.23±8.31**	**86.17±8.187**

Table 5 Summary of tenfold cross validation performance for AUC on all the datasets

Datasets	SVM	C4.5	FT	SMOTE	CDMOTE	NBCD
Breast	0.586±0.102	0.606±0.087	0.604±0.082	**0.717±0.084**	0.705±0.082	**0.799±0.074**
Breast_w	**0.977±0.017**	0.957±0.034	0.949±0.030	0.967±0.025	0.973±0.018	**0.991±0.009**
Colic	0.802±0.073	0.843±0.070	0.777±0.072	**0.908±0.040**	0.900±0.042	**0.958±0.029**
Credit-g	0.650±0.075	0.647±0.062	0.655±0.044	0.778±0.041	**0.788±0.041**	**0.847±0.043**
Diabetes	0.793±0.072	0.751±0.070	0.668±0.051	0.791±0.041	**0.836±0.046**	**0.849±0.040**
Heart-c	**0.843±0.084**	0.769±0.082	0.757±0.069	0.830±0.077	0.822±0.077	**0.913±0.052**
Heart-h	0.852±0.078	0.775±0.089	0.763±0.082	**0.904±0.054**	0.869±0.065	**0.923±0.043**
Heart-stat	**0.864±0.075**	0.786±0.094	0.760±0.085	0.832±0.062	0.822±0.076	**0.870±0.068**
Hepatitis	0.757±0.195	0.668±0.184	0.678±0.139	0.798±0.112	**0.848±0.136**	**0.952±0.056**
Ionosphere	0.900±0.060	0.891±0.060	0.831±0.067	0.904±0.053	**0.949±0.041**	**0.961±0.032**
Kr-vs-kp	0.996±0.005	0.998±0.003	0.906±0.017	**0.999±0.001**	**0.998±0.002**	0.995±0.004
Labor	**0.971±0.075**	0.726±0.224	0.844±0.162	0.833±0.127	0.870±0.126	**0.995±0.024**
Mushroom	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00	1.000±0.00
Sick	**0.990±0.014**	0.952±0.040	0.795±0.053	0.962±0.025	**0.992±0.012**	0.979±0.019
Sonar	0.771±0.103	0.753±0.113	0.859±0.086	0.814±0.090	0.854±0.086	**0.924±0.063**

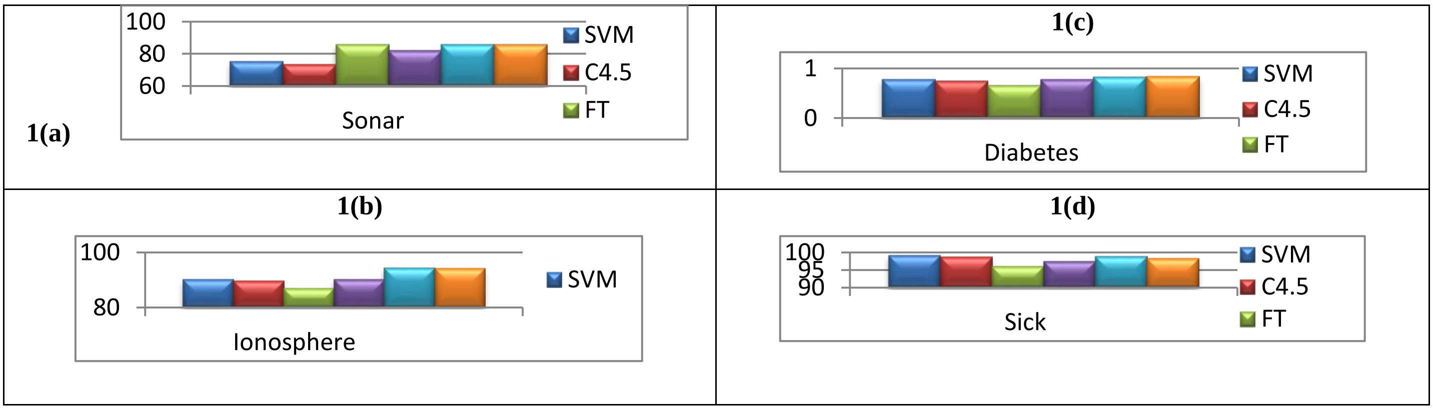


Fig. 1(a) – (d) Test results on AUC on C4.5, CART, FT, REP, SMOTE and CDMOTE for Sonar, Ionosphere, Diabetes and Sick Datasets.

Table 5 Summary of tenfold cross validation performance for Precision on all the datasets

Datasets	SVM	C4.5	FT	SMOTE	CDMOTE	NBCD
Breast	0.745±0.051	0.753±0.042	**0.762±0.051**	0.710±0.075	0.713±0.059	**0.770±0.062**
Breast_w	**0.988±0.019**	0.965±0.026	0.964±0.026	0.974±0.025	0.986±0.021	**0.996±0.011**
Colic	0.845±0.060	0.851±0.055	0.839±0.062	0.853±0.057	**0.864±0.059**	**0.925±0.058**
Credit-g	0.776±0.033	0.767±0.025	0.791±0.027	0.768±0.034	**0.799±0.044**	**0.805±0.052**
Diabetes	0.793±0.037	0.797±0.045	0.764±0.036	0.781±0.064	**0.862±0.050**	**0.826±0.054**
Heart-c	**0.825±0.080**	0.783±0.076	0.776±0.068	0.779±0.082	0.808±0.087	**0.831±0.084**
Heart-h	0.849±0.058	0.824±0.071	0.830±0.063	0.878±0.076	**0.894±0.072**	**0.896±0.070**
Heart-stat	**0.833±0.078**	0.799±0.051	0.796±0.085	0.791±0.081	0.821±0.094	**0.828±0.084**
Hepatitis	0.604±0.271	0.510±0.371	0.546±0.333	0.709±0.165	**0.739±0.200**	**0.791±0.151**
Ionosphere	0.906±0.080	0.895±0.084	0.938±0.073	0.934±0.049	**0.945±0.047**	**0.944±0.051**
Kr-vs-kp	0.991±0.008	0.994±0.006	0.905±0.021	**0.996±0.005**	**0.994±0.005**	0.978±0.023
Labor	0.915±0.197	0.696±0.359	0.802±0.250	0.871±0.151	**0.921±0.148**	**0.938±0.122**
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Sick	**0.997±0.003**	0.992±0.005	0.975±0.007	0.983±0.007	**0.996±0.004**	0.990±0.005
Sonar	0.764±0.119	0.728±0.121	**0.883±0.100**	**0.863±0.068**	0.851±0.090	0.858±0.092

Table 6 Summary of tenfold cross validation performance for F-measure on all the datasets

Datasets	SVM	C4.5	FT	SMOTE	CDMOTE	NBCD
Breast	0.781±0.059	**0.838±0.040**	0.776±0.057	0.730±0.076	0.775±0.049	**0.782±0.056**
Breast_w	0.965±0.019	0.962±0.021	**0.975±0.016**	0.960±0.022	0.967±0.018	**0.980±0.015**
Colic	0.833±0.055	**0.888±0.044**	0.838±0.054	0.880±0.042	0.887±0.043	**0.908±0.045**
Credit-g	**0.802±0.027**	**0.805±0.022**	0.779±0.034	0.787±0.034	0.763±0.039	0.784±0.041
Diabetes	0.778±0.037	**0.806±0.044**	**0.827±0.038**	0.741±0.046	0.808±0.047	0.786±0.044
Heart-c	0.782±0.064	0.792±0.059	**0.827±0.069**	0.772±0.070	0.800±0.069	**0.827±0.065**
Heart-h	0.830±0.063	**0.851±0.061**	**0.859±0.052**	0.841±0.061	0.829±0.066	0.850±0.054
Heart-stat	0.781±0.083	0.806±0.069	**0.841±0.061**	0.791±0.072	0.802±0.076	**0.819±0.077**
Hepatitis	0.469±0.265	0.409±0.272	0.557±0.207	0.677±0.138	**0.693±0.192**	**0.830±0.129**
Ionosphere	0.787±0.098	0.850±0.066	0.855±0.079	0.905±0.048	**0.944±0.039**	**0.942±0.037**
Kv-rs-kp	0.911±0.016	0.995±0.004	0.991±0.005	**0.995±0.004**	**0.994±0.004**	0.981±0.016
Labor	0.794±0.211	0.636±0.312	**0.879±0.195**	0.793±0.132	0.842±0.157	**0.954±0.082**
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.979±0.005	0.993±0.003	**0.996±0.003**	0.987±0.004	**0.995±0.003**	0.991±0.004
Sonar	0.844±0.099	0.716±0.105	0.753±0.102	0.861±0.061	**0.867±0.082**	**0.866±0.080**

Table 7 Summary of tenfold cross validation performance for TP Rate (Recall) (Sensitivity) on all the datasets

Datasets	SVM	C4.5	FT	SMOTE	CDMOTE	NBCD
Breast	0.806±0.091○	0.947±0.060	**0.815±0.095**	0.763±0.117	**0.861±0.101**	.800±0.085
Breast_w	**0.967±0.025**	0.959±0.033	0.962±0.029	0.947±0.035	0.950±0.033	**0.965±0.026**
Colic	0.832±0.075	**0.931±0.053**	0.835±0.077	0.913±0.058	**0.915±0.058**	0.896±0.063
Credit-g	**0.815±0.041**	**0.847±0.036**	0.783±0.052	0.810±0.058	0.733±0.057	0.767±0.051
Diabetes	0.795±0.054	**0.821±0.073**	**0.868±0.065**	0.712±0.089	0.763±0.070	0.753±0.061
Heart-c	0.795±0.095	0.808±0.085	**0.837±0.100**	0.777±0.110	0.802±0.102	**0.831±0.092**
Heart-h	0.835±0.093	**0.885±0.081**	**0.876±0.089**	0.815±0.084	0.783±0.107	0.816±0.088
Heart-stat	0.775±0.113	**0.824±0.104**	**0.857±0.090**	0.803±0.110	0.794±0.102	0.817±0.102
Hepatitis	0.448±0.273	0.374±0.256	0.573±0.248	0.681±0.188	**0.700±0.247**	**0.892±0.149**
Ionosphere	0.689±0.131	0.821±0.107	0.820±0.114	0.881±0.071	**0.946±0.054**	**0.943±0.053**
Kv-rs-kp	0.916±0.021	0.995±0.005	0.990±0.007	**0.995±0.006**	**0.995±0.006**	0.985±0.012
Labor	0.845±0.243	0.640±0.349	**0.885±0.234**	0.765±0.194	0.823±0.227	**0.983±0.073**
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.984±0.006	**0.995±0.004**	**0.995±0.004**	0.990±0.005	0.993±0.004	0.992±0.005
Sonar	0.820±0.131	0.721±0.140	0.757±0.136	0.865±0.090	**0.893±0.109**	**0.883±0.105**

Table 8 Summary of tenfold cross validation performance for TN Rate (Specificity) on all the datasets

Datasets	SVM	C4.5	FT	SMOTE	CDMOTE	NBCD
Breast	0.260±0.141	0.335±0.166	0.151±0.164	**0.622±0.137**	0.464±0.169	**0.634±0.128**
Breast_w	0.932±0.052	0.977±0.037	0.931±0.060	0.975±0.024	**0.984±0.025**	**0.996±0.012**
Colic	0.717±0.119	0.734±0.118	0.731±0.121	**0.862±0.063**	0.841±0.080	**0.918±0.069**
Credit-g	0.398±0.085	0.469±0.098	0.371±0.105	0.713±0.056	**0.772±0.063**	**0.771±0.073**
Diabetes	0.603±0.111	0.574±0.095	0.567±0.105	0.807±0.077	**0.873±0.054**	**0.834±0.063**
Heart-c	0.723±0.119	0.779±0.117	0.717±0.119	**0.861±0.068**	0.809±0.099	**0.830±0.097**
Heart-h	0.655±0.158	0.714±0.131	0.636±0.152	**0.894±0.074**	**0.893±0.079**	0.891±0.085
Heart-stat	0.728±0.131	0.775±0.123	0.677±0.152	**0.862±0.064**	0.812±0.115	**0.820±0.098**
Hepatitis	**0.900±0.097**	0.882±0.092	**0.942±0.093**	0.837±0.109	0.888±0.097	0.896±0.090
Ionosphere	0.940±0.055	**0.949±0.046**	0.933±0.063	0.928±0.057	**0.947±0.047**	0.945±0.054
Kv-rs-kp	0.993±0.007	0.990±0.009	0.987±0.010	**0.998±0.003**	**0.994±0.006**	0.977±0.025
Labor	0.865±0.197	**0.945±0.131**	0.843±0.214	0.847±0.187	0.928±0.138	**0.946±0.106**
Mushroom	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.875±0.071	**0.974±0.026**	0.846±0.080	0.872±0.053	**0.970±0.031**	0.919±0.045
Sonar	0.749±0.134	0.752±0.148	0.762±0.145	0.752±0.113	**0.831±0.113**	**0.839±0.120**

Fig.1(a)-(d) shows the average AUC computed for all approaches, where we can observe that CDMOTE and NBCD has obtained the best AUC and accuracy values in the comparison and therefore it is clearly given the indication of its supremacy. Table 4, 5, 6, 7, 8 and 9 reports the results of AUC, Precision, F-measure, TP Rate, TN Rate and accuracy respectively for fifteen UCI datasets.

The bold indicates a win of that algorithm o other compared algorithms. For each row top two algorithms are marked as bold. The results in the tables show that CDMOTE and NBCD have given a good improvement on all the measures of class imbalance learning. This level of analysis is enough for overall projection of advantages and disadvantages of CDMOTE and NBCD. A two-tailed corrected resampled paired t-test is used in this paper to determine whether the results of the cross-validation show that there is a difference between the two algorithms is significant or not. Difference in accuracy is considered significant when the p-value is less than 0.05 (confidence level is greater than 95%). In discussion of results, if one algorithm is stated to be better or worse than another then it is significantly better or worse at the 0.05 level.

Table 9 Summary of comparison of our proposed approach with a recent published algorithm CILIUS

Datasets	**CILIUS [22]**	**CDMOTE**	**NBCD**
	AUC		
Breast	0.637 ± 0.110	**0.705±0.082**	**0.799±0.074**
Breast_w	0.987 ± 0.016	0.973±0.018	**0.991±0.009**
Colic	0.873 ± 0.082	**0.900±0.042**	**0.958±0.029**
Diabetes	0.826 ± 0.056	**0.836±0.046**	**.849±0.040**
Hepatitis	0.714 ± 0.166	**0.848±0.136**	**0.952±0.056**
Ionosphere	0.917 ± 0.048	**0.949±0.041**	**0.961±0.032**
Labor	0.765 ± 0.217	**0.870±0.126**	**0.995±0.024**
Sick	0.950 ± 0.035	**0.992±0.012**	**0.979±0.019**
Sonar	0.774 ± 0.114	**0.854±0.086**	**0.924±0.063**
	Precision		
Breast	0.736 ± 0.050	0.713±0.059	**0.770±0.062**
Breast_w	0.986 ± 0.020	**0.986±0.021**	**0.996±0.011**
Colic	0.787 ± 0.090	**0.864±0.059**	**0.925±0.058**
Diabetes	0.810 ± 0.048	**0.862±0.050**	**0.826±0.054**
Hepatitis	0.698 ± 0.305	**0.739±0.200**	**0.791±0.151**
Ionosphere	0.922 ± 0.071	**0.945±0.047**	**0.944±0.051**
Labor	0.754 ± 0.337	**0.921±0.148**	**0.938±0.122**
Sick	0.990 ± 0.006	**0.996±0.004**	0.990±0.005
Sonar	0.759 ± 0.112	**0.851±0.090**	**0.858±0.092**
		F-measure	
Breast	**0.812 ± 0.046**	0.775±0.049	0.782±0.056
Breast_w	**0.984 ± 0.014**	0.967±0.018	0.980±0.015
Colic	0.827 ± 0.073	**0.887±0.043**	**0.908±0.045**
Diabetes	**0.836 ± 0.040**	0.808±0.047	0.786±0.044
Hepatitis	0.556 ± 0.238	**0.693±0.192**	**0.830±0.129**
Ionosphere	0.881 ± 0.065	**0.944±0.039**	**.942±0.037**
Labor	0.697 ± 0.307	**0.842±0.157**	**0.954±0.082**
Sick	0.991 ± 0.004	**0.995±0.003**	0.991±0.004
Sonar	0.752 ± 0.103	**0.867±0.082**	**0.866±0.080**
		TP Rate	
Breast	0.325 ± 0.156	**0.861±0.101**	**0.800±0.085**
Breast_w	**0.978 ± 0.030**	0.950±0.033	0.965±0.026
Colic	0.765 ± 0.122	**0.915±0.058**	**0.896±0.063**
Diabetes	0.696 ± 0.096	**0.763±0.070**	**0.753±0.061**
Hepatitis	**0.920 ± 0.092**	0.700±0.247	0.892±0.149
Ionosphere	**0.948 ± 0.052**	0.946±0.054	0.943±0.053
Labor	0.865 ± 0.207	0.823±0.227	**0.983±0.073**
Sick	0.903 ± 0.060	**0.993±0.004**	**0.992±0.005**
Sonar	0.743 ± 0.138	**0.893±0.109**	**.883±0.105**

We can make a global analysis of results combining the results offered by Tables from 4–9 and Fig. 1(a)-(d):

- Our proposals, CDMOTE and NBCD are the best performing one when the data sets are no preprocessed. It outperforms the pre-processing SMOTE methods and this hypothesis is confirmed by including standard deviation variations. We have considered a complete competitive set of methods and an improvement of results is expected in the benchmark algorithms i;e SVM, C4.5, and FT. However, they are not able to outperform CDMOTE and NBCD. In this sense, the competitive edge of CDMOTE and NBCD can be seen.
- Considering that CDMOTE and NBCD behaves similarly or not effective than SMOTE shows the unique properties of the datasets where there is scope of improvement in majority subset and not in minority subset. Our CDMOTE and NBCD can only consider improvements in minority subset which is not effective for some unique property datasets.

The contributions of this work are twofold:

A general strategy to handle class imbalance problem: This is scalable, flexible, and modular, allowing the many existing supervised methods to be as a base algorithm. The method achieves competitive or better results compared to state-of-the-art baselines.Table 9 reports the comparison of our proposed approach with a recent published algorithm CILIUS [22] and our proposed algorithm has performed well. A specific methods based on cluster disjoints for coping up with class imbalance are proposed. This method naturally captures ad removes the noise and weak instances for performing over sampling, which has not been previously examined in the context of supervised approach.

We emphasize that our approach is learner-independent: visualization can be used in conjunction with many of the existing algorithms in the literature. Furthermore, the fact that we select samples in the model space, as opposed to the feature space, is novel and sets it apart from many previous approaches to transfer learning (for both classification and ranking). This allows us to capture the ‘‘functional change’’ assumption and incorporate labeled information in the transfer learning process.Finally, we can say that CDMOTE and NBCD are one of the best alternatives to handle class imbalance problems effectively. This experimental study supports the conclusion that a cluster disjunct approach for cluster detections and elimination can improve the class imbalance learning behavior when dealing with imbalanced data-sets, as it has helped the CDMOTE and NBCD methods to be the best performing algorithms when compared with four classical and well-known algorithms: SVM, C4.5 and FT and a well-established pre-processing technique SMOTE.

IX. CONCLUSION

Class imbalance problem have given a scope for a new paradigm of algorithms in data mining. The traditional and benchmark algorithms are worthwhile for discovering hidden knowledge from the data sources, meanwhile class imbalance learning methods can improve the results which are very much critical in real world applications. In this paper we present the class imbalance problem paradigm, which exploits the cluster disjunct concept in the supervised learning research area, and implement it with C4.5 and NB as its base learners. Experimental results show that CDMOTE and NBCD have performed well in the case of multi class imbalance datasets. Furthermore, CDMOTE and NBCD are much less volatile than C4.5. In our future work, we will apply CDMOTE and NBCD to more learning tasks, especially high dimensional feature learning tasks. Another variation of our approach in future work is to analyze the influence of different base classifier effect on the quality of synthetic minority instances generated.

REFERENCES

[1] J. Wu, S. C. Brubaker, M. D. Mullin, and J. M. Rehg, “Fast asymmetric learning for cascade face detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 3, pp. 369–382, Mar. 2008.

[2] N. V. Chawla, N. Japkowicz, and A. Kotcz, Eds., Proc. ICML Workshop Learn. Imbalanced Data Sets, 2003.

[3] N. Japkowicz, Ed., Proc. AAAI Workshop Learn. Imbalanced Data Sets, 2000.\

[4] G. M.Weiss, “Mining with rarity: A unifying framework,” ACM SIGKDD Explor. Newslett., vol. 6, no. 1, pp. 7–19, Jun. 2004.

[5] N. V. Chawla, N. Japkowicz, and A. Kolcz, Eds., Special Issue Learning Imbalanced Datasets, SIGKDD Explor. Newsl., vol. 6, no. 1, 2004.

[6] W.-Z. Lu and D.Wang, “Ground-level ozone prediction by support vector machine approach with a cost-sensitive classification scheme,” Sci. Total. Enviro., vol. 395, no. 2-3, pp. 109–116, 2008.

[7] Y.-M. Huang, C.-M. Hung, and H. C. Jiau, “Evaluation of neural networks and data mining methods on a credit assessment task for class imbalance problem,” Nonlinear Anal. R. World Appl., vol. 7, no. 4, pp. 720–747, 2006.

[8] D. Cieslak, N. Chawla, and A. Striegel, “Combating imbalance in network intrusion datasets,” in IEEE Int. Conf. Granular Comput., 2006, pp. 732–737.

[9] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, “Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance,” Neural Netw., vol. 21, no. 2–3, pp. 427–436, 2008.

[10] A. Freitas, A. Costa-Pereira, and P. Brazdil, “Cost-sensitive decision trees applied to medical data,” in Data Warehousing Knowl. Discov. (Lecture Notes Series in Computer Science), I. Song, J. Eder, and T. Nguyen, Eds.,

[11] K.Kilic¸,O¨ zgeUncu and I. B. Tu¨rksen, “Comparison of different strategies of utilizing fuzzy clustering in structure identification,” Inf. Sci., vol. 177, no. 23, pp. 5153–5162, 2007.

[12] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, “A methodological approach to the classification of dermoscopy images,” Comput.Med. Imag. Grap., vol. 31, no. 6, pp. 362–373, 2007.

[13] X. Peng and I. King, “Robust BMPM training based on second-order cone programming and its application in medical diagnosis,” Neural Netw., vol. 21, no. 2–3, pp. 450–457, 2008.Berlin/Heidelberg, Germany: Springer, 2007, vol. 4654, pp. 303–312.

[14] Rukshan Batuwita and Vasile Palade (2010) FSVM-CIL: Fuzzy Support Vector Machines for Class Imbalance Learning, IEEE TRANSACTIONS ON FUZZY SYSTEMS, VOL. 18, NO. 3, JUNE 2010, pp no:558-571.

[15] N. Japkowicz and S. Stephen, “The Class Imbalance Problem: A Systematic Study,” Intelligent Data Analysis, vol. 6, pp. 429-450, 2002.

[16] M. Kubat and S. Matwin, “Addressing the Curse of Imbalanced Training Sets: One-Sided Selection,” Proc. 14th Int’l Conf. Machine Learning, pp. 179-186, 1997.

[17] G.E.A.P.A. Batista, R.C. Prati, and M.C. Monard, “A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data,” SIGKDD Explorations, vol. 6, pp. 20-29, 2004.1.

[18] Siti Khadijah Mohamada, Zaidatun Tasir. “Educational data mining: A review”, Procedia - Social and Behavioral Sciences 97 (2013) 320 – 324.

[19] Hongzhou Sha, Tingwen Liu, Peng Qin, Yong Sun, Qingyun Liu.” EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining” Procedia Computer Science 17 (2013) 812 – 818.

[20] M.S.B. PhridviRaj, C.V. GuruRao.” Data mining – past, present and future – a typical survey on data Streams”, Procedia Technology 12 (2014) 255 – 263.

[21] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, Chidchanok Lursinsap.” DBSMOTE: Density-Based Synthetic Minority Over-sampling Technique” Appl Intell (2012) 36:664–684.

[22] Matías Di Martino, Alicia Fernández, Pablo Iturralde, Federico Lecumberry.” Novel classifier scheme for imbalanced problems”, Pattern Recognition Letters 34 (2013) 1146–1151.

[23] V. Garcia, J.S. Sanchez , R.A. Mollineda,” On the effectiveness of preprocessing methods when dealing with different levels of class imbalance”, Knowledge-Based Systems 25 (2012) 13–21.

[24] María Dolores Pérez-Godoy, Alberto Fernández, Antonio Jesús Rivera, María José del Jesus,” Analysis of an evolutionary RBFN design algorithm, CO2RBFN, for imbalanced data sets”, Pattern Recognition Letters 31 (2010) 2375–2388.

[25] Der-Chiang Li, Chiao-WenLiu, SusanC.Hu,” A learning method for the class imbalance problem with medical data sets”, Computers in Biology and Medicine 40 (2010) 509–518.

[26] Enhong Che, Yanggang Lin, Hui Xiong, Qiming Luo, Haiping Ma,” Exploiting probabilistic topic models to improve text categorization under class imbalance”, Information Processing and Management 47 (2011) 202–214.

[27] Alberto Fernández, María José del Jesus, Francisco Herrera,” On the 2-tuples based genetic tuning performance for fuzzy rule based classification systems in imbalanced data-sets”, Information Sciences 180 (2010) 1268–1291.

[28] Z. Chi, H. Yan, T. Pham, Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition, World Scientific, 1996.

[29] H. Ishibuchi, T. Yamamoto, T. Nakashima, Hybridization of fuzzy GBML approaches for pattern classification problems, IEEE Transactions on System, Man and Cybernetics B 35 (2) (2005) 359–365.

[30] J. Burez, D. Van den Poel,” Handling class imbalance in customer churn prediction”, Expert Systems with Applications 36 (2009) 4626–4636.

[31] Che-Chang Hsu, Kuo-Shong Wang, Shih-Hsing Chang,” Bayesian decision theory for support vector machines: Imbalance measurement and feature optimization”, Expert Systems with Applications 38 (2011) 4698–4704.

[32] Alberto Fernández, María José del Jesus, Francisco Herrera,” On the influence of an adaptive inference system in fuzzy rule based classification systems for imbalanced data-sets”, Expert Systems with Applications 36 (2009) 9805–9812.

[33] Jordan M. Malof, Maciej A. Mazurowski, Georgia D. Tourassi,” The effect of class imbalance on case selection for case-based classifiers: An empirical study in the context of medical decision support”, Neural Networks 25 (2012) 141–145.

[34] A. Asuncion D. Newman. (2007). UCI Repository of Machine Learning Database (School of Information and Computer Science), Irvine, CA: Univ. of California [Online]. Available: http://www.ics.uci.edu/∼mlearn/MLRepository.html

[35] J.R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.

[36] T. Jo and N. Japkowicz, "Class Imbalances versus Small Disjuncts," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 40-49, 2004.

[37] N. Japkowicz, "Class Imbalances: Are We Focusing on the Right Issue?" Proc. Int'l Conf. Machine Learning, Workshop Learning from Imbalanced Data Sets II, 2003.

[38] R.C. Prati, G.E.A.P.A. Batista, and M.C. Monard, "Class Imbalances versus Class Overlapping: An Analysis of a Learning

[39] System Behavior," Proc. Mexican Int'l Conf. Artificial Intelligence, pp. 312-321, 2004.

[40] G.M. Weiss, "Mining with Rarity: A Unifying Framework," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7-19, 2004.

[41] N. Chawla, K. Bowyer, and P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[42] Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.

[43] Mohamed Bekkar and Dr. Taklit Akrouf Alitouche, 2013. Imbalanced Data Learning Approaches Review. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.4, July 2013

[44] J. R. Quinlan, C4.5: Programs for Machine Learning, 1st ed. San Mateo, CA: Morgan Kaufmann Publishers, 1993.