Enhanced Classification to Counter the Problem of Cluster Disjuncts

Syed Ziaur Rahman[#1] , Dr.G. Samuel Vara Prasad Raju[*2],

[#]*Research Scholar, Department of CS & SE, Andhra University, Andhra Pradesh , INDIA*
[#]*Professor, Department of CS & SE, Andhra University, Andhra Pradesh , INDIA*

***Abstract*— This paper presets a rigorous yet practical model dubbed as Cluster Disjunct Minority Oversampling Technique (CDMOTE) for learning from skewed training data. This algorithm provides a simpler and faster alternative by using cluster disjunct concept. We conduct experiments using fifteen UCI data sets from various application domains using five algorithms for comparison on six evaluation metrics. The empirical study suggests that CDMOTE have been believed to be effective in addressing the class imbalance problem.**

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

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. 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 [1], [2]. 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.

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 [4-5]. 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 [6-7].

The existence of within-class imbalances is closely intertwined with the problem of small disjuncts, which has been shown to greatly depreciate classification performance [6-7]. 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 [4-5], [7]. 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 [4-5], [7]. 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.

II. LITERATURE REVIEW

In this section, we first review the major research about clustering in class imbalance learning and explain why we choose oversampling as our technique in this paper.

Chumphol Bunkhumpornpat et al. [8] have proposed a hybrid algorithm which uses the preexisting technique of DBSCAN to find clusters based on density and then uses SMOTE algorithm to generate synthetic instances along a shortest path from each positive instance to a pseudo centroid of a minority-class cluster. Matías Di Martino et al. [9] have presented a new classifier developed specially for imbalanced problems, where maximum F-measure instead of maximum accuracy guides the classifier design.

V. Garcia et al. [10] have investigated the influence on different resampling techniques used for balancing the imbalance data. María Dolores Pérez-Godoy et al. [11] have proposed an evolutionary framework for imbalance learning which uses both radial-basis function and the evolutionary cooperative-competitive technique.

Der-Chiang Li et al. [12] have suggested a strategy which both under-samples and oversamples the minority class and

the majority class respectively. 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.

Enhong Che et al. [13] 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. Alberto Fernández et al. [14] 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 named the Chi et al.'s rule generation. The second approach consists of a Fuzzy Hybrid Genetic Based Machine Learning (FH-GBML) algorithm.

J. Burez et al. [15] 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. [16] 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 [17] 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. [18] 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.

III. Cluster Disjunct Minority Oversampling Technique (CDMOTE)

In this section, we first briefly introduce the framework 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 on every cluster disjunct, which removes examples suffering from missing values at first and then removes borderline examples and examples of outlier category. The algorithm 1: CDMOTE 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 method. 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. 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 algorithm is summarized as below.

Algorithm 1 CDMOTE

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:

1: begin
2: $k \leftarrow 1, j \leftarrow 1$.
3: Apply Visualization Technique on subset *N*,
4: Identify cluster disjunct *Cj from* N, j= number of cluster disjunct identified in visualization
5: repeat
6: k=k+1
7: Identify and remove the borderline and outlier instances for the cluster disjunct *Cj*.
8: Until k = j

Phase II: Over sampling Phase

9: Apply Oversampling on *Cj* cluster disjunct from N,
10: repeat
11: k=k+1
12: Generate '*Cj* × s' synthetic positive examples from the minority examples in each cluster disjunct *Cj*.
13: Until k = j

Phase III: Validating Phase

14: Train and Learn A Base Classifier (C4.5) using *P* and *N*
15: end

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

3.1 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.

3.2 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.

3.3 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.

3.4 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 as the base algorithm.

IV. 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),

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

The Precision measure is computed by equation (2),

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{_____ (2)}$$

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

$$F - measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad \text{_____ (3)}$$

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

$$TruePositiveRate = \frac{TP}{TP + FN} \quad \text{_____ (4)}$$

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

$$TrueNegativeRate = \frac{TN}{TN + FP} \quad \text{_____ (5)}$$

V. EXPERIMENTAL FRAMEWORK

In this study CDMOTE is applied to fifteen binary data sets from the UCI repository [19] 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, 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 algorithm, we compared it with the traditional C4.5, CART (Classification and Regression trees), FT (Functional Trees), REP (Reduced Error Pruning Tree) 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

VI. RESULTS

For all experiments, we use existing prototype's present in Weka [20]. We compare the following domain adaptation methods:

We compared proposed method CDMOTE with the C4.5 [21], CART, FT, REP [22] and SMOTE [23] 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 learning algorithms. Second, we compare the classification performance of our proposed CDMOTE algorithm 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 C4.5, CART, FT and REP. we also analyze a pre-processing method SMOTE for performance evaluation of CDMOTE. The complete table of results for all the algorithms used in this study is shown in Table 4 to 8, 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, as it obtains the highest value among all algorithms.

FT, REP and SMOTE and a '○' indicates a loss of CDMOTE method on above said algorithms. The results in the tables show that CDMOTE has 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. A two-tailed corrected resampled paired t-test [46] 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.

Finally, we can say that CDMOTE is 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 methods to be the best performing algorithms when compared with five classical and well-known algorithms: C4.5, CART, FT, REP and a well-established pre-processing technique SMOTE.

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

Datasets	C4.5	CART	FT	REP	SMOTE	CDMOTE
Breast	0.606±0.087●	0.587±0.110●	0.586±0.102●	0.578±0.116●	0.717±0.084○	0.705±0.082

Breast_w	0.957±0.034●	0.950±0.032●	0.977±0.017○	0.957±0.030●	0.967±0.025●	0.973±0.018
Colic	0.843±0.070●	0.847±0.070●	0.802±0.073●	0.844±0.067●	0.908±0.040○	0.900±0.042
Credit-g	0.647±0.062●	0.716±0.055●	0.650±0.075●	0.705±0.054●	0.778±0.041●	0.788±0.041
Diabetes	0.751±0.070●	0.743±0.071●	0.793±0.072●	0.754±0.060●	0.791±0.041●	0.836±0.046
Heart-c	0.769±0.082●	0.810±0.074●	0.843±0.084○	0.806±0.077●	0.830±0.077○	0.822±0.077
Heart-h	0.775±0.089●	0.775±0.088●	0.852±0.078●	0.822±0.074●	0.904±0.054○	0.869±0.065
Heart-stat	0.786±0.094●	0.791±0.094●	0.864±0.075○	0.780±0.089●	0.832±0.062○	0.822±0.076
Hepatitis	0.668±0.184●	0.563±0.126 ●	0.757±0.195●	0.619±0.149●	0.798±0.112●	0.848±0.136
Ionosphere	0.891±0.060●	0.896±0.059●	0.900±0.060●	0.902±0.054●	0.904±0.053●	0.949±0.041
Kr-vs-kp	0.998±0.003	0.997±0.004●	0.996±0.005●	0.998±0.002	0.999±0.001○	0.998±0.002
Labor	0.726±0.224●	0.750±0.248●	0.971±0.075○	0.767±0.232●	0.833±0.127●	0.870±0.126
Mushroom	1.000±0.000	0.999±0.001	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.952±0.040●	0.954±0.043●	0.990±0.014●	0.967±0.030●	0.962±0.025●	0.992±0.012
Sonar	0.753±0.113●	0.721±0.106●	0.771±0.103●	0.746±0.106●	0.814±0.090○	0.854±0.086

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

Datasets	C4.5	CART	FT	REP	SMOTE	CDMOTE
Breast	0.753±0.042○	0.728±0.038○	0.745±0.051○	0.721±0.037○	0.710±0.075●	0.713±0.059
Breast_w	0.965±0.026●	0.968±0.026●	0.988±0.019○	0.965±0.030●	0.974±0.025●	0.986±0.021
Colic	0.851±0.055●	0.853±0.053●	0.845±0.060●	0.857±0.056●	0.853±0.057●	0.864±0.059
Credit-g	0.767±0.025●	0.779±0.030●	0.776±0.033●	0.765±0.025●	0.768±0.034●	0.799±0.044
Diabetes	0.797±0.045●	0.782±0.042●	0.793±0.037●	0.785±0.037●	0.781±0.064●	0.862±0.050
Heart-c	0.783±0.076●	0.792±0.080●	0.825±0.080○	0.780±0.075●	0.779±0.082●	0.808±0.087
Heart-h	0.824±0.071●	0.829±0.073●	0.849±0.058●	0.814±0.064●	0.878±0.076●	0.894±0.072
Heart-stat	0.799±0.051●	0.791±0.083●	0.833±0.078●	0.772±0.079●	0.791±0.081●	0.821±0.094
Hepatitis	0.510±0.371●	0.232±0.334●	0.604±0.271●	0.293±0.386●	0.709±0.165●	0.739±0.200
Ionosphere	0.895±0.084●	0.868±0.096●	0.906±0.080●	0.886±0.092●	0.934±0.049●	0.945±0.047
Kr-vs-kp	0.994±0.006	0.993±0.007●	0.991±0.008●	0.988±0.009●	0.996±0.005○	0.994±0.005
Labor	0.696±0.359●	0.715±0.355●	0.915±0.197●	0.698±0.346●	0.871±0.151●	0.921±0.148
Mushroom	1.000±0.000	0.999±0.002	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.992±0.005●	0.992±0.005●	0.997±0.003○	0.990±0.005●	0.983±0.007●	0.996±0.004
Sonar	0.728±0.121●	0.709±0.118●	0.764±0.119●	0.733±0.134●	0.863±0.068○	0.851±0.090

Table 6 Summary of tenfold cross validation performance for TP Rate (Recall) (Sensitivity) on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	CDMOTE
Breast	0.947±0.060○	0.926±0.081○	0.815±0.095●	0.917±0.087○	0.763±0.117●	0.861±0.101
Breast_w	0.959±0.033○	0.952±0.034○	0.962±0.029○	0.957±0.033○	0.947±0.035●	0.950±0.033
Colic	0.931±0.053○	0.932±0.050○	0.835±0.077●	0.914±0.066●	0.913±0.058●	0.915±0.058
Credit-g	0.847±0.036○	0.869±0.047○	0.783±0.052○	0.872±0.057○	0.810±0.058○	0.733±0.057
Diabetes	0.821±0.073○	0.848±0.066○	0.868±0.065○	0.838±0.072●	0.712±0.089●	0.763±0.070
Heart-c	0.808±0.085○	0.835±0.091○	0.837±0.100○	0.813±0.108○	0.777±0.110●	0.802±0.102
Heart-h	0.885±0.081○	0.856±0.087●	0.876±0.089○	0.868±0.084○	0.815±0.084○	0.783±0.107
Heart-stat	0.824±0.104○	0.832±0.113○	0.857±0.090○	0.830±0.109○	0.803±0.110○	0.794±0.102
Hepatitis	0.374±0.256●	0.169±0.236●	0.573±0.248○	0.187±0.239●	0.681±0.188●	0.700±0.247
Ionosphere	0.821±0.107●	0.830±0.112●	0.820±0.114●	0.826±0.104●	0.881±0.071●	0.946±0.054
Kv-rs-kp	0.995±0.005	0.995±0.006	0.990±0.007●	0.993±0.007●	0.995±0.006	0.995±0.006
Labor	0.640±0.349●	0.665±0.359●	0.885±0.234○	0.665±0.334●	0.765±0.194●	0.823±0.227
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	0.995±0.004○	0.996±0.003○	0.995±0.004○	0.996±0.004○	0.990±0.005●	0.993±0.004
Sonar	0.721±0.140●	0.652±0.137●	0.757±0.136●	0.685±0.192●	0.865±0.090●	0.893±0.109

Table 7 Summary of tenfold cross validation performance for TN Rate (Specificity) on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	CDMOTE
Breast	0.260±0.141●	0.173±0.164●	0.335±0.166●	0.151±0.164●	0.622±0.137○	0.464±0.169

Breast_w	0.932±0.052○	0.940±0.051●	0.977±0.037○	0.931±0.060●	0.975±0.024●	0.984±0.025
Colic	0.717±0.119●	0.720±0.114●	0.734±0.118●	0.731±0.121●	0.862±0.063○	0.841±0.080
Credit-g	0.398±0.085●	0.421±0.102●	0.469±0.098●	0.371±0.105●	0.713±0.056●	0.772±0.063
Diabetes	0.603±0.111●	0.554±0.113●	0.574±0.095●	0.567±0.105●	0.807±0.077●	0.873±0.054
Heart-c	0.723±0.119●	0.729±0.121●	0.779±0.117●	0.717±0.119●	0.861±0.068○	0.809±0.099
Heart-h	0.655±0.158●	0.672±0.162●	0.714±0.131●	0.636±0.152●	0.894±0.074○	0.893±0.079
Heart-stat	0.728±0.131●	0.717±0.135●	0.775±0.123●	0.677±0.152●	0.862±0.064○	0.812±0.115
Hepatitis	0.900±0.097○	0.928±0.094○	0.882±0.092●	0.942±0.093○	0.837±0.109●	0.888±0.097
Ionosphere	0.940±0.055●	0.921±0.066●	0.949±0.046○	0.933±0.063●	0.928±0.057●	0.947±0.047
Kv-rs-kp	0.993±0.007●	0.992±0.008●	0.990±0.009●	0.987±0.010○	0.998±0.003○	0.994±0.006
Labor	0.865±0.197●	0.877±0.192●	0.945±0.131○	0.843±0.214●	0.847±0.187●	0.928±0.138
Mushroom	1.000±0.000	0.999±0.002	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
Sick	0.875±0.071●	0.876±0.078●	0.974±0.026○	0.846±0.080●	0.872±0.053●	0.970±0.031
Sonar	0.749±0.134●	0.756±0.121●	0.752±0.148●	0.762±0.145●	0.752±0.113●	0.831±0.113

Table 8 Summary of tenfold cross validation performance for Accuracy on all the datasets

Datasets	C4.5	CART	FT	REP	SMOTE	CDMOTE
Breast	74.28±6.05○	70.22±5.19	67.21±7.28●	68.99±5.51●	69.83±7.77●	70.23±5.91
Breast_w	95.01±2.73●	94.74±2.60●	96.75±2.00	94.79±2.71●	96.16±2.06	96.58±1.79
Colic	85.16±5.91●	85.37±5.41●	79.78±6.57●	84.64±5.53●	88.53±4.10○	87.92±4.70
Credit-g	71.25±3.17●	73.43±4.00●	68.91±4.46●	72.18±3.31●	76.50±3.38○	75.06±3.89
Diabetes	74.49±5.27●	74.56±5.01●	76.55±4.67●	74.39±4.37●	76.08±4.04●	81.75±4.08
Heart-c	76.94±6.59●	78.68±7.43●	81.02±7.25○	76.92±7.36●	82.99±4.98○	80.57±6.55
Heart-h	80.22±7.95●	79.02±7.18●	81.81±6.20●	78.46±6.52●	85.65±5.46○	83.56±5.81
Heart-stat	78.15±7.42●	78.07±8.58●	82.07±6.88○	76.19±6.68●	83.89±5.05○	80.31±7.75
Hepatitis	79.22±9.57●	77.10±7.12●	81.90±8.38●	78.62±7.13●	78.35±9.09●	83.59±9.65
Ionosphere	89.74±4.38●	88.87±4.84●	90.26±4.97●	89.49±4.58●	90.28±4.73●	94.64±3.74
Kv-rs-kp	99.44±0.37	99.35±0.43	99.02±0.54	99.01±0.55	99.66±0.27	99.45±0.42
Labor	78.60±16.58●	80.03±16.67●	92.40±11.07○	78.10±17.29●	80.27±11.94●	88.33±11.09
Mushroom	100.0±0.00	99.95±0.09	100.0±0.000	99.98±0.08	100.0±0.00	100.0±0.00
Sick	98.72±0.55●	98.85±0.54●	99.26±0.04	98.68±0.58●	97.61±0.68●	99.07±0.50
Sonar	73.61±9.34●	70.72±9.43●	75.46±9.92●	72.55±10.10●	82.42±7.25●	86.23±8.31

Table 4, 5, 6, 7 and 8 reports the results of AUC, Precision, F-measure, TP Rate, TN Rate and accuracy respectively for fifteen UCI datasets. The bold dot '●' indicates a win of CDMOTE method on C4.5, CART,

VII. CONCLUSIONS

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 as its base learners. Experimental results show that CDMOTE has performed well in the case of multi class imbalance datasets. Furthermore, CDMOTE is much less volatile than C4.5.

In our future work, we will apply CDMOTE 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] 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.

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

[3] 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.

[4] 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.
[5] Siti Khadijah Mohamada, Zaidatun Tasir. "Educational data mining: A review", Procedia - Social and Behavioral Sciences 97 (2013) 320 – 324.
[6] 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.
[7] 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.
[8] Chumphol Bunkhumpornpat, Krung Sinapiromsaran, Chidchanok Lursinsap." DBSMOTE: Density-Based Synthetic Minority Over-sampling Technique" Appl Intell (2012) 36:664–684.
[9] Matías Di Martino, Alicia Fernández, Pablo Iturralde, Federico Lecumberry." Novel classifier scheme for imbalanced problems", Pattern Recognition Letters 34 (2013) 1146–1151.
[10] 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.
[11] 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.
[12] 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.
[13] 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.
[14] 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.
[15] J. Burez, D. Van den Poel," Handling class imbalance in customer churn prediction", Expert Systems with Applications 36 (2009) 4626–4636.
[16] 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.
[17] 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.
[18] 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.
[19] 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.
[20] Witten, I.H. and Frank, E. (2005) Data Mining: Practical machine learning tools and techniques. 2nd edition Morgan Kaufmann, San Francisco.
[21] J. R. Quinlan, C4.5: Programs for Machine Learning, 1st ed. San Mateo, CA: Morgan Kaufmann Publishers, 1993.
[22] J.R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81-106, 1986.
[23] N. Chawla, K. Bowyer, and P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.
[24] T. Jo and N. Japkowicz, "Class Imbalances versus Small Disjuncts," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 40-49, 2004.
[25] 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.
[26] R.C. Prati, G.E.A.P.A. Batista, and M.C. Monard, "Class Imbalances versus Class Overlapping: An Analysis of a Learning System Behavior," Proc. Mexican Int'l Conf. Artificial Intelligence, pp. 312-321, 2004.
[27] G.M. Weiss, "Mining with Rarity: A Unifying Framework," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 7-19, 2004.
[28] 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.