1K2-IOS-1b-4

A New Classification for Multiclass Imbalanced Datasets Based on Clustering Approach

Wanthanee Prachuabsupakij[*1] Nuanwan Soonthornphisaj[*2]

Department of Computer Science, Faculty of Science, Kasetsart University, Thailand

The new approach called C-MIEN - Clustering with hybrid sampling approaches for Multiclass Imbalanced classification using Ensemble models is proposed in this paper to improve the performance of classifier for multiclass imbalanced datasets without the decomposition method. We focus on the multiclass imbalance problem because this problem can be found in many real world applications. Multiclass imbalance problem occurs when the number of instances of the one class is much higher than in the remaining classes in the dataset. The aim of this paper is to develop a resampling approach that can effectively classify multiclass imbalanced datasets. Firstly, K-means is used to split the set of instances into two clusters. For each cluster, hybird sampling methods are used. Then, final training sets are used to build an emsemble. Finally, the prediction is obtained by combining the results from both clusters through a majority vote. We have conducted experiments on many multiclass datasets from the UCI. We use different classifiers in order to observe the performance and suitability of our purpose within each classifier. We carry out the experimental study with the several well-known algorithms such as Decision Trees, Naïve bayes, and K-Nearest Neighbors (k=1,3). The performance is measured based on G-mean and F-measure. The experimental results show that C-MIEN achieved higher performance than state-of-the-art methods. Moreover, the empirical results show that C-MIEN algorithm is a practical algorithm since it can be applied to many classifiers. C-MIEN attain better overall performance on Decision tree classifier compared to Naïve Bayes classifier.

Keywords: Multiclass Imbalanced Dataset, Clustering, Sampling, Classification, Data mining

1. Introduction

The classification based on the imbalanced training datasets is one of the most widely found problems in data mining, machine learning domains. In two-class imbalanced dataset, the problem occurs when the number of instances of the one class (majority/negative class) hugely outnumbers another class (minority/positive class). The classification on imbalanced data always causes problems because traditional classification algorithms tend to be overwhelmed by the majority class and ignore the minority class. The result is that predictions based on the majority class have a high possibility to get good performance, whereas the predictions based on minority classes generally have poor performance results, because most classifiers operate on data drawn from the same distribution as the training data. Therefore, the prediction of minority class is more significant in those cases than the prediction of majority class. Datasets obtained from many real-world applications are highly imbalanced, such as text categorization, bioinformatics[Batuwita et al., 2009], intrusion detection and fraud detection. Most methods solved the two-class imbalance problem such as [Benjamin et al., 2008; Yen et al., 2009; Chen et al., 2010]. We found that the problems of multiclass classification on imbalanced data are found in few studies.

The solutions of the class imbalanced problem have been proposed both at the data level [Han et al., 2005; Yen et al., 2009] and algorithm level [Tian et al., 2011; Wang et al., 2011]. The data level is usually based on the resampling method, which deals with the distribution of dataset before the classification process. Resampling method includes oversampling and undersampling. The algorithm level aims to adjust algorithm itself by provide searching for the unequal weights for the minority and majority classes in the training process to force the classifier to recognize the minority class.

In multiclass imbalance problems, the classification is even more complicated. Moreover, the higher degree of class imbalance may increase the difficulty of multiclass classification. There are multiple ways to solve multiclass imbalanced problems. One way is to decompose the multiclass dataset into a series of binary classification problems and then use a two-class learner for classification such as One-Against-One (OAO) [Tian et al., 2011], One-Against-All (OAA) [Chen et al., 2010]. Several decomposition methods use ensemble approach to combine the models obtained from the binary class classifiers. Ensemble models have been more attention because it can average prediction errors and reduce bias and variance of errors. However, using decomposition with sampling technique is not practical for this problem because they are time consuming. Another way is to adjust the original classifiers to multiclass and imbalance cases [Sun, 2006; Shuo et al., 2009; Navarro et al., 2011]. Imbalance learning on multiclass problem is more interesting than two-class problem because this problem can be found in many real world applications. We found that there is very few reported work in literature addressing the multiclass imbalance problem. In spite of that it is a serious problem in data mining.

The aim of this paper is to improve the classification performance based on the multiclass imbalanced datasets. In this paper, we introduce a new resampling approach based on Clustering with hybrid sampling approaches for Multiclass Imbalanced classification using Ensemble models (C-MIEN),

[*1] E-mail address: wanthanee.pk@gmail.com

[*2] Corresponding author, Email address: fscinws@ku.ac.th

which does not apply decomposition technique. The proposed method is a data level approach. We proposed the data level approach because it will not be limited in algorithms themselves. C-MIEN uses the clustering approach to create a new training set for each cluster and apply two resampling technique to rebalance the class distribution. C-MIEN improves the classification performance based on the multiclass imbalanced datasets in three ways. Firstly, k-means is used to split the set of instances into two clusters. Then, for each cluster, two resampling techniques (oversampling and undersampling) are applied on the the training set in order to balance the class distribution. Finally, ensemble approaches are used to combine the models obtained with our method through a majority vote from both clusters.

C-MIEN has been applied from our previous work named KSMOTE [Prachuabsupakij et al., 2012]. KSMOTE operates by decision trees algorithm. Meanwhile, in this paper, we carry out the experimental study with the several well-known algorithms such as Decision Trees, Naïve bayes, and K-Nearest Neighbors (k=1, 3) using ensemble methods. We have conducted experiments on many multiclass datasets from the UCI. These datasets consist of two types of class distribution; high and low. We use different classifiers in order to observe the performance and suitability of our purpose within each classifier. The performance is measured based on G-mean and F-measure.

The experimental results show that C-MIEN achieved better performance than baseline algorithms. Moreover, the empirical results show that C-MIEN algorithm is a practical algorithm since it can be applied to multiclass imbalanced datasets. The characteristic of C-MIEN is to create two new training sets that consist of the new label of instances with similar characteristics. This step is applied to reduce the number of classes then the simpler problem can be easily solved by C-MIEN.

In summary, we proposed a new approach which consists of the contribution as follow:

- We develop a new algorithm to reduce the complexity of multiclass imbalance classification. It creates the new training sets with similar characteristic instances based on clustering approach.
- We carefully design the experiments and analyze the behavior of C-MIEN to demonstrate that our method can be applied to several well-known algorithms when the datasets are multiclass imbalanced problems.

The rest of the paper is organized as follows: Section 2 presents some of the approaches previously applied to deal with the class imbalance problem. Section 3 describes our approach while Section 4 we describe our benchmark datasets, the experimental and report on the experimental results. Finally, Section 5 is the conclusion.

2. Related work

2.1 The class imbalance problem

Recently, many real world applications have the imbalanced class distribution problem such as text classification, fraud detection, information retrieval and so on. Researchers have proposed many classifiers to solve this problem such as decision tree, k-nearest, Naïve Bayes , and Support Vector Machines. They found that the performance of the majority classes are high whereas the prediction performance of the minority classes tends to be low, nevertheless the prediction of minority class is more significant in some domains.

Two different approaches to solve the class imbalance problem are data level and algorithm level methods. Data level methods aim to solve problems by manipulate the distribution of a training set, including over-sampling and under-sampling methods. Both methods decrease the overall level of class imbalance. Sometimes this can involve a combination of the two methods. Algorithm level methods adapt existing learning algorithms to pay more attention to the minority classes. In this study, we will focus on the class imbalance problem at the data level methods.

Oversampling reduces the degree of imbalanced distribution by increase the size of minority class either by duplicates or interpolates minority instances. The basic oversampling method is random oversampling (ROS). It balances the class distribution by randomly duplicates minority instances into the minority class. SMOTE is one of the famous oversampling methods by Chawla [Chawla et al., 2002]. SMOTE produce synthetic minority class instances by interpolating between minority examples that lie together. It makes the decision regions larger towards majority class and less specific. Synthetic examples are introduced along the line segment between each minority class example and one of its k minority class nearest neighbors. SMOTE reduce the imbalanced class distribution without causing overfitting as shown in many studies [Chawla et al., 2002; Chawla et al., 2003]. Jo *et al.* [Jo et al., 2004] proposed cluster-based oversampling algorithm. It creates the independent clusters from the minority and majority classes, and then randomly does oversampling for each of the majority clusters, except the largest cluster. This is done with replacement until all of the majority clusters contain the same number of instances as the largest cluster. The algorithm then oversamples each of the minority clusters with replacement until the number of instances in each minority cluster is equal to the number of instances in a majority cluster after oversampling divided by the number of minority clusters.

On the other hand, undersampling is supposed to reduce the number of instance from the majority class in order to achieve a more balanced class distribution. The simple undersampling method is random undersampling (RUS). It randomly discard instances of a majority class until the ratio between the minority and majority class is at the desired level. Another method uses partitioning and various techniques to break the majority class into n disjoint partitions, and combining the n models to make a final classification by Yan *et al* [Yan et al., 2003]. Yen *et al.* proposed a cluster-based undersampling to determine the number of selected majority class samples in each cluster by using expression, and then randomly select the majority class samples in each cluster. Then, the algorithm selects the majority class samples randomly from each cluster and combines them with the minority class samples to form a new dataset.

2.2 Ensemble classifiers for imbalanced datasets

In recent years, ensemble learning is primarily used to improve the performance of the imbalanced classification [Opitz et al., 1999; Sun, 2007; Lin et al., 2009; Yun et al., 2010; Tian et al.,

2011]. Two well-known ensemble methods are Bagging [Breiman, 1996] and Boosting [Freund et al., 1996], which are very successful in improving the accuracy of the certain classifiers. In imbalanced problems, there are several methods that combine both ensemble learning algorithms and resampling techniques.

Chawla *et al* proposed the SMOTEBoost algorithm [Chawla et al., 2003]. In each iteration of boosting, it utilize SMOTE to add the new minority class and increase the sampling weights for the minority class instance. Another method is SMOTEBagging [Shuo et al., 2009]. This method combines three popular resampling methods; undersampling, oversampling, and SMOTE; into the ensemble model based on Bagging for diversity analysis. Yang *et al* presented the EnSVM [Yang et al., 2010]. This algorithm concerned with improving the performance of the Support Vector Machines (SVMs) on imbalanced datasets. It integrates two types of sampling methods by starting with oversampling the minority class to a moderate extent. For undersampling, it uses the bootstrap sampling approach. The size of the new majority class is the same as that of the minority class obtained from SMOTE. The ensemble of SVMs is employed to boost the performance.

2.3 Multiclass classification in imbalanced datasets

Many researchers focus on the imbalanced dataset concentrated for two-class classification [Shengguo et al., 2009; Yen et al., 2009; Geiler et al., 2010; Nguwi et al., 2010; Yang et al., 2010]. In case of multiclass datasets, there are two or more minority classes with respect to one majority class. Therefore, this problem can be solved in multiple ways. One typical way is the decomposition techniques, which decompose the multi-class classification into several binary classifications such as One-Against-One and One-Against-All. However, some two-class techniques were not useful when applied to multiclass problem directly, especially in the case of imbalanced datasets [Zhu, 2007].

Consider the decomposition method, there are some methods that combine both resampling and binary classification approaches. One of these methods was introduced by Fernndez et al.[Fernández et al., 2010]. They applied an over-sampling step before the pair-wise learning process. The quality of this method can be tested using the linguistic fuzzy rule based classification system and fuzzy hybrid genetics-based machine learning algorithm.

There are not many works addressing the genuinely imbalance multi-class problem. One of these approaches uses a dynamic over-sampling method that incorporated into a memetic algorithm (MA) and uses RBFNNS as the classification model [Navarro et al., 2011]. The authors propose two different methodologies which add an over-sampling method: the static smote radial basis function (SSRBF) and the dynamic smote radial basis function (DSRBF). DSRBF modifies the oversampling procedure within the learning process. AdaC2.M1 [Sun, 2006] develops a cost-sensitive boosting algorithm to improve the classification performance of imbalanced data involving multiple classes. AdaC2.M1 extended the original AdaBoost [Freund et al., 1997] and AdaC2 [Sun et al., 2005] algorithms to multiclass cases. Another method was proposed by Shuo *et al*.[Shuo et al., 2009], who explores the impact of diversity on each class and overall performance. They combine undersampling, oversampling and SMOTE methods into ensemble model based on both two-class and multiclass datasets. In multiclass dataset, the algorithm controls the resampling rate with *a%*. It refers to sampling rate of majority class and other classes. Therefore, for each class, i^{th}, algorithm resample instances with replacement at the rate of $(N_C/N_i)a\%$, where N_C is the number of class of majority class and the i^{th} class has N_i number of training instances.

3. C-MIEN

In this section, we present a new sampling method based on clustering approach called C-MIEN. The algorithm does clustering with hybrid sampling approaches for Multiclass Imbalanced datasets using Ensemble method. C-MIEN aims to improve the performance of multiclass learning from an imbalanced dataset. In case of multiclass imbalanced dataset, it is more difficult to define the majority and minority classes. Therefore, we want to reduce the number of classes in the multiclass training set without the decomposition method. The training patterns include three steps. The first step is a reclustering process using the k-mean algorithm. The main idea is that the multiclass dataset is divided into small subsets based on clustering approach. Due to the use of the clustering method, most instances for each subset seem to have similar characteristics. Second, we apply two resampling techniques in order to rebalance the class distribution called rebalancing process. The benefit of doing two resampling methods in C-MIEN is mitigating the overfitting and information loss problems.

The final step is to train base classifiers independently on every subset of the new training dataset for each cluster and combine all models using a new majority vote from both clusters. The main reason is that multiple classifiers are expected to be more robust and satisfactory than a single classifier. The framework of C-MIEN is shown in Fig. 1 and the C-MIEN algorithm is detailed in Table 1.

3.1 Reclustering Process

Given the multiclass training set, the reclustering process is performed using k-means algorithm. The main idea of this step is to split all instances into certain number of clusters fixed a priori. C-MIEN divides all instances into two clusters by setting *k* to be 2. In order to measure the distance between two instances, we use the Euclidean distance, which is a useful measure of distance and also the simplest.

In this step, we assume that applying the clustering algorithm on multiclass datasets may improve the performance of the resulting classifier; because the members in the cluster have similar characteristics. Moreover, split training sets can also decrease complicated sampling in multiclass dataset.

Table 1 The pseudo-code for C-MIEN algorithm.

Algorithm 1: C-MIEN

Input:

1) Given S $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ $x_i \in X$, with labels $y_i \in Y = \{1\ldots L\}$
2) y_{mk} = the over majority class of S_k, Ny_i = the number of instances of class y_i
3) k = number of classes for the cluster (k=2)
4) IR = imbalance ratio between y_{Lk} and remaining classes of S_k

Begin:

1) C= Kmeans(S,k)
2) let C_1 = cluster1 ,C_2 = cluster2
3) **for each** classLabel y_i
4) if(Ny_i in C_1) >(Ny_i in C_2) then xy_i is assigned to C_1
5) else xy_i is assigned to C_2
6) **end for**
7) temp = C_2 //temp contains all instances in cluster, C_2
8) **for** k = 1 to 2
9) R_k ,T_k = ∅
10) **for each** x_i in C_k
11) x_i = relabel(x_i)
12) $R_k = R_k \cup x_i$
13) **end for**
14) $S_k = R_k \cup$ temp
15) temp = C_1
16) **for** each class i_k do
17) **If** IR > 1.5 then y_{ik}^{new}= SMOTE(y_{ik}) else $y_{ik}^{new} = y_{ik}$
18) $T_k = T_k \cup y_{ik}^{new}$
19) **end** for
20) y_{Lk}^{new} = Random undersampling(y_{Lk}) with d instances
21) $T_k = T_k \cup y_{mk}^{new}$
22) **for** j = 1 to M do
23) h_{kj} = BaseClassifier(T_k)
24) **end for**
25) **end for**

Output: The output hypothesis H* is calculated as follows:
if majority vote of $h_1 = R_2$ then
H* = majority vote of h_2
else H* = majority vote of h_1

Considering the instances for each cluster, let Ny_i denotes the number of data instances of class y_i in training dataset. Let C_1 and C_2 denote the first cluster and the second cluster respectively. If Ny_i in C_1 is greater than Ny_i in C_2 then all instances of class y_i in both clusters are assigned to C_1. On the other hand, if Ny_i in C_2 is greater than Ny_i in C_1 then all instances of class y_i in both clusters are assigned to C_2. Consequently, we get two set of the samples, which are different classes. After that, the classes from both clusters are combined using relabeling of classes group. For example, classes (Y_1, Y_2 and Y_3) in the first cluster are combined with all classes in the second cluster that were relabeled as same label (R_2). Meanwhile, classes (Y_4, and Y_5) in the second cluster are combined with all classes in the first cluster that were relabeled (R_1) as well. The output of this process is two set of new sample, S_1 and S_2. Note that, S_1 consists of classes Y_1, Y_2, Y_3, and R_2. On the other hand S_2 consists of classes Y_4, ,Y_5 and R_1. Table 2 shows an example on pageblocks dataset, which is adjusted in the reclustering process.

Table 2 An example of the reclustering approach on pageblock dataset.

Clusters	Classes	Number of Instances per class	Combined and relabled classes		
			Sample sets	Classes	Number of Instances per class
1 (C_1)	text	4913	1 (S_1)	text	4913
	graphic	28		graphic	28
	picture	115		picture	115
	total	*5056*		R_1	*417*
2 (C_2)	horiz	329	2 (S_2)	horiz	329
	vert	88		vert	88
	total	*417*		R_2	*5056*

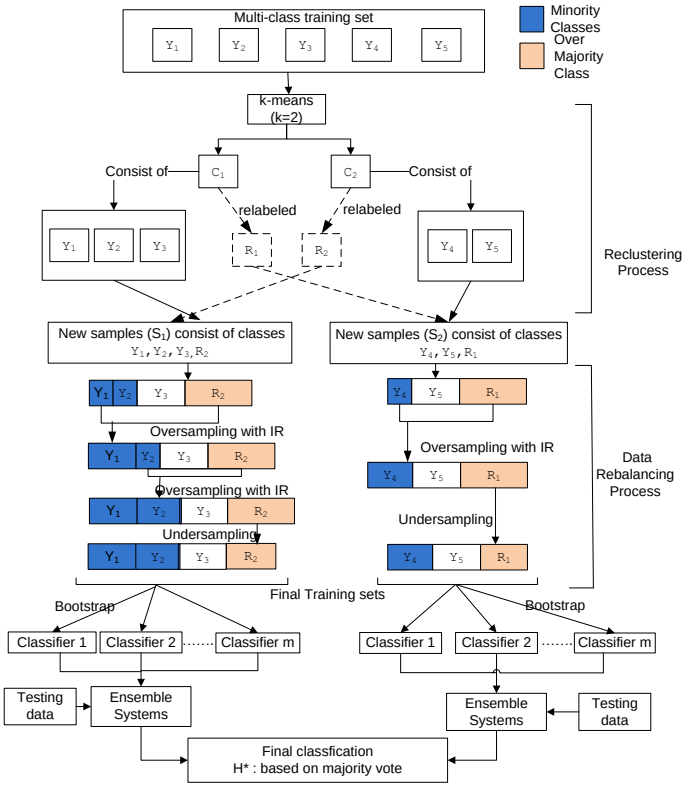


Fig.1 The framework of C-MIEN algorithm (suppose that there are five classes)

3.2 Data Rebalancing Process

After the re-clustering process is finished, the data rebalancing process is started. (We get two set of new sample, S_1 and S_2.) In order to rebalance the class distribution, we integrate two sampling techniques, SMOTE and random under sampling. The

Table 3 Characteristics of 7 datasets used for experimentation.

Datasets	Size	No. of Features	No. of Classes	Over majority (OMa)	Under minority (UMi)	Name of OMa	Name of Omi	Number of instances per class	IR
balance-scale	625	4	3	288	49	L	R	288, 49, 288	5.88
car1	1728	7	4	1210	65	unacc	v-good	1210, 384, 69, 65	18.62
ecoli	336	8	8	143	2	cp	imL	143, 77, 52, 35, 20, 5, 2, 2	71.50
glass	214	10	6	76	9	two	six	70, 76, 17, 13, 9, 29	8.44
new-thyroid	215	6	3	150	30	1	3	150, 35, 30	5.00
page-blocks	5473	11	5	4913	28	text	graphic	4913, 329, 28, 88, 115	175.46
yeast	1484	9	10	463	5	CYT	ERL	463, 429, 244, 163, 51, 44, 37, 30, 20, 5	92.60

benefit of SMOTE is to alleviates the overfitting problem.

Given a training dataset in *k* cluster (*k*=2) $\{x_i,y_i\}$, i = 1 to n, where $x_i \in X$ is the i^{th} instance and $y_i \in \{1,2,\ldots,L\}$ is the i^{th} class label. X_{yi} is all instances of class y_i. These classes are sorted by the number of classes. Therefore, N_{yL} is the number of instances of class having the largest number of instances called "*over majority class*". For other classes, the typical of class may be either a minority class or a normal class depending on the class distribution. The minority class with the smallest number of instances is called "*under minority class*". Suppose that there are H minority classes, SMOTE algorithm is produced in C-MIEN (L-H) times.

For each cluster, the process starts by sorting the number of classes. Then, over majority class is defined. Considering resampling rate in multiclass cases, we use imbalance ratio (IR) [Orriols-Puig et al., 2009; Fernndez et al., 2010]. The imbalanced ratio is defined as the fraction between the number of instances of the over majority and the minority classes (N_{yL} / N_{yi}). If other classes ($y_i \ldots y_{L-1}$) contain a value of IR higher than 3 (a distribution of 75-25%), the oversampling method is applied for instances of class y_{ik}. We got new synthetic instances of class y_{ik} (y_{ik}^{new}). This step is called "firstover". After that, the imbalanced ratio of firstover is examined. In case that its value is higher than 1.5 (a distribution of 60-40 %), x_{yi} are interpolated by doing oversampling, which is "lastover" step. For the final training set (T_k), we use random undersampling technique to reduce *d* instances of the over majority class, where *d* is the different number of instances among the under minority class obtained from the rebalancing process (firstover and lastover steps). Therefore, we got totally two new final training sets from both clusters (T_1 and T_2).

3.3 Classification process using ensemble methods

In this process, we improve the performance of classifier using an ensemble approach that can reduces the variance and/or bias of a set of classifiers. It has been demonstrated in many studies. Moreover, generalization ability of sampling technique with a single classifier is always unsatisfactory and robustness are often poor [Haifeng et al., 2010]. In this way, we build 3 classifiers as ensemble members for each cluster. We get totally six hypotheses from two clusters.

The prediction is done using majority vote method from all hypotheses of two clusters. In this paper, we present new implementations of the majority vote. The detail of majority vote are implemented as follows: Given a test example, if the final prediction obtained from the majority vote among three hypotheses of S_1 is equal to R_2 then the classification is depend on the majority vote of hypotheses of S_2. Otherwise, the prediction will rely on the majority vote of three hypotheses of S_1. In our experiments reported in Section 5, we set M to be 3, and the class of a instance is assigned through the majority vote from two clusters.

4. Experiments

4.1 Benchmark Datasets and experimental design

C-MIEN is applied to 7 datasets taken from the UCI Repository for Machine Learning [Asuncion et al., 2007]. Four of them are highly imbalanced datasets whereas the imbalance ratios of the rest three datasets are low. The selected datasets are multiclass problems and different numbers of instances, features and classes. The four datasets with highly imbalance are car, ecoli, page-blocks and yeast. Besides, three datasets with low imbalance ratios are balance-scale, glass, and new-thyroid. All datasets were separated into two clusters to reduce sampling complexity. Since the aim of C-MIEN is to show the effectiveness of C-MIEN to improve the performance of multiclass learning from an imbalanced datasets, the datasets selected have a considerable on imbalance rate and the number of classes. Table 3 lists the information of each dataset.

The experimental software was developed based on WEKA 3.6.0 framework [Witten et al., 2005]. All experiments, the parameters were optimized using a 10-fold cross validation strategy. Euclidean distance was used to compute the distance between instances and cluster in the k-means algorithm. For each cluster, the number of iterations of the ensemble method is equal to 3 (M).

4.2 Algorithms used for the study

In the empirical study, we have selected several well-known Machine Learning algorithms as base classifiers including decision tree, k-nearest neighbor, and naïve bayes. These algorithms can be applied to solve the multiclass problems without decomposition techniques. Nevertheless, these traditional learning algorithms do not perform well on imbalance problems especially those which are multiclass dataset.

We chose these algorithms because they are commonly used, and represent completely different learning mechanisms.

Table 4 Parameter specification for base classifiers employed in our experiment.

Classifiers	Parameters
Decision tree	Prune = true Confidence level = 0.25 Minimum number of item-sets per leaf = 2
KNN	K= 1, K=3 Nearest neighbor search algorithm = LinearNNSearch Distance metric = Euclidean distance

Table 5 Confusion matrix for a classifier in the multi-class classification..

Actually Class	**Predicted class**			
	C_1	C_2	……	C_k
C_1	n_{11}	n_{12}	……	n_{1k}
C_2	n_{21}	n_{22}	……	n_{2k}
.	.	.	.	.
.	.	.	.	.
C_k	n_{k1}	n_{k2}	……	n_{kk}

Table 6 Measures for multi-class classification using the notation of Table 4 [Sun, 2006].

Measure	**Formula**	**Detail**
Precision_i (P_i)	$\frac{n_{ii}}{\sum_{j=1}^{k} n_{ji}}$	Precision of class C_i
Recall_i (R_i)	$\frac{n_{ii}}{\sum_{j=1}^{k} n_{ij}}$	Recall of class C_i
F_i-measure	$\frac{2\,R_i P_i}{R_i + P_i}$	F-measure of class C_i
G-mean	$\left(\prod_{i=1}^{k} R_i\right)^{1/k}$	G-mean of recall values of every classes

Decision tree is an induction approach for building classification model. K-nearest neighbor is an instance-based learning algorithm and Naïve bayes is an incremental learning algorithm. Moreover, these classifiers are very popular and are applied to solve the imbalanced problems. However, in order to estimate the performance of C-MIEN, we study the effects of our method with various different classifier algorithms.

The configuration parameters for the base classifiers are shown in Table 4. For KNN algorithm, we considered two configurations, the first is the one nearest neighbor and the second is the three nearest neighbors, so we analyzed them as two different base classifiers 1NN and 3NN. For Naïve bayes classifier, we set the default parameters from WEKA for implementation.

The C-MIEN method is compared to different algorithms:

- The single baseline algorithms without re-sampling data (C4.5, 1NN, 3NN, and NB).
- The single baseline algorithm with oversampling using SMOTE (SC4.5, S1NN, S3NN, and SNB).
- Ensemble of baseline algorithms (EC4.5, E1NN, E3NN, and ENB).
- Ensemble of baseline algorithm with oversampling using SMOTE (ESC4.5, ES1NN, ES3NN, and ESNB).

We chose SMOTE algorithm because it generally shows better performance than new intelligent sampling approaches as shown in many previous works. [Han et al., 2005; Seiffert et al., 2010]. For ensemble approach, it is chosen because it was used to solve many imbalanced data problems [Yan et al., 2003; Benjamin et al., 2008]. Moreover, KNN, NB and C4.5 algorithms are sensitive to the amount of negative training examples.

4.3 Evaluation measures

In our experiments, we used two evaluation measures: F-measure, and Geometric mean (G-mean). Since we focus on multi-class classification, the confusion matrix has been applied as shown in Table 5. Where C_i denotes the class label of the i^{th} class and k is the number of classes. The evaluation measure of multi-class classification was proposed by Y.Sun [Sun, 2006] as shown in Table 6. Kubat *et al* [Kubat et al., 1998] suggested to use the G-mean as the geometric means of recall values of two classes. In multiclass cases, Sun *et al* [Sun, 2006] defined G-mean of recall values of every classes as shown in Table 6. As each recall value representing the classification performance of a specific class is equally accounted, G-mean is capable to measure the balanced performance among classes of a classification output. The G-mean and F-measure value are in the [0, 1] range. If G-mean value is equal to 1, it means that all minority class instances are identified. On the other hand, if its value is equal to 0, it means that none of the minority class instances are predicted correctly.

4.4 Results and Analysis

In this subsection, the C-MIEN method is compared to baseline algorithms and oversampling approaches on seven datasets. The purpose of this section is to show that incorporating the sampling approaches before the learning algorithm can improve the performance of classifiers in multiclass imbalanced dataset, especially in the class which is more difficult to classify. Moreover, to show the effectiveness of C-MIEN, we perform the experimental study with four algorithms; C4.5, 1NN, 3NN, NB.

4.4.1 Effectiveness of C-MIEN on C4.5

We first compare the performance of C-MIEN with C4.5 method. Table 7 and Table 8 show the performance of each method in terms of the F-measure and G-mean respectively. The results show that C-MIEN mitigate the imbalanced data problem and achieves the highest F-measure score on all datasets. These results indicate that C-MIEN is the most suitable algorithm for multiclass imbalanced datasests when C4.5 is used as a base classifier. Consider on Table 8, we found that G-mean values of C-MIEN are superior to other methods on most of the datasets. Meanwhile, the G-mean value of SC4.5 seems to provide better result on ecoli dataset. Analyzing the results on Table 7 and Table 8, C-MIEN obtains a mean of F-measure values of 0.90 while the mean of F-measure values of the other classifiers rage between 0.71 and 0.78. In addition, the mean of G-mean values of C-MIEN is equal to 0.93. It is better than ESC4.5, which got 0.86 on mean of G-mean value, which is the second-best performance. On balance_scale, C4.5 obtains 0.00 on G-mean value, whereas up C4.5 with oversampling method improves G-mean value a bit (24.67%). However, C-MIEN achieve the

Table 7 F-measure comparison among four methods and C-MIEN on Decision tree classifier (C4.5).

Datasets	C4.5	SC4.5	EC4.5	ESC4.5	C-MIEN
balance-scale	0.7497	0.7975	0.7835	0.8130	**0.8415**
car1	0.8585	0.9301	0.8577	0.9154	**0.9831**
ecoli	0.8360	0.9153	0.8350	0.9221	**0.9342**
glass	0.7090	0.7035	0.7176	0.6820	**0.8248**
new-thyroid	0.9469	0.9644	0.9477	0.9642	**0.9772**
page-blocks	0.9845	0.9778	0.9723	0.9798	**0.9995**
yeast	0.5288	0.4956	0.5437	0.5147	**0.9244**

Table 8 G-mean comparison among four methods and C-MIEN on Decision tree classifier (C4.5).

Datasets	C4.5	SC4.5	EC4.5	ESC4.5	C-MIEN
balance-scale	0.0000	0.7844	0.2467	0.7989	**0.8250**
car1	0.9153	0.9713	0.9143	0.9592	**0.9872**
ecoli	0.9534	0.9689	0.9519	**0.9707**	0.9611
glass	0.7845	0.8232	0.7933	0.8051	**0.8785**
new-thyroid	0.9094	0.9715	0.9051	0.9720	**0.9762**
page-blocks	0.9174	0.9829	0.9186	0.9839	**0.9996**
yeast	0.6953	0.6849	0.6997	0.6934	**0.9615**

Table 9 F-measure comparison among four methods and C-MIEN on K-nearest neighbor classifier (1NN).

Datasets	1NN	S1NN	E1NN	ES1NN	C-MIEN
balance-scale	0.8420	0.7990	0.8210	0.8060	**0.8475**
car1	0.8795	0.9013	0.8497	0.8876	**0.9543**
ecoli	0.9318	0.9206	0.9351	0.9264	**0.9353**
glass	0.7324	0.7064	0.7177	0.7138	**0.8389**
new-thyroid	0.9720	0.9480	**0.9830**	0.9750	0.9825
page-blocks	0.9794	0.9580	0.9792	0.9565	**0.9962**
yeast	0.4956	0.5027	0.4956	0.5104	**0.9065**

Table10 G-mean comparison among four methods and C-MIEN on K-nearest neighbor classifier (1NN).

Datasets	1NN	S1NN	E1NN	ES1NN	C-MIEN
balance-scale	0.0000	0.8655	0.0000	0.8598	**0.8648**
car1	0.9433	0.9495	0.9098	0.9329	**0.9634**
ecoli	0.0000	0.9491	0.0000	0.9546	**0.9579**
glass	0.8040	0.8158	0.7934	0.8177	**0.8834**
new-thyroid	0.9617	**0.9794**	0.9466	0.9742	0.9722
page-blocks	0.8742	0.9680	0.8729	0.9669	**0.9977**
yeast	0.6628	0.6676	0.6677	0.6744	**0.9462**

best performance in term G-mean on balance_scale dataset (82.50%).

4.4.2 Effectiveness of C-MIEN on 1NN

The results of F-measure and G-mean values with 1NN classifier are summarized in Table 9 and Table 10 respectively. The confidence of the base classifier is always 1 for the predicted class. Therefore the results using VOTE are completely equivalent. The results show that the F-measure of C-MIEN is the best F-measure performance on most datasets except for E1NN which outperforms C-MIEN on new-thyroid dataset. (which has the lowest imbalance ratio of 5.) On new-thyroid dataset, C-MIEN is a bit less than E1NN (the difference is equal to 0.005). From Table 9, we found that using a simple oversampling and oversampling with ensemble methods are not the best strategy to solve the multiclass imbalance data in term of 1NN as a classifier. Consider the page-block dataset, which has the highest imbalance ratio (175.46). We found that, the F-measure of C-MIEN is higher than other methods (1NN, S1NN, E1NN, and ES1NN) about 1.68%, 3.82%, 1.7%, and 3.97% respectively. These results of highly imbalance ratio confirm that SMOTE and SMOTE with Ensemble algorithms do not improve the classifier performance when 1NN as a classifier. Table 10 reveals that C-MIEN obtained the best G-mean results on most of datasets compared to other algorithms. These results indicate that C-MIEN can make correct prediction on the minority class efficiently than other methods. However, SMOTE obtains the highest G-mean value on new-thyroid dataset. Analyzing from Table 9 and Table 10, we found that the mean of both measures of C-MIEN (0.92) is superior to other methods (0.76). On new-thyroid dataset, both F-measure and G-mean values of C-MIEN is less than E1NN and S1NN respectively. Consider on balance-scale and ecoli datasets, G-mean values of both 1NN and E1NN methods are equal to 0.00, these results indicate that both methods predict very poor on the minority class instances.

4.4.3 Effectiveness of C-MIEN on 3NN

We conducted another experiment using 3 nearest neighbors and found that the performance of C-MIEN is better than other methods (see Table 11, 12)**.** Consider the glass dataset which has a large number of classes and features but it has a few number of instances per class, we found that C-MIEN outperforms other methods. C-MIEN obtains 81.93%, whereas S3NN got 72.68% on F-measure value which is the second-best performance. In term of F-measure, the results reveal that C-MIEN obviously improves the performance of classification on multiclass imbalanced dataset.

Table 12 presents the performance of all methods measured in term of G-mean. From Table 12, the performance of C-MIEN is better than state-of-the-art methods in most imbalanced datasets. However, S3NN obtained the best result on balance-scale dataset, which the imbalance ratio is low (5.88). On car dataset, G-mean value of C-MIEN is greatly better than other methods. Therefore, it confirms that, C-MIEN can perform very well on the minority class instances with car dataset. Consider on three datasets including glass, new-thyroid, and yeast, classification based on ensemble method cannot improve the performances of baseline classifier. On the other hand, ensemble method slightly reduces the classification performance. Consider on ecoli dataset, 3NN and E3NN obtain 0.00 on G-mean value like balance-scale

Table 11 F-measure comparison among four methods and C-MIEN on K-nearest neighbor classifier (3NN).

Datasets	3NN	S3NN	E3NN	ES3NN	C-MIEN
balance-scale	0.8420	0.8640	0.8280	0.8570	**0.8685**
car1	0.9450	0.8240	0.8497	0.9192	**0.9745**
ecoli	0.8450	0.9300	0.8400	0.9410	**0.9430**
glass	0.7247	0.7268	0.7162	0.7215	**0.8193**
new-thyroid	0.9340	0.9733	0.9340	0.9733	**0.9785**
page-blocks	0.9794	0.9540	0.9791	0.9530	**0.9948**
yeast	0.5222	0.5263	0.5332	0.5188	**0.9041**

Table 12 G-mean comparison among four methods and C-MIEN on K-nearest neighbor classifier (3NN).

Datasets	3NN	S3NN	E3NN	ES3NN	C-MIEN
balance-scale	0.0000	**0.8644**	0.0760	0.8572	0.8573
car1	0.7066	0.7693	0.7777	0.6666	**0.9656**
ecoli	0.0000	0.9263	0.0000	0.9104	**0.9458**
glass	0.8022	0.8375	0.7966	0.8242	**0.8692**
new-thyroid	0.8856	0.9731	0.8780	0.9731	**0.9752**
page-blocks	0.8487	0.9648	0.8511	0.9632	**0.9969**
yeast	0.7265	0.7311	0.7003	0.6851	**0.9455**

Table 13 F-measure comparison among four methods and C-MIEN on Naïve bayes classifier (NB).

Datasets	NB	SNB	ENB	ESNB	C-MIEN
balance-scale	**0.8674**	0.7924	0.8591	0.7713	0.8485
car1	0.8422	**0.8872**	0.8456	0.8868	0.7305
ecoli	0.9524	0.9488	**0.9534**	0.9492	0.9390
glass	0.5524	0.3419	0.5496	0.3907	**0.6439**
new-thyroid	0.9676	0.9524	0.9728	0.9538	**0.9765**
page-blocks	0.9152	0.7368	0.9100	0.7381	**0.9290**
yeast	0.6147	0.5417	0.6137	0.5583	**0.9029**

Table 14 G-mean comparison among four methods and C-MIEN on Naïve bayes classifier (NB).

Datasets	NB	SNB	ENB	ESNB	C-MIEN
balance-scale	**0.9462**	0.8932	0.9375	0.8876	0.8119
car1	0.8028	**0.8625**	0.7949	0.8621	0.7011
ecoli	0.9590	0.9749	0.9597	**0.9750**	0.9695
glass	0.6458	0.5330	0.6454	0.5828	**0.7561**
new-thyroid	0.9528	0.9664	0.9559	0.9674	**0.9766**
page-blocks	0.8287	0.8403	0.8315	0.8413	**0.9162**
yeast	0.7522	0.6539	0.7502	0.6735	**0.9417**

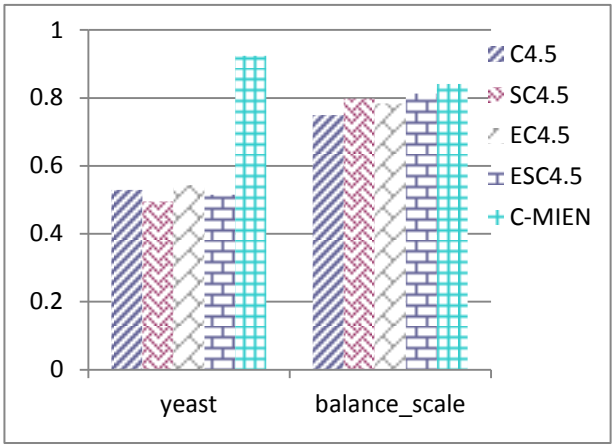
dataset, G-mean value of 3NN is equal to 0.00. These results show that 3NN without oversampling methods are not the good predictive model for classification on the minority class instances in case of both datasets.

4.4.4 Effectiveness of C-MIEN on NB

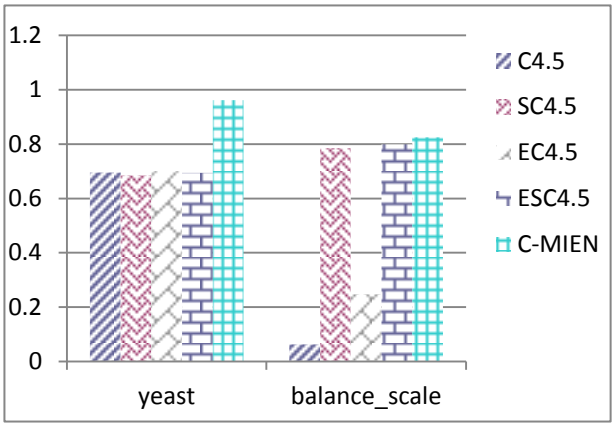
In this subsection, the performance of C-MIEN is compared to other methods based on NB classifier as shown in Table 13 and Table 14. The results in Table 13 show that C-MIEN performs better than other algorithms in terms of F-measure in four out of seven datasets. There are three datasets that the F-measure of C-MIEN is lower than other methods including balance-scale, car, and ecoli. Consider on Table 14, we found that G-mean values of C-MIEN are superior to other methods in four out of five datasets as well. We suspected that these outcomes occur because the types of attribute in these datasets are categorical attributes. In addition, C-MIEN cannot work well in these datasets due to attribute dependency when NB is used as a base classifier. However, C-MIEN gets the mean of F-measure and G-mean values of 0.78, while the mean of its values of NB, SNB, ENB, and ESNB are 0.73, 0.69, 0.73, and 0.70 respectively. Consider on yeast dataset, which is a highly imbalance ratio and a largest number of classes. Adjustment the NB classifier with basic oversampling and ensemble methods cannot improve the performance of NB classifier. On the other hand, these methods also reduce the efficiency of NB classifier on yeast dataset.

4.4.5 Discussions on overall results

In this section, we present the overall results of C-MIEN that can improve the performance of classifier in multiclass imbalanced dataset. The experimental results show that C-MIEN achieves the best performances in all datasets in terms of F-measure when C4.5 and 3NN are used as base classifiers. However, C-MIEN with NB classifier obtains the lowest mean value in both measures because most datasets have attribute dependency. Therefore, using NB as a classifier will not perform well. In addition, G-mean results show that the prediction performance of C-MIEN is better than state-of-the-art methods in the minority class. Moreover, we compare the performance of C-MIEN with other methods in term of imbalance ratio. We first select the yeast and balance-scale datasets which have very different imbalance ratios and number of classes. Yeast dataset has the maximum amount of classes (10 classes) and has high imbalance ratio (92.60). On balance-scale dataset, it has the lowest number of classes (3 classes) and has low imbalance ratio (5.88) as well. We finally compare the performance in terms of F-measure and G-mean when C4.5 is used as classifier, and the results are shown in Fig 2. In case that the imbalance ratio is low, the performance results of C-MIEN on the balance scale dataset is similar to other methods. However, our values are highest in the baseline C4.5 method. On yeast dataset, the imbalance ratio is high as 92.60. It is clear that, the performance of C-MIEN is more sharply than other methods. The resulting classifications on other classifiers (1NN, 3NN, and NB) are not different as well. The results show that C-MIEN outperforms the current multiclass imbalanced data problem solving methods especially in case high multiclass and high imbalance ratio datasets. In addition, we also make the following observation on the data studied.



a) F-measure



b) G-mean

Fig 2. Comparison of F-measure and G-mean of C4.5, SC4.5, EC4.5, ESC4.5 and proposed method for yeast and balance-scale datasets based on C4.5 classifier.

- Using SMOTE to oversampling the minority class instances is better than a baseline without oversampling on the most of the datasets.
- Between the ensemble and SMOTE, SMOTE performs better in multiclass imbalanced dataset.
- C-MIEN outperforms SMOTE and SMOTE with ensemble methods in the most of datasets and the most of base classifiers especially when the dataset has high imbalance ratio and large the number of classes.
- The best base classifiers for C-MIEN are 3NN and C4.5, whereas C-MIEN with NB classifiers provide the lowest mean of both values.

5. Conclusion

In this research, we proposed a new resampling approach to learn from multiclass imbalanced dataset based on clustering and hybrid sampling approaches. In our approach, K-means is employed to separate all the instances into two clusters in order to reduce complicated sampling in multiclass dataset. After that, hybrid sampling approaches is used to reduce the degree of imbalanced distribution in sub training set. Finally, we proposed to use ensembles of several well-known classifiers; Decision tree, K-nearest neighbor, and Naïve bayes; to enhance the prediction performance. The findings from several UCI multiclass imbalanced datasets indicate that C-MIEN is very promising.

From the experimental result, we found that C-MIEN obtained a big improvement of classification performance when the dataset has high imbalance ratio and large the number of classes. In addition, the results reveal that choosing the best base classifiers is an important issue to increase the performance of C-MIEN algorithm. We found that the best base classifiers for C-MIEN in this study are 3NN and C4.5. 3NN provides the highest mean of F-measure score, whereas C4.5 provides the highest mean of G-mean value. Therefore, C-MIEN with C4.5 and 3NN classifiers make an impressive improvement in prediction performance, not only for the minority class, but also for the majority class. On the other hand, C-MIEN with NB classifiers provide the lowest mean of both values.

In particular, we found that C-MIEN is a practical algorithm for multiclass imbalanced datasets in three ways: (1) the performance of base classifier can be improved by C-MIEN in both minority and majority classes. (2) C-MIEN reduce the complexity of decision regions in multiclass takes using clustering and relabeling approaches (3) C-MIEN inherits the strength of two strategies; ensemble and SMOTE. Therefore, it reduces variance/bias and alleviates the overfitting problems. In future work, we will apply C-MIEN to other datasets.

Acknowledgement

This research is supported by the Department of Computer Science, Faculty of Science, Kasetsart University and National Science and Technology Development Agency under Ministry of Science and Technology of Thailand.

References

Asuncion A., & Newman D.). UCI machine learning repository. Retrieved December 15,2010., from http://archive.ics.uci.edu/ml/datasets.html, 2007.

Batuwita R., & Palade V., *A New Performance Measure for Class Imbalance Learning. Application to Bioinformatics Problems*. Paper presented at the Machine Learning and Applications, 2009. ICMLA '09. International Conference on, 545-550, 2009.

Benjamin W., & Nathalie J., Boosting Support Vector Machines for Imbalanced Data Sets. *Foundations of Intelligent Systems, 4994*, 38-47, 2008.

Breiman L., Bagging predictors. *Mach. Learn., 24*(2), 123-140, 1996.

Chawla N. V., Bowyer K. W., Hall L. O., & Kegelmeyer W. P., SMOTE: synthetic minority over-sampling technique. *Journal artificial intelligence research, 16*(1), 321-357, 2002.

Chawla N. V., Lazarevic A., Hall L. O., & Bowyer K. W., *SMOTEBoost: improving prediction of the minority class in boosting*. Paper presented at the the Principles of Knowledge Discovery in Databases, PKDD-2003, Cavtat-Dubrovnik,Croatia, 107-119, 2003.

Chen S., He H., & A. G. E., RAMOBoost: Ranked Minority Oversampling in Boosting. *IEEE Transactions on Neural Networks, 21*(10), 1624-1642, 2010.

Fernández A., Jesús M. J. d., & Herrera F., *Multi-class imbalanced data-sets with linguistic fuzzy rule based*

classification systems based on pairwise learning. Paper presented at the Proceedings of the Computational intelligence for knowledge-based systems design, and 13th international conference on Information processing and management of uncertainty, 2010.

Fernndez A., Jesus M. J. D., & Herrera F., *Multi-class imbalanced data-sets with linguistic fuzzy rule based classification systems based on pairwise learning*. Paper presented at the Proceedings of the Computational intelligence for knowledge-based systems design, and 13th international conference on Information processing and management of uncertainty, 2010.

Freund Y., & Schapire R. E., *Experiments with a New Boosting Algorithm*. Paper presented at the In proceeding of the 13th. Internation Conference on Machine Learning, 148-156, 1996.

Freund Y., & Schapire R. E., A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences, 55*(1), 119-139, 1997.

Geiler O. J., Hong L., & jian G. Y., *An Adaptive Sampling Ensemble Classifier for Learning from Imbalanced Data Sets*. Paper presented at the International MultiConference of Engineers and Computer Scientist, Hong Kong, 2010.

Ghanem A. S., Venkatesh S., & West G., *Multi-class Pattern Classification in Imbalanced Data*. Paper presented at the Proceedings of the 2010 20th International Conference on Pattern Recognition, 2010.

Haifeng S., Bingru Y., Yun Z., Wu Q., & Bing A., *The problem of classification in imbalanced data sets in knowledge discovery*. Paper presented at the Computer Application and System Modeling (ICCASM), 658-661, 2010.

Han H., Wang W.-Y., & Mao B.-H.,Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning, 2005.

Jo T., & Japkowicz N., Class imbalances versus small disjuncts. *SIGKDD Explor. Newsl., 6*(1), 40-49, 2004.

Kubat M., Holte R. C., & Matwin S., Machine Learning for the Detection of Oil Spills in Satellite Radar Images. *Mach. Learn., 30*(2-3), 195-215, 1998.

Lin S.-C., Chang Y.-c. I., & Yang W.-N., Meta-learning for imbalanced data and classification ensemble in binary classification. *Neurocomput., 73*(1-3), 484-494, 2009.

Murphey Y. L., Wang; H., Ou; G., & Feldkamp L. A., *OAHO: an Effective Algorithm for Multi-Class Learning from Imbalanced Data*. Paper presented at the International Joint Conference on Neural Networks, 406 - 411, 2007.

Navarro F. F., Martinez C. H., & Gutierrez P. A., A dynamic over-sampling procedure based on sensitivity for multi-class problems. *Pattern Recogn., 44*(8), 1821-1833, 2011.

Nguwi Y. Y., & Cho S. Y., An unsupervised self-organizing learning with support vector ranking for imbalanced datasets. *Expert Syst. Appl., 37*, 8303-8312, 2010.

Opitz D., & Maclin R., Popular Ensemble Methods: An Empirical Study. *Journal of Artificial Intelligence Research, 11*, 169-198, 1999.

Orriols-Puig A., & Bernadó-Mansilla E., Evolutionary rule-based systems for imbalanced data sets. *Soft Computing - A Fusion of Foundations, Methodologies and Applications, 13*(3), 213-225, 2009.

Prachuabsupakij W., & Soonthornphisaj N.,Clustering and Combined Sampling Approaches for Multi-class Imbalanced Data Classification, Advances in Information Technology and Industry Applications. Springer Berlin Heidelberg, 2012.

Seiffert C., Khoshgoftaar T. M., Van Hulse J., & Napolitano A., RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on, 40*(1), 185-197, 2010.

Shengguo H., Yanfeng L., Lintao M., & Ying H., *MSMOTE: Improving Classification Performance When Training Data is Imbalanced*. Paper presented at the, 13-17, 2009.

Shuo W., & Xin Y., *Diversity analysis on imbalanced data sets by using ensemble models*. Paper presented at the Computational Intelligence and Data Mining, 2009. CIDM '09. IEEE Symposium on, 324-331, 2009.

Sun Y., *Boosting for learning multiple classes with imbalanced class distribution*. Paper presented at the In 2006 IEEE International Conference on Data Mining 2006.

Sun Y. *Cost-Sensitive Boosting for Classification of Imbalanced Data*. University of Waterloo, Canada, 2007.

Sun Y., Wong A. K. C., & Wang Y., *Parameter inference of cost-sensitive boosting algorithms*. Paper presented at the Proceedings of the 4th international conference on Machine Learning and Data Mining in Pattern Recognition, 2005.

Tian J., Gu H., & Liu W., Imbalanced classification using support vector machine ensemble. *Neural Computing &Applications, 20*(2), 203-209, 2011.

Wang S., & Yao X. *Ensemble diversity for class imbalance learning.* University of Birmingham, 2011.

Witten I. H., Frank E., & Hall M. A., *Data Mining: Practical Machine Learning Tools and Techniques* (Third Edition ed.), San Francisco: Morgan Kaufmann, 2005.

Yan R., Liu Y., Jin R., & Hauptmann A., *On Predicting Rare Classes With Svm Ensembles In Scene Classification* Paper presented at the IEEE International conference on Acoustics, Speech and Signal Processing, 2003.

Yang L., Xiaohui Y., Xiangji H. J., & Aijun A., Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. *Information Processing & Management, In Press, Corrected Proof*, 2010.

Yen S.-J., & Lee Y.-S., Cluster-based under-sampling approaches for imbalanced data distributions. *Expert Syst. Appl., 36*(3), 5718-5727, 2009.

Yun Z., Bingru Y., Nan M., & Da R., *New construction of Ensemble Classifiers for imbalanced datasets*. Paper presented at the Intelligent Systems and Knowledge Engineering (ISKE), 2010 International Conference on, 228-233, 2010.

Zhu X., *Lazy Bagging for Classifying Imbalanced Data*. Paper presented at the Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, 2007.