Addressing the Class Imbalance Problem in Medical Datasets

M. Mostafizur Rahman and D. N. Davis

***Abstract*—A well balanced dataset is very important for creating a good prediction model. Medical datasets are often not balanced in their class labels. Most existing classification methods tend to perform poorly on minority class examples when the dataset is extremely imbalanced. This is because they aim to optimize the overall accuracy without considering the relative distribution of each class. In this paper we examine the performance of over-sampling and under-sampling techniques to balance cardiovascular data. Well known over-sampling technique SMOTE is used and some under-sampling techniques are also explored. An improved under sampling technique is proposed. Experimental results show that the proposed method displays significant better performance than the existing methods.**

***Index Terms*—Class imbalance, under-sampling, over-sampling, clustering, SMOTE.**

I. Introduction

A balanced dataset is very important for creating a good training set. Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. They aim to optimize the overall accuracy without considering the relative distribution of each class [1]. Typically real world data are usually imbalanced and it is one of the main causes for the decrease of generalization in machine learning algorithms [2]. Conventional learning algorithms do not take into account the imbalance of class. They give the same attention to the majority class and the minority class. When the imbalance is massive, it is hard to build a good classifier using conventional learning algorithms [3]. The cost in miss-predicting minority classes is higher than that of the majority class for many class imbalanced datasets; this is particularly so in medical datasets where high risk patients tend to be the minority class. Therefore, there is a need of a good sampling technique for medical datasets.

Sampling strategies have been used to overcome the class imbalance problem by either eliminating some data from the majority class (under-sampling) or adding some artificially generated or duplicated data to the minority class (over-sampling) [4].

Over-sampling techniques [5] increase the number of minority class members in the training set. The advantage of over-sampling is that no information from the original training set is lost since all members from the minority and majority classes are kept. However, the disadvantage is that the size of the training set is significantly increased [5]. If the time taken to resample is not considered, under-sampling betters over-sampling in terms of time and memory complexity [1].

Random over-sampling is the simplest approach to over-sampling, where members from the minority class are chosen at random; these randomly chosen members are then duplicated and added to the new training set [6]. Chawla[5] proposed an over-sampling approach called SMOTE in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with duplicated real data entries. SMOTE blindly generates synthetic minority class samples without considering majority class samples and may thus cause overgeneralization [7]. Over-sampling may cause longer training time and over-fitting. Drummond and Holte[8] show that random under-sampling yields better minority prediction than random over-sampling.

The alternative to over-sampling is under-sampling. Under-sampling is a technique to reduce the number of samples in the majority class, where the size of the majority class sample is reduced from the original datasets to balance the class distribution. One simple method of under-sampling (random under-sampling) is to select a subset of majority class samples randomly and then combine them with minority class sample as a training set [7]. Many researchers have proposed some advanced way of under-sampling the majority class data.

According to Chyi[9] the under-sampling approach based on distance uses distinct modes: the nearest, the farthest, the average nearest, and the average farthest distances between Minority and Majority, as four standards to select the representative samples from the majority class. For every minority class sample in the dataset, the first method (''nearest") calculates the distances between all majority class samples and the minority class samples, and selects k majority class samples which have the smallest distances to the minority class sample. If there are *n* minority class samples in the dataset, the ''nearest" method would finally select $(k \times n)$ majority class samples $(k > 1)$. However with this method, some samples within the selected majority class samples might be duplicated. The ''farthest" method selects the majority class samples which have the farthest distances to each minority class sample. For every majority class sample in the dataset, the third method (''average nearest") calculates the average distances between one majority class sample and all minority class samples. This method selects the majority class samples which have the smallest average distances. The last method ''average farthest" is similar to the ''average nearest" method; it selects the majority class samples which have the farthest

Manuscript received March 3, 2013; revised April 29, 2013.

The authors are with Department of Computer Science, University of Hull, UK (e-mail: M.M.Rahman@2009.hull.ac.uk, D.N.Davis@hull.ac.uk).

DOI: 10.7763/IJMLC.2013.V3.307

average distances from all the minority class samples. These under-sampling approaches based on distance, expend a lot of time selecting the majority class samples in the large dataset, and they are not efficient in real applications [7].

In this paper we have used a SMOTE oversampling and cluster based under-sampling technique to balance cardiovascular data and compare the results.

II. Smote over Sampling

Chawla[5] proposed an over-sampling approach called SMOTE in which the minority class is over-sampled by creating "synthetic" examples rather than by over-sampling with duplicated real data entries. Depending upon the amount of over-sampling required, neighbours from the k nearest neighbours of a record are randomly chosen. Our implementation currently uses five nearest neighbours. The SMOTE algorithm is:

```
Algorithm SMOTE(T, N, k)
Input:  Number of minority class samples T;
        Amount of SMOTE N%;
        Number of nearest neighbors k
Output: (N/100)* T synthetic minority class samples
(∗ If N is less than 100%, randomize the minority classsamples as only a random percent of them will be SMOTEd. ∗)
ifN <100
thenRandomize the T minority class samples
        $T = (N/100) * T$
        $N = 100$
endif
$N = (int)(N/100)$
        (∗ Amount of SMOTE is in integral multiples of100. ∗)
k= Number of nearest neighbors
numattrs= Number of attributes
Sample[ ][ ]: array for original minority class samples
newindex: keeps a count of number of synthetic samples generated, initialized to 0
Synthetic[ ][ ]: array for synthetic samples
(∗ Compute k nearest neighbors for each minority class sample. ∗)
fori ←1 to T
        Compute k nearest neighbors for i,
                and save the indices inthe nnarray
        Populate(N, i, nnarray)
endfor
Populate(N, i, nnarray)
        (∗Function to generate the synthetic samples. ∗)
whileN != 0
        Choose a random number between 1 and k, call it nn.
        (*This step chooses one ofthe k nearest neighbors of i.*)
        for attr ←1 to numattrs
        Compute: dif= Sample[nnarray[nn]][attr] –
    Sample[i][attr]
        Compute: gap = random number between 0 and 1
Synthetic[newindex][attr] = Sample[i][attr] + gap ∗dif
endfor
        newindex++
        N = N -1
endwhile
return(∗ End of Populate. ∗)
```

III. Cluster Based Under-Sampling

Down-sizing the majority class results in a loss of information that may result in overly general rules [10]. In order to overcome this drawback of the under-sampling approach Yen and Lee (2009) proposed an unsupervised learning technique for supervised learning called cluster-based under-sampling. Their approach is to first cluster all the training samples into K clusters (they have run the experiment with different K values to observer the outcome) then chose appropriate training samples from the derived clusters. The main idea is that there are different clusters in a dataset, and each cluster seems to have distinct characteristics. If a cluster has more majority class samples and less minority class samples, it will behave like a majority class sample. On the other hand, if a cluster has more minority class samples and less majority class samples, it does not hold the characteristics of the majority class samples and behaves more like the minority class samples. Therefore, their approach selects a suitable number of majority class samples from each cluster by considering the ratio of the number of majority class samples to the number of minority class samples in the derived cluster [7]. They first cluster the full data to K clusters. A suitable number (*M*) of majority class samples from each cluster are then selected by considering the ratio of the number of majority class samples to the number of minority class samples in the cluster. The number *M* is determined by equation 1, and they randomly choose the *M* number of majority class samples from each cluster. In the i^{th} cluster ($1 \leq i \geq K$) the $Size_{MA}^{i}$ will be:

$$Size_{MA}^{i} = (\mathrm{m} \times Size_{MI}) \times \frac{Size_{MA}^{i} / Size_{MI}^{i}}{\Sigma_{i-1}^{K} Size_{MA}^{i} / Size_{MI}^{i}} \qquad (1)$$

This approach may be suitable for datasets where class labels are confidently defined and truly reflect the property of the labeled class. But as mentioned earlier that in some cases, especially for medical datasets, there is no guarantee that the given class labels are truly reflect the actual class of that record.

IV. Modified Cluster Based Under-Sampling Method

In this research the approach of Yen and Lee (2009) is taken and modified. The data is first separated in to two sets; one subset has all the majority class samples and the other subset has the entire minority class samples. The majority class samples are then separated into *K* clusters (*K*> 1), where each cluster is considered to be one subset of the majority class. The aim was not to make the majority and minority class ratio to 1:1; we just wanted to reduce the gap between the numbers of majority class samples to the numbers of minority class samples. All the subsets of the majority class are separately combined with the minority class samples to make *K* different training data sets (The K value is dependent on the data domain, in our implementation the K value was 3). All the combined datasets are classified using a decision tree[11] and a Fuzzy Unordered Rule Induction Algorithm[12]. The datasets with the highest accuracy over the majority of the classifiers were kept for further data mining processes.

For experiments several datasets were prepared using different clusters and then classified using decision tree. The experimental outcomes are discussed in the result section.

V. Experiments and Algorithms

Two cardiovascular datasets from Hull and Dundee clinical sites were used. K-Means [13] clustering is used to cluster the majority samples. For choosing the best subset, decision tree [11] and Fuzzy Unordered Rule Induction Algorithm [12]were used as classifiers. SM OTE is used as an oversampling technique.

A. Cardiovascular Data

The Hull site data includes 98 attributes and 498 cases of cardiovascular patients and the Dundee site data includes 57 attributes, and 341 cases from cardiovascular patients. After combining the data from the two sites, 26 matched attributes are left.

From these two data sets, a combined dataset having 26 attributes with 823 records was prepared. Out of 823 records, 605 records have missing values and 218 records do not have any missing values. The missing values are imputed by the machine learning technique proposed in previous work [14]. Among all the records 120 patients are dead and 703 patients are alive; a majority to minority ratio of 6:1. For this experiment according to clinical risk prediction model (CM1) [15], patients with status "Alive" are consider to be "Low Risk" and patients with status "Dead" are consider to be "High Risk".

B. Classifier Performance Evaluation

The performance of the classification is evaluated by accuracy (ACC); sensitivity (Sen); specificity (Spec) rates, and the positive predicted value (PPV) and negative predicted value (NPV), based on confusion matrix.

Assume that the cardiovascular classifier output set includes two typically risk prediction classes as: *"High risk",* and *"Low risk".* Each pattern x_i ($i=1,2..n$) is allocated into one element from the set (P, N) (positive or negative) of the risk prediction classes. Hence, each input pattern might be mapped onto one of four possible outcomes as Table I true positive- true high risk (TP)- when the outcome is correctly predicted as High risk; true negative- true low risk (TN)- when the outcome is correctly predicted as Low risk; false negative-false Low risk (FN)- when the outcome is incorrectly predicted as Low risk and in fact is High risk; or false positive- false high risk (FP) - when the outcome is incorrectly predicted as High risk and in fact is Low risk. The set of (P, N) and the predicted risk set can be built as a confusion matrix.

TABLE I: Confusion Matrix

		Predicted classes	
		High risk	Low risk
Expected/Actual Classes	High risk	TP	FN
	Low risk	FP	TN

The accuracy of a classifier is calculated by:

$$ACC = \frac{TP+TN}{TP+FP+TN+FN} \qquad (2)$$

The sensitivity is the rate of number correctly predicted "High risk" over the total number of correctly predicted "High risk" and incorrectly predicted "Low risk". It is given by:

$$Sen = \frac{TP}{TP+FN} \qquad (3)$$

The specificity rate is the rate of correctly predicted "Low risk" over the total number of expected/actual "Low risk". It is given by:

$$Spec = \frac{TN}{TN+FP} \qquad (4)$$

Higher accuracy does not always reflect a good classification outcome. For clinical data analysis it is important to evaluate the classifier based on how well the classifier predicts the "High Risk" patients. In many cases it has been found that the classification outcome is showing good accuracy as it can predict well the low risk patients (majority class) but failed to predict high risk patients (the minority class).

VI. Results

Different methods were tried in preparing a closely balanced datasets by under-sampling using different methods and over-sampling using SMOTE. The aim was to reduce the ratio gap between the majority classes with the minority class. The results are presented in Table I and II.

Four datasets were made using under-sampling by clustering and random under-sampling and named as USD1…USD4, as described in Table I. For exploring different alternatives the ratio gap of majority class samples to minority class samples were also reduced further. In order to understand the quality of the training set, the minority samples were separated into three clusters and then grouped in different combination with the clusters of majority class samples; an example of such a dataset is USD2.

One further dataset was created using the under-sampling by clustering method proposed by Yen and Lee (2009). The first dataset (K3M1Yen) was produced by separating the full data to 3 clusters and collected the majority class samples using equation (1) with the majority and minority ratio 1:1 (M=1).

SMOTE is used to create one more dataset, where the minority samples were oversampled by 485% to make the ratio 1:1. The detail descriptions of the datasets are given the following Table II.

TABLE II: The Descriptions of the Datasets

Data	Ratio	Description
USD1	2 : 1	Data consist of all the minority class samples ("dead") and one cluster of majority class records out of three clusters made by K-Mean.
USD2	2.4 : 1	Data consist of combination of two clusters of the minority class samples and one cluster of majority class samples. Clusters are made with simple k-mean for both of the classes (K=3).
USD3	3 : 1	Data consist of combination of all the minority class samples with randomly (random cut 1) selected samples from majority class sample.
USD4	3: 1	Data consist of combination of all the minority class samples with randomly (random cut2) selected samples from majority class sample.
OD	6 :1	Original data with full samples.
K3M1Yen	1: 1	Majority and minority ratio 1:1 (M=1) using Yen and Lee (2009)
SMOTE	1:1	Data were balanced using SMOTE Over-sampling

TABLE III: CLASSIFICATION OUTCOME OF FURIA

Data Sets	ACC%	SEN%	SPEC%	PPV%	NPV%
USD1	85.89	64.17	98.12	95.06	82.94
USD2	92.11	79.78	97.21	92.21	92.07
USD3	74.68	11.67	96.29	51.85	76.07
USD4	70.82	15.83	89.52	33.93	75.78
USD5	66.71	30.00	72.97	15.93	85.93
K3M1Yen	61.48	67.50	55.65	59.56	63.89
SMOTE	85.28	83.78	86.77	86.36	84.25

TABLE IV: CLASSIFICATION OUTCOME OF DECISION TREE

Data Sets	ACC%	SEN%	SPEC%	PPV%	NPV%
USD1	84.08	67.50	93.43	85.26	83.61
USD2	92.05	83.15	95.77	89.16	93.15
USD3	67.66	35.83	78.57	36.44	78.13
USD4	66.60	33.33	77.90	33.90	77.46
USD5	79.59	20.00	89.76	25.00	86.80
K3M1Yen	51.64	52.50	50.81	50.81	52.50
SMOTE	85.78	84.21	87.34	86.93	84.69

From the Table III and IV the original unbalance dataset USD5 has accuracy of 66.71% with FURIA classification and 79.59 % with decision tree classification. But for both of the classifiers the sensitivity value is very poor (30% and 20%). The accuracy is high because the classifier was able to classify the majority class (*Alive*) sample well (72.97% and 89.76%) but failed in classifying the minority set. Dataset USD1 where data are balanced by clustering the majority class samples and combining all the minority samples shows better classification outcome than the original unbalance data. With the FURIA and decision tree classification of the USD1 dataset, the sensitivity value is 0.642 with the decision tree and 0.675 with the FURIA. The classification outcome of the USD1 is 2 to 3 times higher than the original datasets. The dataset prepared by the method proposed by Yen and Lee (2009) could show some increase in the sensitivity value but the accuracy dropped and overall performance was not good. Under sampling by random cut USD3 and USD4 also disappointed in its accuracy and sensitivity values.

The SMOTE over-sampling technique shows a good classification outcome for both the classifiers decision tree and FURIA. But the performance of under-sampling using the proposed approach is better in terms of classification accuracy and training time of the classifier. Out of 10 runs the average tanning time of the dataset prepared by SMOTE was 3.84 second and with our approach 0.1 to 0.31.

It is observed from the experiments that the majority and minority ratio is not the only issue in building a good prediction model. There is also a need for good training samples that display data properties consistent with the class label assigned to them. Most of the time the records of clinical datasets do not truly reflect data properties consistent with the outcome label. The majority and minority ratio of USD1 and USD2 are very close but the classification outcomes are not similar. Although the majority minority ratio is almost same, there is a big difference of the classification accuracy, sensitivity and specificity of the datasets; as can be noticed in the Table I and II. The dataset "K3M1Yen" prepared by the method proposed by Yen and Lee (2009) has 1:1 ratio but still have less classification outcome than other datasets.

The ROC [16]space for all datasets classified with decision tree is plotted in Fig. 2 and for FURIA plotted in Fig. 3. These show that overall accuracy of all the datasets are above the random line, and the datasets USD1, USD2 and SMOTE with highest accuracy than all the other datasets.

Over-sampling using SMOTE shows good classification outcome and some cases it is very close to the performance of the proposed approach to under-sampling. However, the training time of the datasets prepared by SMOTE is much higher than datasets prepared by under sampling techniques.

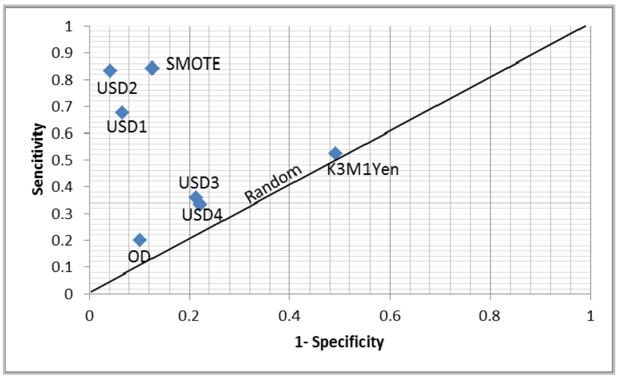

Fig. 2. ROC of Decision Tree Classification.

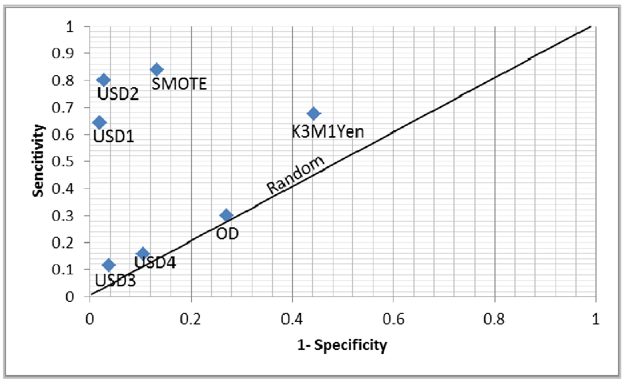

Fig. 3. ROC of FURIA Classification

VII. CONCLUSIONS

Class imbalance is a common problem with most medical datasets. Most existing classification methods tend not to perform well on minority class examples when the dataset is extremely imbalanced. Sampling strategies have been used to overcome the class imbalance problem by either over-sampling or under-sampling. Many researchers proposed different methods of over-sampling and under-sampling the majority class sample to balance the data. We examined the popular over-sampling technique SMOTE and some under-sampling techniques over cardiovascular data. This paper proposed a modified cluster based under-sampling method that not only can balance the data but also can generate good quality training sets for building classification models. The outcome labels of most of the clinical datasets are not consistent with the underlying data. If we consider the current data set, where cardiovascular risk is based on whether previous patients records display dead or alive of, it appears some of the patients may have died due to causes other than cardiovascular risk; conversely some high risk cardiovascular patients appear to be alive by chance. Both situations confound the class imbalance problem. The conventional over-sampling and under-sampling technique may not always be appropriate for such datasets. The proposed method is found to be useful for such datasets where the class labels are not certain and can also help to overcome the class imbalance problem of clinical datasets and also for other data domains.

REFERENCE

[1] Y. Liu *et al.*, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Information Processing & Management,* vol. 47, no. 4, pp. 617-631, Jul, 2011.

[2] M. S. Kim, "An Effective Under-Sampling Method for Class. Imbalance Data Problem," in *Proc. 8th International Symposium on Advance intelligent System (ISIS 2007)*, 2007.

[3] Y. P. Zhang , L. N. Zhang, and Y. C. Wang "Cluster-based majority under-sampling approaches for class imbalance learning," pp. 400-404, 2010.

[4] R. Laza *et al.*, "Evaluating the effect of unbalanced data in biomedical document classification," *Journal of integrative bioinformatics,* vol. 8, no. 3, pp. 177, 2011 Sep, 2011.

[5] N. V. Chawla *et al.*, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research,* vol. 16, pp. 321–357, 2002.

[6] Y. Zhai *et al.*, "An Effective Over-sampling Method for Imbalanced Data Sets Classification," *Chinese Journal of Electronics,* vol. 20, no. 3, pp. 489-494, Jul, 2011.

[7] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications,* vol. 36, pp. 5718–5727, 2009.

[8] C. Drummond and R. C. Holte, "C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling," in Workshop on Learning from Imbalanced Data Sets II, 2003.

[9] Y. M. Chyi, "Classification analysis techniques for skewed class distributionproblems," Department of Information Management, National Sun Yat-Sen University, 2003.

[10] J. Zhang and I. Mani, "kNN approach to unbalanced data distributions: A case study involving information extraction," presented at ICML'2003 workshop on learning from imbalanced datasets, 2003.

[11] R. C. Barros *et al.*, "A Survey of Evolutionary Algorithms for Decision-Tree Induction," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 42, no. 3, pp. 291-312, 2012.

[12] F. Lotte, A. Lecuyer, and B. Arnaldi, "FuRIA: A Novel Feature Extraction Algorithm for Brain-Computer Interfaces using Inverse Models and Fuzzy Regions of Interest," in *Proc. 3rd International IEEE/EMBS Conference on Neural Engineering, CNE '07*, pp. 175-178, 2007.

[13] J. Han and M. Kamber, Data mining: concepts and techniques, Amsterdam [u.a.]: Elsevier/Morgan Kaufmann, 2011.

[14] M. M. Rahman and D. N. Davis, "Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for K-Mean Clustering on Real Cardiovascular Data," presented at ICDMKE_5:The 2012 International Conference of Data Mining and Knowledge Engineering, World Congress on Engineering 2012 (WCE 2012), July 4-6, 2012, London.

[15] D. N. Davis and T. T. T. Nguyen, "Generating and Veriffying Risk Prediction Models Using Data Mining (A Case Study from Cardiovascular Medicine)," presented at European Society for Cardiovascular Surgery 57th Annual Congress of ESCVS, Barcelona Spain, 2008.

[16] T. C. W. Landgrebe and R. P. W. Duin, "Efficient Multiclass ROC Approximation by Decomposition via Confusion Matrix Perturbation Analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 30, no. 5, pp. 810-822, 2008.

M. Mostafizur Rahman received a BSc degree in Computer Science from International University of Business Agriculture and Technology and a Master's degree in Applied Artificial Intelligence from Dalarna University Sweden. After graduation he worked as Assistant Professor at the Leading University Bangladesh and later he joined Eastern University Bangladesh as an assistant professor. He is currently on study leave from Eastern University and doing his PhD in The University of Hull, UK. He has over 10 years of teaching and research experience. He is a (co-)author of published papers in journals and conference proceedings. His research interests include medical systems, data mining, machine learning, artificial intelligence and applications of notions of fuzzy logic.

Darryl Davis is director of research in the Department of Computer Science at the University of Hull, UK. He has a B.Sc. in Psychology, M.Sc. in Knowledge Base Systems and Ph.D. in Diagnostic and Investigative Medicine (The Application of AI in Medicine). He has over 20 years experience in artificial intelligence systems. Successful applications include classification and computer vision problems in business, medicine and geology. Application research includes the use of agents as a vehicle for domain applications, and the application of computational intelligence to medical domains and machine vision. He has over 20 years experience in data mining, first working on Decision Trees at Edinburgh and the Turing Institute (Glasgow) in 1987. Medical applications of data mining and decision support include Urology, Orthodontics, Stomach cancer, Cervical cancer, Bowel cancer, Neuroanatomy and Cardiovascular medicine. He has published widely with over 100 publications.