Journal of Network and Innovative Computing
ISSN 2160-2174, Volume 1 (2013) pp. 332-340

A Review of Class Imbalance Problem

Shaza M. Abd Elrahman[1] and Ajith Abraham[2]

[1]Faculty of Computer Science & Information Technology, Sudan University of Science and Technology, Khartoum, Sudan
shaza.merghani@sustech.edu

[2]Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, WA, USA
ajith.abraham@ieee.org

Abstract- Class imbalance is one of the challenges of machine learning and data mining fields. Imbalance data sets degrades the performance of data mining and machine learning techniques as the overall accuracy and decision making be biased to the majority class, which lead to misclassifying the minority class samples or furthermore treated them as noise. This paper proposes a general survey for class imbalance problem solutions and the most significant investigations recently introduced by researchers.

I. Introduction

Class imbalance problem is a hot topic being investigated recently by machine learning and data mining researchers. It can occur when the instances of one class outnumber the instances of other classes. The class have overwhelmed called the majority class while the other called minority class. However, in many applications the class has lower instances are the more interesting and important one. The imbalance problem heightens whenever the class of interest is relatively rare and has small number of instances compared to the majority class. Moreover, the cost of misclassifying the minority class is very high in comparison with the cost of misclassifying the majority class for example; consider cancer versus non-cancer or fraud versus un-fraud [1].

The class imbalance can be intrinsic property or due to limitations to obtain data such as cost, privacy and large effort [2]. Many real world applications such as medical diagnosis, fraud detection (credit card, phone calls, insurance), network intrusion detection, pollution detection, fault monitoring, biomedical, bioinformatics and remote sensing (land mine, under water mine) suffer from these phenomena.

There are different difficulties caused by imbalance classes and they also hinder the performance of machine learning and data mining techniques:

The class distribution, in standard classifiers such as decision trees and neural networks assume that the training samples are equally distributed among classes. However, in many real applications the ratio of the minority class is very low (1:100, 1:1000 or may exceeded to 1:10000). Due to lack of data, few samples of minority class in training set tends the classifiers to falsely detect them and the decision boundary be far from the true one. There are also issues due to concept complexity or overlapping, which refers to level of separability between data classes. High overlapped classes and high noise level produced higher complexity. Moreover, the discriminating rules can be difficult to induce if the examples of each class are overlapping at different levels in some feature space [3]. Finally the existence of small disjuncts in a data set adds more complexity to the problem. Furthermore, in most imbalance problems the cost of errors for different classes is uneven and usually it is unknown.

The rest of the paper is structured as follows. In Section II, we demonstrate the feature selection methods used in imbalanced classes. Section III, explain various evaluation metrics used in imbalanced classes. In Section IV, we explain various solutions introduced for dealing with imbalance class's problem

II. Feature Selection in Imbalance Problems

Feature selection is another critical issue in machine learning and data mining. It aims to select important features that improve the accuracy and performance of the classifier. High dimensional data and irrelevant features may reduce the performance of the classifier and increase the misclassification rate especially in imbalance data sets [4], [5]. Feature selection metrics can be categorized as one-sided or two-sided based on whether they select only positive features (most indicative of membership on the target class) or combine both positive and negative features [6]-[7]. Also, feature selection metrics can be categorized as binary or continuous feature selection metrics depend on the data type. For example; Chi square, Information Gain (IG) and odds ratio (OR) can handle both binary and nominal data. But Pearson correlation coefficient, feature assessment by sliding threshold (FAST) and signal to noise ratio (S2N) can handle continuous data [5]-[9].

Nguwi and Cho [10] presented a weight vector sensitivity feature selection method derived from SVM (Support Vector Machine). They used ranking criteria and eliminated those features less contributed on enhancing the generalization capability of classifier. The emergent Self-Organizing Map (ESOM) was used to cluster the ranker features so as to provide clusters for unsupervised classification.

Alibeigi et al. [6] presented a feature ranking approach based on the probability density estimation of features for small sample size and high dimensional imbalanced data sets. Density Based Feature Selection (DBFS) taking the advantage of features' distributions over classes with their correlations.

III. Evaluation Metrics Used in Imbalanced Classes

Evaluation metrics is a critical issue in machine learning, which used as indicator for the performance of machine learning algorithms. The standard evaluation metrics used are accuracy and error rate however, these metrics are not proper to handle imbalance classes as the overall accuracy be biased to the majority class regardless of the minority class with lower samples which leads to poor performance on it. For the two class problem, common metrics are derived from a confusion matrix as shown in Table 1.

Table 1: 2x2 Confusion Matrix

		Predicted Class	
		+ ve	**- ve**
Actual Class	**+ ve**	True Positive (TP)	False Negative (FN)
	- ve	False Positive (FP)	True Negative (TN)

The most evaluation metrics related to imbalance classes are recall (sensitivity) (1), specifity (2), precision (3), F-measure(4),(5) , geometric mean (g-mean) (6) [11]. Sensitivity and specificity are used to monitor the classification performance on each individual class. While precision is used in problems interested on highly performance on only one class, F-measure and G-mean are used when the performance on both classes –majority and minority classes- needed to be high [12].

$$\text{Sensitivity (true positive rate)} = \frac{TP}{TP + FN} \quad (1)$$

$$\text{Specifity (true negative rate)} = \frac{TN}{TN + FP} \quad (2)$$

$$\text{precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{F} = \frac{(1 + \beta^2).(precision.TPrate)}{\beta^2.precision + TPrate} \quad (4)$$

where as β is non negative constant and generally is set to 1 so:

$$\text{F} = \frac{2.precision.\text{Sensitivity}}{precision + \text{Sensitivity}} \quad (5)$$

$$G - mean = \sqrt{TPrate.TNrate} \quad (6)$$

Another popular metrics used in imbalance classes are ROC (Receiver Operating Characteristic) curve and AUC. ROC curve is a tool for visualizing and summarizing the performance of classifier based on trade off between true positive rate (Y-axis) and false positive rate (X-axis). AUC (7) is the area under the ROC curve, which is computed by:

$$\text{AUC} = \frac{TPrate + TNrate}{2} \quad (7)$$

There are another cost sensitive metrics are used the cost misclassifying errors when they are known such as cost matrix and cost curve. A cost matrix is a matrix that identify the cost of classifying samples where C(i,j) define the cost of classifying an instance from class ***i*** as class ***j.***

To evaluate both sampling and cost sensitive classifiers, the total cost computed by (8):

$$Total\ Cost = (FN.C_{FN}) + (FP.C_{FP}) \quad (8)$$

IV. Solutions Proposed for Dealing with Class Imbalance

Several methods proposed for solution of imbalance class problems include re- sampling and feature selection at the data level and other ones at the algorithm level such as cost sensitive and single class learning.

A. Sampling methods

Sampling methods is a preprocessing of data which handle the imbalance problem by constructing balanced training data set and adjusting the prior distribution for minority and majority class [2], [13]. Sampling methods include under sampling and over sampling methods. Under sampling balance the data by removing samples from majority class and over sampling balance the data by create copies of the existing samples or adding more samples to the minority class. Significant differences between majority and minority class can be handled by oversampling when data is highly imbalanced [14].

However, under sampling may cause loss of useful information by removing significant patterns and over sampling may cause over fitting and may introduce additional computational task. To tackle this problem Chawla et al. [15] proposed a synthetic minority over sampling technique (SMOTE) by generating a synthetic examples rather than replacement with replication. This technique identifies more specific regions in the feature space for the minority class. The

proposed technique maximized the performance of the classifier and the learning biased towards the minority class. However, SMOTE is applicable only for binary class problems with a continuous feature space. Another drawback of SMOTE is appeared when the number of examples of minority class is not adequate for estimating the accurate probability distribution for the actual data [2].

Chawla et al. [16] applied SMOTE in imbalanced data that severe from high sparsity in addition to high class skew. They used the Naïve Bayes and the decision tree algorithms. Their results explained the effectiveness of SMOTE in sparse dataset.

Tafta et al. [17] proposed a technique that combined (SMOTE) and the particle swarm optimization (PSO) and radial basis function (RBF) classifier. They applied PSO algorithm to determine the structure and the parameters of RBF kernels. The results explained the competitive performance of SMOTE PSO.

Fernandez-Navarro et al. [18] proposed two oversampling methods; static SMOTE radial basis function method and a dynamic SMOTE radial basis function procedure incorporated into a memetic algorithm that optimizes radial basis functions neural networks. The experiments showed the highest accuracy and sensitivity level of the dynamic over-sampling method comparing to other neural networks methods.

Mazurowskia et al. [19] investigated the use of under sampling and oversampling with two different neural network learning methods backpropagation (BP) and particle swarm optimization (PSO). They concluded that PSO was more sensitive to class imbalance, small training sample size and large number of features.

Cohena et al. [20] proposed a resampling approach using both oversampling and under-sampling with synthetic instances. Also, they introduced class-dependent regularization parameters for tuning SVM and obtained asymmetrical soft margin (larger margin on the side of the smaller class).

An under sampling approach based on ant colony optimization (ACO) was proposed by Yua et al. [21]. The more significant and informative majority samples estimated according to their selection frequency. This method produced the optimal majority balance set however; it is time consuming than the simple sampling approach.

Yong [22] proposed an oversampling method based on the K-means cluster and the genetic algorithm. K-means algorithm was used to cluster the minority class samples and the genetic algorithm was used to gain the new sample from each cluster. The results showed the effectiveness of the proposed method using nearest neighbor (KNN) and support vector machine (SVM) as classifiers.

An investigation of the effects of imbalance ratio and the classifier was presented by Garcia et al. [14]. They evaluated several sampling methods RUS (random under sampling) and WE+ MSS (Wilson's editing with MSS condensing over the negative instances) as under sampling methods and SMOTE and gg-SMOTE (Gabriel- graph-based SMOTE). Their results showed that over sampling was outperformed under sampling in highly imbalance datasets as under sampling causes loss of significant patterns. And the performance of evaluated methods (under sampling plus over sampling) was alike when the imbalance ratio was low.

Li et al. [23] used granular support vector machines repetitive under sampling method (GSVM-RU). This method balances the majority class by extract important samples and removes those unimportant ones. It significantly improved the efficiency of SVM model and reduced the computational cost.

Kamei et al. [24] evaluated the effects of four sampling methods (random over sampling, SMOTE, random oversampling and one sided selection) using four models (linear discrimination analysis, logistic regression analysis, neural network and classification tree). However, the sampling methods improved the prediction performance of linear and logistic models but there was no effect on neural network or classification tree performance.

Yen and Lee [25] introduced a cluster based under sampling approach. The training data was divided into clusters and then the representative data for majority class samples were selected from each cluster regarding the ratio of majority class samples to minority class samples. Their results showed that the cluster based under sampling improved prediction accuracy and it was more stable than other under sampling approach.

Kerdprasop and Kerdprasop [26] used Random over sampling and SMOTE to improve the performance of the learned model using decision tree induction, regression analysis, neural network and SVM. The highest sensitivity model given by random over sampling while SMOTE gives the highest specificity model. Moreover, they applied a cluster based feature selection which added a significant improvement to the predicting accuracy for the learned models.

Ramentol et al. [27] introduced a new hybrid sampling method SMOTE with fuzzy rough set theory (SMOTE-FRST). They improved the performance of SMOTE by eliminating the synthetic minority class samples which they had lower degree to the fuzzy region. To evaluate the proposed method, C4.5 was used as a classifier. SMOTE-FRST performance surpassed other SMOTE approaches.

Fernándeza et al. [28] used fuzzy rule classification systems. They extracted hierarchical rule base (HRB) from the initial rule base and studied the effects of using SMOTE on the performance of the hierarchical fuzzy rule base classification system (HFRBCS). The best cooperative rules from the HRB were selected using genetic algorithm.

Pérez-Godoy et al. [29] studied the effect of SMOTE on performance of CO2 RBFN (evolutionary cooperative–competitive model for the design of radial-basis function networks) and extended their experiments to ANN, C4.5 decision tree and fuzzy rule-based classification system

(HFRBCS). The best results obtained when combined CO2RBFN with SMOTE.

A hybrid under sampling technique for mining unbalanced datasets was proposed by Ravi and Vasu [30]. (KRNN) was applied to detect the outliers and K-means clustering on the majority class. The proposed method was tested using several classifiers such as SVM, logistic regression (LR), radial basis function network (RBF), genetic programming and decision tree (J48). Their results showed that the proposed under sampling technique increased the classifier's performance.

Although of SMOTE's advantages for balancing the data effectively however, it may bring noise and other problems. Recently, Mi [31] proposed an active learning SMOTE. He introduced SVM into a SMOTE learning frame. Their results showed that the proposed method outperformed other learning models.

Also, a hybrid feature selection method was proposed by Biodgloi and Parsa [32] by combining re-sampling and feature subset approaches. They used SMOTE for re-sampling and consistency subset evaluation method and genetic search for finding the optimal feature space and removing irrelevant features. The proposed method improved the classifier performance and outperformed the other feature selection method.

Thammasiria et al. [33], analyzed the interrelationships of the performance among sampling methods and classifiers. They tested three sampling techniques over sampling, under-sampling and SMOTE with four classifications methods logistic regression, decision trees, neuron networks and support vector machines. Their results showed the better combination was SVM with SMOTE.

Zhou [34] investigated the effect of six different sampling methods with number of samples in training set on highly imbalance data. Their experiments showed that, there is no difference on performance and the proper sampling methods depend on the number of training set samples. Although oversampling is time consuming however, SMOTE is a better choice if the training sample size is too large.

For enhancing C4.5 and PART rule induction algorithms, Garcia et al. [33] proposed an under sampling method guided by evolutionary algorithms to perform the training set selection. The proposed method outperformed standard under-sampling methods and the prediction model became smaller in number of leaves or rules and more interpretable.

A decision tree method based on Kolmogorov–Smirnov statistic (K–S tree) was proposed by[36]. This K–S tree have two benefits: first, it selects relevant variables and remove redundant variables. Then it reduced the effects of imbalance in training data by segmenting the complex problems into sub problems, which is less severe from class imbalance. They rebalanced the data at each segment using under-sampling and oversampling methods.

Although most of learners benefits from sampling techniques, the performance of sampling techniques depend on the dataset size imbalance ratio [37].

Some works [38-39] investigated the impacts of noise with class imbalance. Hulse and Khoshgoftaar [38], found that the impact of noise depend on the complexity of learning algorithm whereas the simple learners such as naïve Bayes and nearest neighbor learners are often more robust than more complex learners such as support vector machines or random forests. Also, they concluded that sampling improves the performance of class imbalance and noise classifiers. Moreover, they found that simple sampling methods are the most effective, WE and RUS are generally the two best techniques and RUS performed very well with higher levels of minority class noise.

Seiffert et al. [39] used different classification algorithms including decision trees, nearest neighbors, neural networks and Bayesian learners. Also they analyzed the relationship between classification performance, data sampling, learner selection, class imbalance and class noise. They concluded the following results:

- RBF proved to be most sensitive to imbalance, most learners and sampling techniques actually improved in performance as imbalance was increased.
- The reduction of noise had a more significant impact on sampling technique performance than the increase in imbalance.
- WE proved to be the best sampling technique achieving the highest AUC in most cases.
- RUS performed best when combined with four classification algorithms (C4.5, RBF, RIPPER and SVM) at all levels of noise and imbalance.
- NB and SVM consistently perform best on all datasets for all levels of imbalance and noise.

Recently, Thanathamathee and Lursinsap [40] proposed a a method that combined boundary data generation and boosting procedures. Firstly, they eliminated the imbalanced error by identifying all relevant class boundary data using Hausdorff distance. Secondly, they expanded the distribution of training data space using the concept of bootstrapping to estimate new region of each sub-cluster and synthesize the new boundary data. AdaBoost algorithm was used for classifying all new synthesized data

B. Cost sensitive learning

In many imbalance problems, not only the data distribution is skewed but also the misclassification error cost is uneven. The cost learning techniques take the misclassification cost in its account by assigning higher cost of misclassification to the positive class (minority class) i.e. $C(+,-) > C(-,+)$ and generate the model with lowest cost [3]. However, the misclassification

errors costs are often unknown and furthermore, cost sensitive learning may lead to over fitting [32]. Another cost sensitive learning approach used in unbalance dataset is adjusting the decision threshold of the standard machine learning algorithms, wherever the selection of threshold is an effective factor on the performance of learning algorithms [41].

Thach et al. [42] proposed accuracy- based learning (XCS) with cost sensitive. They identified a constraint reward function which maximizing the total reward of the positive class samples and improve the performance of XCS in unbalance data. Alejo et al. [43] proposed a hybrid method based on Gabriel graphs technique and modified back propagation algorithm. They proposed new cost function based on minimum square error (MSE).

Yang et al. [44] proposed cost sensitive SVM that modified margins and lopsided them to achieve a highly unbiased decision boundary. The modification employed a penalty regularization by adopting an inversed proportional regularized penalty to re-weight the imbalanced classes. Then margin compensation was applied to lead the margin to be lopsided, which enables the decision boundary drift. The proposed method achieved highly unbiased accuracy in a complex imbalanced dataset.

Uyar et al. [41] examined the classification performance when using oversampling, under sampling and adjusting the decision threshold. Their results showed that the optimum true positive rate and false positive rate could be obtained easily by adjusting the decision threshold. Also, Yan et al. [45] proposed an adjustment method for threshold based on Fisher discrimination. The proposed method improves the accuracy.

Maalouf and Trafalis [46] proposed a weighted Kernel Logistic Regression by combining rare events corrections to Logistic Regression (LR) with truncated Newton methods. They explained the strength and the accuracy of Weighted Kernel Logistic Regression (RE-WKLR) even if the datasets is imbalanced or not linearly separable. Moreover, they explained the benefit of less complex of unconstrained optimization of RE-WKLR compared with constrained optimization methods, such as SVM.

Maratea et al. [47], proposed a modification for Support Vector Machine algorithm to be effectively cope with data imbalance using approximate solution and kernel transformation, they compensated data skewness by enlarge asymmetrically space around the class boundary. Also, they proposed an accuracy measure, named AGF, which is a generalization of the F-measure.

A weighted maximum margin criterion to optimize data-dependent kernel was proposed in [48]. The optimization based on the maximization of the weighted average margin between the majority and minority classes, which made the minority class more clustered in the induced feature. Hwang et al. [49], proposed a weighted Lagrangian support vector machine (WLSVM). They embedded weight parameters in the Lagrangian SVM formulation. This method speed up the training and improve the performance of LSVM on imbalance problem

Oh [50] introduced a new error back-propagation function, which intensified weight updating for minority class. Al-Haddad et al. [51] compensated for the imbalance data by using a posteriori probabilities to adjust the neural networks. Another probability based weighting approach proposed in [52]. SVMs and Naive Bayes classifiers were used to test the proposed method. Their approach boosts the performance over skewed data.

A modular neural network (MNN) based on divide-and-conquer technique was proposed in [53]. MNN was used to solve complex imbalance multi-class problem by transforming an imbalanced multi classification problem into symmetrical sets two-class problems. Their results showed that the proposed method attained the better performance and it was less complexity and time consumptions compared with other neural networks so as the modified back-propagation technique.

C. Recognition based methods

In recognition-based method (one-class learning) the classifier learn on the just minority class samples (target class). This approach improves the performance of the classifier on unseen data and recognize only those belong to that class. Raskutti and Kowalczyk [54] investigated the effect of sampling, and weighted learning of a single class. They concluded that one-class learning can be a robust technique when dealing with unbalanced data and highly dimensional noisy feature space. One-class learning can perform better under certain conditions such as high dimensional data, however, many classifiers such as decision trees and Naive Bayes cannot be built by one class learning.

D. Ensemble- based Methods

Ensemble is a combination of multiple classifiers so as to improve the generalization ability and increase the prediction accuracy. The most popular combining techniques are boosting and bagging. In boosting, each classifier is dependent on the previous one, and focuses on the previous one's errors. Examples that are misclassified in previous classifiers are chosen more often or weighted more heavily. Whereas, in bagging, each model in the ensemble votes with equal weight. In order to promote model variance, bagging trains each model in the ensemble using a randomly drawn subset of the training set [55]. Kang and Cho [56] proposed an ensemble of under sampled SVM (EUS SVMs). They integrated the good generalization ability of SVM by boosting ensemble scheme. Their proposed method overcame the drawback of under sampling method and reduced the time complexity of oversampling method.

An investigation on the performance of random sampling and advanced under sampling (CUBE) and two modeling techniques (gradient boosting and weighted random forests) was introduced by Burez and Poel [57]. They concluded that under sampling improved the prediction accuracy comparably with sophisticated under sampling which had no any effect on the performance. Also, they found that Boosting is a robust classifier but not surpassed the other techniques and weighted random forest performed better than random forest.

Gue and Viktor [58] proposed an ensemble based learning approach (DataBoost-IM) that combined boosting with data generation. The hard examples were identified then they were used to generate synthetic examples for both classes to be focus by the next classifier component in the boosting procedure. The synthetic examples prevented boosting from over fitting on hard examples. Another ensemble in a hierarchical frame was proposed by Zhang and Luo [59]. They proposed a parallel classification method to improve classifying speed; two classifiers (simple one and complicated one) were trained serially but worked in parallel. The results showed that their proposed approach effectively improved performance and speed. An approach based on repeated sub-sampling proposed by Khalilia et al. [60]. They compared the performance of SVM, bagging, boosting and Random Forest (RF). They emphasized the effectiveness of repeated sub-sampling in dealing with highly imbalance data sets. However, RF outperformed other methods plus its ability to estimate the importance of each variable in classification process.

Chena et al. [61] employed cost-sensitive Random Forest by both sampling and thresholding. First search, which is Correlation based feature selection algorithm, was applied to select the best subset of features. Their experiments showed the enhancement of performance on rare classes but a little degradation on the majority class however which can be adjusted by using a cost matrices.

Lia and Suna [62] incorporated nearest neighbour (NN) in a new oversampling approach and a bagging ensemble. The new samples generated using random distance to nearest neighbour (NN). Then the nearest-neighbour support vector machines (NNsSVM) were generated and assembled using bagging algorithm.

To improve the performance, Liu et al. [63] integrated both sampling methods; oversampling and under sampling with ensemble of SVM (EnSVM) model. Moreover, to boost the performance they developed a genetic algorithm-based model for classifier selection. Their results showed the effectiveness of the proposed model.

Zhang et al. [64] proposed ensemble methods combined with cluster based under-sampling, which remove the nearest clusters by calculating distance from each cluster to the minority class. To improve imbalanced classification, bagging and Adaboost were used to train two ensembles of SVMs and ANNs (Artificial Neural Networks). The results explained the effective performance of ensemble.

Zhang et al. [65] introduced a dynamic classifier ensemble method for imbalanced data (DCEID). The proposed ensemble adaptively selected the proper dynamic ensemble between dynamic classifier selection (DCS) and dynamic ensemble selection (DES). Also, they proposed a new cost-sensitive selection criteria constructed for DCS and DES. Their results showed that DCEID outperformed some static ensemble strategies such as weighted random forests and improved balanced random forests.

Recently [66] introduced an ensemble an online ensemble of NN for classifying non -stationary and imbalanced data streams. The proposed ensemble consisted of two layers, the first layer handled class imbalance using cost sensitive nn. In the second layer new weighting method proposed for handling both class imbalance and non-stationary data streams

Xiao et al. [66] introduced a new algorithm for classifying imbalanced data called LFC (linear F-measure classification). Their results showed the effectiveness of LFC algorithm especially with the overlapped distributions comparing with one class SVM and cost sensitive SVM.

Table 2 summarizes the advantages and drawbacks of the proposed methods for dealing with imbalance problem: sampling, cost sensitive learning, one class learning and ensemble approaches.

V. Conclusions

Class imbalance is a hot topic being investigated recently by machine learning and data mining researchers. The researchers for solving the imbalance problem have proposed various approaches. However, there is no general approach proper for all imbalance data sets and there is no unification framework. This paper summarizes various solutions for dealing with class imbalance problems.

Table 2: The advantages and drawbacks of the proposed methods for dealing with imbalance problem

Method	Advantages	Limitations
Under-sampling	• Independent on underlying classifier. • Can be easily implemented	• May remove significant patterns and cause loss of useful information
Over-sampling		• Time consuming: Introduce additional computational cost • May lead to over-fitting
Cost sensitive	• Minimize the cost of misclassification (by biasing the classifier toward the minority class)	• The misclassification costs (the actual cost of errors) often are unknown
Recognition based	• Have better performance especially on high dimensional data	• Many classifiers such as decision trees and Naive Bayes cannot be built by one class learning.
Ensemble	• Better classification performance than individual classifiers • More resilience to noise	• Time consuming • Over fitting

References

[1] N. Satuluri and M. R. Kuppa, " A Novel Class Imbalance Learning Using Intelligent Under-Sampling," *International Journal of Database Theory and Application*, vol. 5, pp. 25-35, 2012.

[2] S. L. Phung, A. Bouzerdoum and G. H. Nguyen (2009) Learning Pattern classification tasks with imbalanced data sets [Online].Available: ro.uow.au/infopapers/792

[3] Y. Sun, M. S. Kamel, A. K.C. Wong and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," *PATTERN RECOGNITION*, vol. 40,pp. 3358-3378, 2007.

[4] K. Chomboon , K. Kerdprasop and N. Kerdprasop, Rare Class Discovery Techniques for Highly Imbalance Data. Proc. International multi conference of engineers and computer scientists, vol. 1, 2013.

[5] L. Lusa and R. Blagues, " The Class-imbalance for high-dimensional class prediction," in 11th International Conference on Machine Learning and Application, *IEEE*, 2012.

[6] M. Alibeigi, S. Hashemi and A. Hamzeh," DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets," *Data & Knowledge Engineering, vol.* 81–82, pp. 67–103, 2012

[7] H. Ogura, H. Amano and M. Kondo," Comparison of metrics for feature selection in imbalanced text classification," *Expert Systems with Applications*, vol. 38, pp. 4978–4989, 2011.

[8] M. Wasikowski and X. Chen, "Combating the Small Sample Class Imbalance Problem Using Feature Selection," *IEEE Transactions on Data Engineering*, vol. 22, 2010.

[9] I. Jamali, M. Bazmara and S. Jafari, "Feature Selection in Imbalance data sets," *International Journal of Computer Science Issues*, vol. 9, 2012.

[10] Y. Nguwi and S. Cho," An unsupervised self-organizing learning with support vector ranking for imbalanced datasets," *Expert Systems with Applications*, vol. 37,pp. 8303–8312, 2010

[11] X. Gue, Y. Yin, C. Dong, G. Yang and G. Zhou, "On the Class Imbalance Problem," in Fourth International Conference on Natural Computation, *IEEE*,2008.

[12] V. Garcia, J. S. Sanchez and R. A. Mollineda, "On the effective of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, vol. 25, pp. 13-21, 2012.

[13] B. Lu, X. Wang, Y. Yang and Zhao, " Learning from imbalanced data sets with a Min-Max modular support vector machine," *Front. Electr. Electron. Eng.*, vol. 6, pp.56-71, 2011.

[14] V. Garcia, J.S. Sanchez and R.A. Mollineda, " On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Systems*, vol. 25, pp. 13-21, 2012.

[15] N. V. Chawla, K. W. Bowyer, L. O. Hall and W. P. Kegelemer, "SMOTE: Synthetic Minority Over-Sampling Technique," *Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[16] L.M. Tafta, et al., "Countering imbalanced datasets to improve adverse drug event predictive models in labor and delivery," *Journal of Biomedical Informatics*, vol. 42, pp. 356–364, 2009

[17] M. Gaoa , X. Honga, S. Chenb and C. J. Harrisb," A combined SMOTE and PSO based RBF classifier for two-class imbalanced problems," *Neurocomputing*, vol. 74, pp. 3456–3466, 2011

[18] F. Fernandez-Navarro, C. Hervas-Martinez and P. A. Gutierrez," A dynamic over-sampling procedure based on sensitivity for multi-class problems," *Pattern Recognition*, vol. 44, pp. 1821–1833, 2011

[19] M. A. Mazurowski, P. A. Habas , J. M. Zurada , J. Y. Lob , J. A. Baker and G. D. Tourassi, Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance, *Neural Networks*, (21) pp.427-436, 2008.

[20] G. Cohena, M. Hilariob , H. Saxc , S. Hugonnetc and A. Geissbuhlera, "Learning from imbalanced data in surveillance of nosocomial infection," *Artificial Intelligence in Medicine*, vol. 37, pp. 7—18, 2006

[21] H. Yua, J. Nib and J. Zhaoc, "ACOSampling: An ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data," *Neurocomputing*, vol. 101, pp. 309–318, 2013

[22] Y. Yong, "The Research of Imbalanced Data Set of Sample Sampling Method Based on K-Means Cluster and Genetic Algorithm," in 2012 International Conference on Future Electrical Power and Energy Systems, Energy Procedia, vol. 17, pp. 164 – 170, 2012

[23] Q. Li, Y. Wang and S. H. Bryant, " A novel method for mining highly imbalanced high-throughput screening data in PubChem,"*Bioinformatics*, vol. 25, pp. 3310-3316, 2009.

[24] Y. Kamei, A. Monden, S. Matsumoto, T. Kakimoto and K. Matsumoto," The Effects of Over and Under Sampling on Fault-Prone Module Detection," *IEEE*, pp. 196-204, 2007.

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

[26] N. Kerdprasop and K. Kerdprasop, "On the Generation of Accurate Predictive Model from Highly Imbalanced Data with Heuristics and replication Technologies," *International Journal of Bio-Science and Bio-Technology*, vol. 4,pp. 49-64,2012.

[27] E. Ramentol, N. Verbiest, Y. Caballero and C. Cornelis, SMOTE-FRST: A New Resampling Method Using Fuzzy Rough Set Theory, *proc: WSPC*, 2012.

[28] A. Fernándeza, M. J. Jesusb and F. Herrera," Hierarchical fuzzy rule based classification systems with genetic rule selection for imbalanced data-sets," *International Journal of Approximate Reasoning*, vol 50,pp. 561–577, 2009

[29] M. D. Pérez-Godoy, A. Fernández and A. J. Rivera, M. J. Jesus," Analysis of an evolutionary RBFN design algorithm, CO2 RBFN, for imbalanced data sets," *Pattern Recognition Letters*, vol. 31,pp. 2375–2388, 2010

[30] M. Vasu and V. Ravi, "A hybrid under-sampling approach for mining unbalanced datasets: applications to banking and insurance," *Int J. Data Mining Modelling and Management*, vol. 3,pp. 75-105,2011.

[31] Y. Mi, "Imbalanced Classification Based on Active Learning SMOTE," *Research Journal of Applied Science Engineering and Technology*, vol. 5, pp. 944-949,2013.

[32] A. Biodgloi and M.N. Parsa, " AHyprid Feature Selection by Resampling, Chi squared and Consistency Evaluation Techniques", *World Academy of Science, Engineering and Technology*, vol. 68, 2012.

[33] D. Thammasiria , D. Delenb,, P. Meesadc and N. Kasapd, "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition," *Expert Systems with Applications*, 2013

[34] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," *Knowledge-Based Systems*, vol. 41, pp. 16–25, 2013

[35] S. Garcia, A Fernandez and F. Herrera, "Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems," *Applied Soft Computing*, vol. 9, pp. 1304–1314, 2009

[36] R. Gong and S. H. Huang, "A Kolmogorov–Smirnov statistic based segmentation approach to learning from imbalanced datasets: With application in property refinance prediction," *Expert Systems with Applications*, vol. 39, pp. 6192–6200, 2012

[37] A.O. Puig and E.B. Mansilla, "Evolutionary rule-based systems for imbalanced data sets," *Soft Computing*, vol. 13,pp. 213-225, 2009.

[38] J. V. Hulse, T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data & Knowledge Engineering*, vol. 68, pp. 1513–1542, 2009

[39] C. Seiffert, T. M. Khoshgoftaar, J. V. Hulse and A. Folleco, "An empirical study of the classification performance of learners on imbalanced and noisy software quality data," *Information Sciences*, 2011

[40] P. Thanathamathee, C. Lursinsap, "Handling imbalanced data sets with synthetic boundary data generation using bootstrap re-sampling and AdaBoost techniques," *Pattern Recognition Letters*, vol. 34, pp. 1339–1347, 2013

[41] A. Uyar, A. Bener, H. N. Ciracy and M. Bahceci, "Handling the Imbalance Problem of IVF Implantation Prediction," *IAENG International Journal of Computer Science*, 2006.

[42] N. H. Thach, P. Rojanavasu and O. Pinngern, " Cost-sensitive XCS Classifier System Addressing Imbalance Problems," in Fifth International Conference on Fuzzy Systems and Knowledge Discovery, *IEEE*, pp. 132-136, 2008.

[43] R. Alejo, R.M. Valdovinos, V. Garcia and J.H. Pacheco-Sanchez, " A hyprid method to face class overlap and class imbalance on neural networks and multi-class scenario," *Pattern Recognition Letters*, vol. 34, pp. 3880-3888,2013.

[44] C. Yang, J. Yang and J. Wang, "Margin calibration in SVM class-imbalanced learning," *Neurocomputing*, vol. 73, pp. 397–411, 2009

[45] L. Yan, D. Xie and Z. Du, " A new Method of Support vector Machine for Class Imbalance Problem," in *International Joint Conference on Computational Science and Optimization, IEEE*, pp. 904- 907, 2009.

[46] M. Maalouf and T. B. Trafalis, "Robust weighted kernel logistic regression in imbalanced and rare events data," *Computational Statistics and Data Analysis*, vol. 55, pp. 168 183, 2011

[47] A. Maratea, A. Petrosino and Mario Manzo, "Adjusted F-measure and kernel scaling for imbalanced data learning", *Information Sciences*, 2013

[48] Z. Zhao, P. Zhong and Y. Zhao, "Learning SVM with weighted maximum margin criterion for classification of imbalanced data," *Mathematical and Computer Modelling*, vol. 54 pp.1093–1099, 2011

[49] J. P. Hwang, S. Park and E. Kim," A new weighted approach to imbalanced data classification problem via support vector machine with quadratic cost function," *Expert Systems with Applications*, vol. 38, pp. 8580–8585, 2011

[50] S.H. Oh, "Error back-propagation algorithm for classification of imbalanced data," *Neurocomputing,* vol. 74, pp. 1058–1061, 2011

[51] L. Al-Haddad , C. W. Morris and L. Boddy, "Training radial basis function neural networks: effects of training set size and imbalanced training sets," *Journal of Microbiological Methods*, vol. 43, pp. 33–44, 2000

[52] Y. Liua, H. T. Lohb and A. Sunc, "Imbalanced text classification: A term weighting approach," *Expert Systems with Applications*, vol. 36, pp. 690–701, 2009

[53] Z.Q. Zhao, "A novel modular neural network for imbalanced classification problems," *Pattern Recognition Letters*, vol. 30, pp. 783–788,2009

[54] B. Raskutti and A. Kowalczyk, "Extreme Re-balancing for SVMs: a case study," *sigkdd Explorations,* vol. 6, pp. 60-69,2004.

[55] M. Galar, A. Fernando, E. Barrenechea, H. Business and F. Herrera, " A Review on ensembles for the class Imbalance Problem," IEEE Transactions on Systems, Man, and Cybernetics- Part C: Applications And Review, vol. 42, 2012.

[56] P. Kang and S. Cho, EUS SVMs: *Ensemble of Under-Sampled SVMs for Data Imbalance Problems. Proc. International Conference on Neural Information Processing*, ser. Lecture Notes in Computer Science, 2006.

[57] J. Burez and D. V. Poel, "Handling class imbalance in customer churn prediction," *Experts System with Applications*, vol. 36, pp. 4626-4636, 2009.

[58] H. Gue and H. Viktor, "Learning from Imbalanced Data Sets with Boosting and Generation: The DataBoost-IM Approach," *Sigkdd Explorations*, vol. 6,pp. 30-39.

[59] Y. Zhang and B. Luo, *Parallel Classification Ensemble with Hierarchical Machine Learning for Imbalanced Classes. Proc: the seventh International conference on Machine Learning and Cybernetics*, Kunming, 12-15 July,2008.

[60] M. Khalilia, S. Chakraborty and M. Popescu, (2011) Predicting disease risks from highly imbalanced data using random forest [Online].Available:biomedicalcentral.com/1472-67947/11/51.

[61] J. Chena, Y. Y. Tanga, B. Fanga and C. Guoa, "In silico prediction of toxic action mechanisms of phenols for imbalanced data with Random Forest learner," *Journal of Molecular Graphics and Modelling*, vol. 35, pp. 21–27, 2012

[62] H. Lia and J. Suna, "Forecasting business failure: The use of nearest-neighbour support vectors and correcting imbalanced samples e Evidence from the Chinese hotel industry," *Tourism Management*, vol. 33, pp. 622-634, 2012

[63] Y. Liua , X. Yua, J. X. Huangb and A. An, "Combining integrated sampling with SVM ensembles for learning from imbalanced datasets," *Information Processing and Management*, vol. 47 , pp. 617–631, 2011

[64] Y. Zhang et al.," Using ensemble methods to deal with imbalanced data in predicting protein–protein interactions," *Computational Biology and Chemistry*, vol. 36, pp. 36–41, 2012

[65] J. Xiao, L. Xieb , C. Hea and X. Jiangc, "Dynamic classifier ensemble model for customer classification with imbalanced class distribution," *Expert Systems with Applications*, vol. 39, pp. 3668–3675, 2012

[66] A. Ghazikhania, R. Monsefia and H. S. Yazdi, "Ensemble of online neural networks for non-stationary and imbalanced data streams," *Neurocomputing*, 2013

[67] M. D. Martino, A. Fernández, P. Iturralde and F. Lecumberry, "Novel classifier scheme for imbalanced problems," *Pattern Recognition Letters*, vol. 34, pp. 1146–1151, 2013.