A novel ensemble classifier by combining sampling and genetic algorithm to combat multiclass imbalanced problems

Archana Purwar* and Sandeep Kumar Singh

Department of Computer Science and Information Technology,
JIIT Noida, India
Email: archana.purwar@gmail.com
Email: sandeepk.singh@jiit.ac.in
*Corresponding author

Abstract: To handle datasets with imbalanced classes is an exigent problem in the area of machine learning and data mining. Though a lot of work has been done by many researchers in the literature for two-class imbalanced problems, the multiclass problems still need to be explored. In this paper, we propose sampling and genetic algorithm based ensemble classifier (SA-GABEC) to handle imbalanced classes. SA-GABEC tries to find the best subset of classifiers for a given sample that is precise in predictions and can create an acceptable diversity in features subspace. These subsets of classifiers are fused together to give better predictions as compared to a single classifier. Moreover, this paper also proposes modified SA-GABEC which performs the feature selection before applying sampling and outperforms SA-GABEC. The performance of the proposed classifiers is evaluated and compared with GAB-EPA, Adaboost and bagging using minority class recall and extended G-mean.

Keywords: feature extraction; diversity; genetic algorithm; ensemble learning; multiclass imbalanced problems.

Reference to this paper should be made as follows: Purwar, A. and Singh, S.K. (2020) 'A novel ensemble classifier by combining sampling and genetic algorithm to combat multiclass imbalanced problems', *Int. J. Data Analysis Techniques and Strategies*, Vol. 12, No. 1, pp.30–42.

Biographical notes: Archana Purwar has been working as an Assistant Professor in Jaypee Institute of Information Technology, Noida, India. During her teaching career of more than 13 years, she taught subjects such as database systems, software engineering, object oriented programming, computer architecture and organisation. Her area of interest lies in data mining and information retrieval.

Sandeep Kumar Singh has been working as an Associate Professor in Jaypee Institute of Information Technology, Noida, India. During his teaching career of more than 15 years, he taught subjects such as object oriented programming, software engineering, data structures, computer programming, e-commerce and social web and high performance software engineering. He participated in many international and national conferences in India and abroad and actively involved as a reviewer for many international journals and conference proceedings. He has been actively involved in organising International Conference on Contemporary Computing (IC3) in year 2008–2017 as web and payment gateway chair. He has guided approximately 100 post graduate and

graduate project thesis. He also has been involved with many universities of India as question paper setter, head examiner, thesis evaluator and expert of selection committee. Currently, he is supervising four PhD scholars. He has successfully completed one PhD under his supervision.

1 Introduction

Classification is one of the vital errands in the area of machine learning and data mining which provides a wide range of tools and techniques to extract the useful information (Singh et al., 2015; Nag et al., 2015). In recent times, imbalanced class distribution has drawn a lot of attention in both academia and industry. Most benchmark classification algorithms have shown that skewed class distribution of a dataset favours the majority class causing poor accuracy for minority class (Tan et al., 2015; Dietterich and Bakiri, 1991). Majority class has a large number of examples as compared to minority class in an imbalanced dataset. An uneven distribution of class in the training procedure can leave the poor precision for the minority class (es) however high accuracy for majority class (es). Classical classification algorithms do not produce good results in case of imbalanced datasets as they are developed to generalise from training data samples and produce the easiest hypothesis that fits the data.

Various solutions have been put forward to handle the class imbalance problem at data as well as algorithmic level. Data level approaches balance the skewed class distribution using a pre-processing step such as sampling. The algorithmic approaches try to modify classical algorithm by adding bias to the minority class examples. Besides, cost sensitive (Freitas et al., 2007) as well as ensemble techniques such as BEV (Li, 2007), SMOTEBoost (Chawla et al., 2003) have emerged as prevalent and key means to handle class imbalance problem efficiently. Nonetheless, most of researchers so far have developed the approaches on two-class imbalance problems in the literature.

Many real life problems consist of more than two classes with skewed class distributions, such as protein fold classification (Zhao et al., 2008; Chen et al., 2006; Tan et al., 2003) and weld flaw classification (Liao, 2008). These multiclass imbalance problems face new challenges that are not spotted in binary class problems. Zhou et al. demonstrated that adding of misclassifications costs to different classes in multiclass problems is tough as compared to binary class problems (Jin and Zhang, 2007). Consequently, researchers have tried to find how to make use of theoretical and empirical performance advantages of two class algorithms for problems having multiple classes. One conventional approach performing so is to use class decomposition methods. These methods split the multi-class problems into a group of binary class problems. As a result, researchers may train classical two-class classifiers on every set of binary problems that can later on be fused into an ensemble to resolve multiclass classes present in the dataset. These techniques include ‘one-versus-all’ (OVA) and ‘error correcting output codes’ (ECOC) (Dietterich and Bakiri, 1991). OVA method builds ‘C’ classifiers, if dataset consists of C classes. Each classifier in OVA assumes one of the classes as positive class and rest are considered into negative class. For each test instance, all classifiers return a probability estimate for the instance. Then, result will be computed by overall probability estimate. ECOC method employs error correcting codes to learn a decomposition ensemble of classifiers. Although these approaches are able to resolve multiple class

issue, these methods provoke imbalanced distributions (Tan et al., 2003). Moreover, fusing results from different classifiers for different sub problems may generate errors to classify the examples (Jin and Zhang, 2007; Valizadegan et al., 2008). Consequently, there is a need to design a more effective and efficient method to handle multiclass imbalance problems (Purwar and Singh, 2014). In this paper, we intend to design an approach to tackle multiclass imbalanced problem.

The rest of this paper is organised into four sections. Section 2 describes the research progress in imbalanced learning. Section 3 introduces a proposed approach to classify the data in the presence of skewed class distribution. Section 4 shows the experimental studies of the proposed approach with the results. Lastly, Section 5 concludes this paper with future directions.

2 Review of imbalanced classification

Class distributions, i.e., the ratio in which number of instances are present in the dataset, have a significant role in learning of classifiers. Class imbalance problem arises when examples of one class (majority) are overrepresented as compared to second class (minority). Numerous approaches developed by the researchers can be categorised into three categories namely data group, algorithmic group and cost sensitive group. Data group approaches make use of a pre-sampling method. Different strategies of pre-sampling alter skewed class proportions (Chawla et al., 2003; Lin et al., 2013; Tao et al., 2006). These pre-sampling methods remove imbalance class distribution in the training dataset. These do so either by oversampling the minority class or by under sampling the majority class. In particular, random oversampling (ROS) method generates minority class examples randomly. The use of ROS may also bring in redundancy and trigger over-fitting (He and Garcia, 2009). As an alternative of randomly duplicating examples, synthetic minority oversampling technique (SMOTE) (Chawla et al., 2003; Anil Kumar and Ravi, 2008) and its enhanced forms namely Borderline-SMOTE (Han et al., 2005) and adaptive synthetic sampling (Adasyn) (He et al., 2008), produce synthetic samples for the minority class which make the decision boundary more general. Whereas, undersampling methods, such as random undersampling (RUS) and one-sided selection (OSS) (Kubat and Matwin, 1997) trim majority class examples and make distribution of classes balanced. In the algorithmic group, existing approaches are modified to take into account the importance of minority examples (Quinlan, 1991; Zadrozny and Elkan, 2001; Wu and Chang, 2005). Finally, cost-sensitive approaches combines data as well as algorithm approaches (Freitas et al., 2007; Chawla et al., 2008). In addition to this one, ensemble classifiers are employed to get better performance of a single classifier. Adaboost and bagging are the most common ensemble classifiers (Wang and Yao, 2012). Zhang et al. (2014) also designed a novel ensemble classifier IRUST using Inverse random sampling and random tree (IRUST) to create more diversity in feature subspace. Meanwhile, they have advised to generalise it to multiclass imbalanced datasets. The feature selection approaches (Zheng et al., 2004; Wasikowski and Chen, 2010) have also been investigated to deal the class imbalance problem where dataset consists of high-dimensions.

Aforementioned approaches were mainly designed for two-class problems of imbalanced datasets, while their efficiency was not tested well for multiclass problems. Sun et al. (2006) proposed a cost-sensitive ensemble algorithm, named AdaC2.M1genetic

to handle multiclass problems in which they employed genetic algorithm to explore the best cost setup for of each class. To deal with multiclass problem, a novel negative correlation-based algorithm, known as AdaBoost.NC was designed by Wang and Yao (2012) in which a negative correlation learning, is added to basic AdaBoost. However, establishing a cost matrix for a class imbalance problem in such methods is a tough task. It has been proved that cost-sensitive methods are not realistic as one would expect for imbalance classification problems (Sun et al., 2006).

There exist some cost free learning approaches in the class imbalance problems. Zhang and Hu (2014) developed a cost insensitive approach as cost free learning (CFL) that makes the best use of normalised mutual information of the targets as well as decision results of classifiers. They had applied Powell algorithm to optimise normalised mutual information that is a non-linear optimisation without calculating the derivatives. But, this approach adds an extra computational cost over the existing techniques.

Another set of techniques arises when the use of ensembles of classifiers is taken into consideration. Ensembles are developed to enhance the power of weak classifier as it improves the accuracy of the prediction model extensively. It has also been also proved to be useful to handle the imbalanced dataset problem (Tao et al., 2006; Abdi and Hashemi, 2013). Diversity and accuracy are two crucial factors in designing of ensemble classifier (Chandra and Yao, 2006; Ho, 1998). These factors look for most suitable subset of classifiers that are precise while making their predictions. Moreover, they are capable of producing a tolerable diversity if fused together as an ensemble classifier (Abdi and Hashemi, 2013).Therefore, an ensemble classifier which has high accuracy, is composed of highly accurate and diverse ensemble members. Abdi and Hashemi (2013) had designed genetic algorithm-based ensemble pruning algorithm (GAB-EPA) to handle multi-class problems that use weighted vote of ensemble members for the prediction of test instance. Further, Li et al. also developed adaptive multiple classifier system to handle the problem due to imbalanced lanced classes (Yijing et al., 2016).

We have used sampling, feature selection as well as ensemble technique and designed SA-GABEC that combines sampling with genetic algorithm and uses C4.5 as base classifier to develop an ensemble classifier.

3 Proposed algorithm: SA-GABEC

This paper proposes sampling with Genetic algorithm based ensemble classifier (SA-GABEC) to handle imbalanced classes in predictive classification present in the dataset. Proposed algorithm is developed by fusing sampling with genetic algorithm to tackle imbalanced problem present in multiclass datasets. Sampling is done like asymmetric bagging proposed by Tao et al. (2006) to deal class imbalanced problem. This method firstly undersamples the minority class and creates multiple balanced training sets (multiple designs). These different designs are used to train the model. Random subspace method (Ho, 1998) makes use of bootstrapping and aggregation as bagging does. In contrast to bagging (Akour et al., 2017) that bootstraps training samples, random subspace method bootstrap the features present in the dataset. Therefore, we have employed genetic algorithm to generate different feature subspaces from bootstrapped sample set to provide the diversity. C4.5 classifier is used as a base classifier. The pseudo-code of SA-GABEC is illustrated in Algorithm 1.

Algorithm 1 Proposed algorithm: SA-GABEC

Step 1: Build M sets (models) of training datasets. Each set consists of N_{min} samples present in minority class along with N_{max} samples from each remaining classes using sampling with and without replacement. The value M is calculated by the following equation.

$$M = \left\lceil N_{maj} \Big/ N_{\min} \right\rceil \tag{1}$$

where, N_{maj} and N_{min} shows the count of instances in the largest majority class and smallest minority class of training dataset respectively. N_{max} is the number of instances from other classes (except smallest minority class) whose value is same as N_{min}.

Step 2: Apply the genetic algorithm on each set to find the optimal set of features.

Generate initial population of size N by creating randomly and evaluated using a fitness function. Each individual (chromosome) in the population has length L. Each chromosome is a binary string of 1 and 0. The value '1' specifies that feature indexed by particular position in chromosome is present and '0' shows that feature is not present at that position.

For each generation, do the following steps till the stopping condition (Stall Limit and Number of generations) meets:

a Choose best parents, do crossover and apply mutation.

b Generate new population N by elite, crossover and mutation children.

c Calculate fitness of all chromosomes using KNN (K-nearest neighbor).

d Transfer N best chosen individuals to create new population for next generation.

Step 3: Train these M models (after feature selection) using C4.5 algorithm.

Step 4: Allocate a weighted score to each model in ensemble using its averaged receiver operating characteristics (ROC) of multiple classes, and then normalise the weights using following equation.

$$wi = \frac{AUC_i}{\sum_{i=1}^{M} AUC_i} \tag{2}$$

$$AUC(Area\ under\ ROC\ Curve) = \frac{1 + True\ Positive\ Rate(TPR) - False\ Positive\ Rate(FPR)}{2}$$

where TPR = TP / (TP + FN) and FPR = FP / (FP + TN)

TP, FN, FP, TN are true positive, false negative, false positive and true negative respectively which are outcomes of a binary classifier.

Step 5: For each test data, class of each testing instance will be obtained with majority / weighted vote of these models.

Further, we also have modified SA-GABEC shown in Algorithm 2. It is a modified version of Algorithm 1. Modified SA-GABEC applies the genetic algorithm in the step 1 and then applies sampling (i.e.) after the step 1. Algorithm 2 stands on concept that the size of the different training dataset after sampling may be lower than total numbers of features in the dataset, that may lead to problem of overfitting. To avoid the problem of overfitting, feature selection may be applied to the imbalanced dataset before sampling.

To tackle the feature selection, we have chosen meta heuristic algorithm as genetic algorithm in our proposed work. As it is one of most enveloping heuristic search technique to get the best optimised solution for a given problem based on inheritance, mutation, selection and some other techniques.

Algorithm 2 Proposed modified SA-GABEC

Step 1:	First, apply the genetic algorithm on entire training dataset to find the optimal set of features as done in step 2 of SA-GABEC.
Step 2:	Build M sets (models) of training datasets using equation 1 from dataset obtained from above step.
Step 3:	Do similar to SA-GABEC.
Step 4:	Do similar to SA-GABEC.
Step 5:	Do similar to as SA-GABEC.

4 Experimental studies

4.1 Datasets

We had accomplished experiments to assess the efficiency of designed method SA-GABEC. We have also compared it with existing methods such as bagging, boosting and GAB-EPA approach) (Wang and Yao, 2012; Abdi and Hashemi, 2013) on two datasets. These were gathered using the UCI machine learning repository (Frank and Asuncion, 2010). Table 1 shows the main characteristics of the datasets used in our experiments. For each dataset, it records name, number of classes, number of instances, number of features, and class distribution. Class distribution column shows the count of instances for every class separated by '/'.

Table 1 Characteristics of benchmark datasets

Name of dataset	*Number of classes*	*Number of instances*	*Number of features*	*Class distribution*
Dermatology	6	366	34	112/72/61/52/49/20
Satimage	6	6,435	36	1533/703/1358/626/707/1508

4.2 Experimental settings

To test the efficiency of proposed work, we used twenty and eighty percent of data for testing and training respectively.

The genetic algorithm used in proposed method was implemented using Matlab software with version R2013. The initial population (N) is chosen as 100. Length of chromosome (L) is set as number of features present in dataset. The stopping condition for genetic algorithm is set by using stall limit and maximum number of generations whose value is set as 150 and 300 respectively. The value of other parameters such as crossover, probability of crossover, mutation , mutation probability, selection scheme, elite count are set as arithmetic crossover, 0.8, uniform mutation, 0.1, tournament size of 2 and two respectively.

Table 2 Selected features of different sets of dermatology dataset for Algorithm 1

Model	*Selected features with index no.*
1	1, 3, 4, 5, 6, 7, 10, 14, 15, 16, 17, 18, 21, 23, 25, 27, 30, 33
2	1, 2, 4, 6, 8, 9, 11, 13, 15, 16, 17, 25, 26, 27, 28, 30, 31
3	5, 6, 9, 10, 12, 13, 18, 19, 21, 26, 30, 31, 32
4	1, 2, 4, 6, 8, 9, 11, 13, 15, 16, 17, 25, 26, 27, 28, 30, 31
5	2, 4, 5, 7, 9, 10, 13, 14, 18, 21, 23, 26, 27, 28, 30, 33
6	1, 3, 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 18, 21, 23, 27, 28, 30, 33

The results after applying genetic algorithm on different sets of proposed Algorithm 1 are mentioned in Table 2 for dermatology dataset. Each model after extraction of features is trained using base learner J48 and weighted ROC using Weka is calculated shown in Table 3. We have implemented J48 classifier in Weka (Witten and Frank, 2005), with the default settings such as confidence factor, minimum number of instances in leaf node and others.

Table 3 Weighted ROC and normalised ROC of models for dermatology dataset using Algorithm 1

Model	*Weighted ROC*	*Normalised ROC*
1	0.938	0.156
2	0.951	0.159
3	0.972	0.162
4	0.983	0.164
5	0.913	0.152
6	0.961	0.16

Table 4 Selected features of dermatology dataset for Algorithm 2

Index no. of selected features
2, 4, 5, 6, 8, 10, 11, 13, 15, 16, 18, 19, 21, 23, 25, 27, 28, 31

Table 5 Selected features of satimage dataset for Algorithm 2

Index no. of selected features
4, 5, 10, 12, 18, 20, 25, 27, 33, 36

Table 6 Weighted ROC and normalised ROC of models for dermatology dataset using Algorithm 2

Model	*Weighted ROC*	*Normalised ROC*
1	0.928	0.164
2	0.976	0.173
3	0.952	0.168
4	0.91	0.161
5	0.938	0.166
6	0.951	0.168

Table 7 Weighted ROC and normalised ROC of models for satimage dataset using Algorithm 2

Model	*Weighted ROC*	*Normalised ROC*
1	0.937	0.329
2	0.996	0.349
3	0.915	0.321

We have also done experiments to assess the efficiency of the Algorithm 2. Features obtained after applying genetic algorithm on datasets using proposed Algorithm 2 are shown in Tables 4 and 5 for dermatology and satimage datasets respectively. ROC scores of each model (using weka) after applying the C4.5 are tabulated in Tables 6 and 7.

4.3 Evaluation metrics

In case of imbalanced learning scenario, traditional evaluation metric such as classification accuracy (Sakthivel et al., 2011; Mokeddem et al., 2016; Purwar and Singh, 2015) is not an appropriate choice because results may be biased toward the majority class (Chawla et al., 2003; He and Garcia, 2009). For example, when a smallest class is characterised by only 1% of training data, it is expected to attain a high accuracy of 99%. However; this output is misleading to those problems such as credit fraud detection where our focus is to classify the smallest class correctly. Hence, we have adopted two major assessment metrics for analysis shown by (3) and (4).

4.3.1 Recall (Abdi and Hashemi, 2013)

Recall is a single class performance and it is a measure of completeness. It is capable of finding that how many instances are correctly labelled by the classifier.

$$Recall = \frac{TP_j}{TP_j + FN_j} \tag{3}$$

where

j 1, 2.., C

C count of classes in dataset

TP_j count of correctly classified instances of class j

FN_j count of misclassified instances of class j.

4.3.2 Extended G-mean (Sun et al., 2006)

Extended G-mean is an evaluation metric to evaluate the overall performance of a classifier in case of multiple classes (Wang and Yao, 2012). It shows that how good a classifier can balance the identification among different classes.

$$Extended\ G-mean = \sqrt[C]{\prod_{j=1}^{C} Recall} \tag{4}$$

4.4 Results and analysis

We have assessed the performance of proposed Algorithm 1 and its modified version, i.e., Algorithm 2 on benchmark datasets under study in terms of recall mentioned in Tables 8, 9 and 10 respectively. Results show that recall of the minority class (class having minimum instances) is 1 and 0.8 for dermatology and satimage datasets respectively. It indicates that every instance of minority class for dermatology dataset and 80% of minority instances for satimage dataset are correctly classified. Proposed algorithm is also compared with existing methods namely bagging, boosting and GAB-FPA using extended G-mean metric and recall of minority class. Comparison results are summarised in Tables 11 and 12. It is apparent from tabulated results that recall of minority class is improved by Algorithm 2 as compared to other methods.

Table 8 Recall values for each class in dermatology dataset for Algorithm 1

Class	*Recall*
Psoriasis	0.86
Seboreic dermatitis	0.80
Lichen planus	0.90
Pityriasis rosea	0.85
Cronic dermatitis	1
Pityriasis rubra pilaris (minority class)	1

Table 9 Recall values for each class in dermatology dataset for Algorithm 2

Class	*Recall*
Psoriasis	1
Seboreic dermatitis	0.88
Lichen planus	0.90
Pityriasis rosea	1
Cronic dermatitis	1
Pityriasis rubra pilaris (minority class)	1

Table 10 Recall values for each class in satimage dataset for Algorithm 2

Class	*Recall*
Red soil	0.98
Cotton crop	0.96
Grey soil	0.95
Damp grey soil (minority class)	0.80
Soil with vegetation stubble	0.95
Very damp grey soil	0.87

Table 11 Minority class (smallest class) recall and G-mean by bagging, Adaboost, GAB-EPA, and proposed method for dermatology dataset

	Recall (minority class)	*G-mean*	*Reference*
Proposed method-Algorithm 2	1.0	0.96	Proposed method-2
Proposed method-Algorithm1	1.0	0.90	Proposed method-1
GAB-EPA	0.80	0.93	Abdi and Hashemi (2013)
Adaboost	0.92	0.92	Witten and Frank (2005)
Bagging	0.80	0.73	Witten and Frank (2005)

Table 12 Minority class (smallest class) recall and G-mean by bagging, Adaboost, GAB-EPA, and proposed method for Satimage dataset

	Recall (minority class)	*G-mean*	*Reference*
Proposed method-Algorithm 2	.80	0.96	Proposed method-2
GAB-EPA	0.60	0.87	Abdi and Hashem (2013)
Adaboost	0.57	0.85	Witten and Frank (2005)
Bagging	0.56	0.85	Witten and Frank (2005)

5 Conclusions with future work

A new sampling with Genetic algorithm based ensemble classifier, SA-GABEC and modified SA-GABEC are designed to tackle class imbalanced problem of classification. The key characteristics of SA-GABEC are three fold. First, it creates the multiple sets by using sampling with and without replacement such that each set consists of all the samples of minority class. Secondly, it applies genetic algorithm on these training sets to get the possible set of features of a given dataset. Finally extracted feature set from each training model is used for learning purpose of classifier and then merge them through normalised ROC. Modified SA-GABEC first applies the genetic algorithm on the dataset. Then, it applies undersampling the majority class (es) to produce different datasets that are used in learning process of the classifier. Lastly, different designs are fused together to test the training dataset. Proposed approaches are assessed through Recall and extended G-mean over two datasets. Results have shown that modified SA-GABEC is performing better than SA-GABEC and other existing methods taken under study.

In future, we will focus to investigate performance of SA-GABEC with different base classifier other than C4.5 (used in this work) like nearest neighbour, bayesNet, support vector machine, neural network and others. Further, different combination of base classifiers in each training model of proposed algorithm could be examined such as C4.5 with SVM and MLP. In addition to this, impact of noise on proposed approach can also be examined.

References

Abdi, L. and Hashemi, S. (2013) 'GAB-EPA: a GA based ensemble pruning approach to tackle multiclass imbalanced problems', in *Asian Conference on Intelligent Information and Database Systems*, Springer.

Akour, M., Alsmadi, I. and Alazzam, I. (2017) 'Software fault proneness prediction: a comparative study between bagging, boosting, and stacking ensemble and base learner methods', *International Journal of Data Analysis Techniques and Strategies*, Vol. 9, No. 1, pp.1–16.

Anil Kumar, D. and Ravi, V. (2008) 'Predicting credit card customer churn in banks using data mining', *International Journal of Data Analysis Techniques and Strategies*, Vol. 1, No. 1, pp.4–28.

Chandra, A. and Yao, X. (2006) 'Evolving hybrid ensembles of learning machines for better generalisation', *Neurocomputing*, Vol. 69, No. 7, pp.686–700.

Chawla, N.V. et al. (2003) 'SMOTEBoost: improving prediction of the minority class in boosting', in *European Conference on Principles of Data Mining and Knowledge Discovery*, Springer.

Chawla, N.V. et al. (2008) 'Automatically countering imbalance and its empirical relationship to cost', *Data Mining and Knowledge Discovery*, Vol. 17, No. 2, pp.225–252.

Chen, K., Lu, B-L. and Kwok, J.T. (2006) 'Efficient classification of multi-label and imbalanced data using min-max modular classifiers', in the *2006 IEEE International Joint Conference on Neural Network Proceedings*, IEEE.

Dietterich, T.G. and Bakiri, G. (1991) 'Error-correcting output codes: a general method for improving multiclass inductive learning programs', in *AAAI*, Citeseer.

Frank, A. and Asuncion, A. (2010) *UCI Machine Learning Repository*, University of California, Irvine, School of Information and Computer Sciences [online] http://archive.ics.uci.edu/ml.

Freitas, A., Costa-Pereira, A. and Brazdil, P. (2007) 'Cost-sensitive decision trees applied to medical data', in *International Conference on Data Warehousing and Knowledge Discovery*, Springer.

Han, H., Wang, W-Y. and Mao, B-H. (2005) 'Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning', in *International Conference on Intelligent Computing*, Springer.

He, H. and Garcia, E.A. (2009) 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9, pp.1263–1284.

He, H. et al. (2008) 'ADASYN: Adaptive synthetic sampling approach for imbalanced learning', in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, IEEE.

Ho, T.K. (1998) 'The random subspace method for constructing decision forests', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp.832–844.

Jin, R. and Zhang, J. (2007) 'Multi-class learning by smoothed boosting', *Machine Learning*, Vol. 67, No. 3, pp.207–227.

Kubat, M. and Matwin, S. (1997) 'Addressing the curse of imbalanced training sets: one-sided selection', in *Proceedings of the Fourteenth International Conference on Machine Learning*.

Li, C. (2007) 'Classifying imbalanced data using a bagging ensemble variation (BEV)', in *Proceedings of the 45th Annual Southeast Regional Conference*, ACM.

Liao, T.W. (2008) 'Classification of weld flaws with imbalanced class data', *Expert Systems with Applications*, Vol. 35, No. 3, pp.1041–1052.

Lin, M., Tang, K. and Yao, X. (2013) 'Dynamic sampling approach to training neural networks for multiclass imbalance classification', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 24, No. 4, pp.647–660.

Mokeddem, S., Atmani, B. and Mokaddem, M. (2016) 'An effective feature selection approach driven genetic algorithm wrapped Bayes naïve', *International Journal of Data Analysis Techniques and Strategies*, Vol. 8, No. 3, pp.220–243.

Nag, B.N., Han, C. and Yao, D-q. (2015) 'Information enhancement in data mining: a study in data reduction', *International Journal of Data Analysis Techniques and Strategies*, Vol. 7, No. 1, pp.3–20.

Purwar, A. and Singh, S.K. (2014) 'Issues in data mining: a comprehensive survey', in *2014 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, IEEE.

Purwar, A. and Singh, S.K. (2015) 'Hybrid prediction model with missing value imputation for medical data', *Expert Systems with Applications*, Vol. 42, No. 13, pp.5621–5631.

Quinlan, J.R. (1991) 'Improved estimates for the accuracy of small disjuncts', *Machine Learning*, Vol. 6, No. 1, pp.93–98.

Sakthivel, N. et al. (2011) 'Decision support system using artificial immune recognition system for fault classification of centrifugal pump', *International Journal of Data Analysis Techniques and Strategies*, Vol. 3, No. 1, pp.66–84.

Singh, A., Rana, A. and Ranjan, J. (2015) 'Data mining techniques and its effect in customer relationship management', *International Journal of Data Analysis Techniques and Strategies*, Vol. 7, No. 4, pp.406–427.

Sun, Y., Kamel, M.S. and Wang, Y. (2006) 'Boosting for learning multiple classes with imbalanced class distribution', in *Sixth International Conference on Data Mining (ICDM'06)*, IEEE.

Tan, A.C., Gilbert, D. and Deville, Y. (2003) 'Multi-class protein fold classification using a new ensemble machine learning approach', *Genome Informatics*, Vol. 14, pp.206–217.

Tan, S.C. et al. (2015) 'Evolutionary fuzzy ARTMAP neural networks for classification of semiconductor defects', *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 26, No. 5, pp.933–950.

Tao, D. et al. (2006) 'Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 28, No. 7, pp.1088–1099.

Valizadegan, H., Jin, R. and Jain, A.K. (2008) 'Semi-supervised boosting for multi-class classification', in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer.

Wang, S. and Yao, X. (2012) 'Multiclass imbalance problems: analysis and potential solutions', *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 42, No. 4, pp.1119–1130.

Wasikowski, M. and Chen, X-w. (2010) 'Combating the small sample class imbalance problem using feature selection', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 22, No. 10, pp.1388–1400.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann.

Wu, G. and Chang, E.Y. (2005) 'KBA: kernel boundary alignment considering imbalanced data distribution', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 17, No. 6, pp.786–795.

Yijing, L. et al. (2016) 'Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data', *Knowledge-Based Systems*, Vol. 94, pp.88–104.

Zadrozny, B. and Elkan, C. (2001) 'Learning and making decisions when costs and probabilities are both unknown', in *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM.

Zhang, C.-X. et al. (2014) 'IRUSRT: a novel imbalanced learning technique by combining inverse random under sampling and random tree', *Communications in Statistics-Simulation and Computation*, Vol. 43, No. 10, pp.2714–2731.

Zhang, X. and Hu, B-G. (2014) 'A new strategy of cost-free learning in the class imbalance problem', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 12, pp.2872–2885.

Zhao, X.M. et al. (2008) 'Protein classification with imbalanced data', *Proteins: Structure, Function, and Bioinformatics*, Vol. 70, No. 4, pp.1125–1132.

Zheng, Z., Wu, X. and Srihari, R. (2004) 'Feature selection for text categorization on imbalanced data', *ACM Sigkdd Explorations Newsletter*, Vol. 6, No. 1, pp.80–89.