On the Generation of Accurate Predictive Model from Highly Imbalanced Data with Heuristics and Replication Techniques

Nittaya Kerdprasop and Kittisak Kerdprasop
Data Engineering Research Unit, School of Computer Engineering,
Suranaree University of Technology, Nakhon Ratchasima 30000 Thailand
nittaya@sut.ac.th, kittisakThailand@gmail.com

Abstract

Recent advancement in the field of life science data mining has inspired researchers and healthcare professionals to apply this novel technology to obtain descriptive patterns and predictive models from biomedical and healthcare databases. The discovery of hidden biomedical patterns from large clinical database can uncover potential knowledge to support prognosis and diagnosis decision makings. However, clinical application of data mining algorithms has a severe problem of low predictive accuracy rate that hampers their wide usage in the clinical environment. We thus focus our study on the improvement of predictive accuracy of the models created from the data mining algorithms. Our main research interest concerns the problem of learning a classification model from a multiclass data set with low prevalence rate of some minority classes. With such data characteristics, directly applying classification data mining techniques such as decision tree induction, regression analysis, neural networks, or support vector machines yields a suboptimal model in terms of predictive accuracy rate. To remedy the imbalanced class distribution among data instances, we apply random over-sampling and synthetic minority over-sampling (SMOTE) techniques to increase the predictive performance of the learned model. In our preliminary study, we consider specific kinds of primary tumors occurring at the frequency rate less than one percent as rare and minority classes. From the experimental results, the SMOTE technique gave a high specificity model, whereas the random over-sampling produced a high sensitivity classifier. The precision performance of a classification model obtained from the random over-sampling technique is on average much better than the model learned from the original imbalanced data set. We then extend our study by designing the heuristic based method to cope with the abundance of irrelevant feature that causes the decrease in learning time and sometimes lower the accuracy rate. The over-sampling technique and the heuristic-based feature selection are coupled as a data preparation method to deal with imbalanced data sets with many irrelevant features. The experimental results on arrhythmia and communities-and-crime data sets show significant improvement on the predicting accuracy, specificity, and sensitivity of the induced models.

Keywords: *Classification model, Learning from imbalanced data, Heuristic-based feature selection, Data replication, Over-sampling technique, Rare case prediction*

1. Introduction

The discovery of hidden patterns from large databases can uncover knowledge to support decision making. Researchers and practitioners in several areas have successfully applied data mining technology to obtain descriptive patterns and predictive models from their databases. However, data mining application in some

specific areas such as biomedical [1, 2] and clinical professions is still in a limited scope due to a severe problem of low predictive accuracy of the induced model. Low accuracy of the induced model is due to the multi-features and imbalanced characteristics inherent in the dataset.

Data mining is about building a model that can best characterize underlying data and accurately predict the class of unlabelled data. The quality of data mining model depends directly on the quality of the training data. Data of low quality are those that contain noise, missing values, and class imbalance. A data set is imbalanced if the number of data instances in one class is much more than those in other classes. In the presence of class imbalance, data mining models are biased toward the majority class in such a way that the models can predict the majority class correctly but data instances from the minority class tend to be incorrectly predicted. This research issue of learning from highly imbalance datasets has recently gained much attention from the data mining and machine learning community [3, 4, 5, 6, 7]. We refer to this problem as rare class prediction.

To solve the problem of biased learning toward the majority class, many researchers consider the sampling techniques for manipulating class distribution such that rare cases could be sufficiently represented in the training data. The basic sampling techniques that have been applied are under-sampling and over-sampling. Under-sampling alters the class distribution by removing data instances from the minority class, whereas over-sampling duplicates data instances in the minority class [7, 8, 9]. The under-sampling technique may remove good representatives, while over-sampling may cause the over-fitting problem. The classification model shows over-fitting characteristics when the model can classify extremely well on the training data, but perform poorly on other data sets or the unseen data.

We propose the unsupervised feature selection technique to be applied to the training data prior to the application of over-sampling technique replicating the rare case instances to the same proportion to the majority cases. We use a hold-out method that separates test data from the train data to assess the model performance. Our experimental studies on several datasets yield satisfactory results in that the proposed method can induced accurate models for predicting both majority and minority test data instances without incurring the over-fitting problem.

2. Accuracy Measurement Metrics on Classification Models

In data classification, the classifier is evaluated by a confusion matrix. For a binary class problem (positive and negative classes), a matrix is a square of 2×2 as shown in Figure 1. The column represents the outcomes of classifier. The row is a real value of class label. The numbers appeared in each cell of the matrix has different names, that is, TP, FN, FP, and TN. Each acronym can be explained as follows:

TP = true positive, that is, the number of positive cases that are correctly identified as positive,

FP = false positive, that is, the number of negative cases that are incorrectly identified as positive cases,

FN = false negative, that is, the number of positive cases that are misclassified as negative cases, and

TN = true negative, that is, the number of negative cases that are correctly identified as negative cases.

Actual class \ Predicted class	Class = +	Class = –
Class = +	TP	FN
Class = –	FP	TN

Figure 1. A Confusion Matrix to Evaluate Model of the Binary Classification Task

We assess the model performance based on the five metrics: true positive rate (recall or sensitivity), false positive rate, specificity, precision, and F-measure. The computation methods of these metrics are as follows [3, 10]:

$$\text{TP rate (or Recall, Sensitivity)} = \frac{TP}{TP + FN}$$

$$\text{FP rate} = \frac{FP}{TN + FP}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{F - measure} = \frac{2TP}{2TP + FP + FN}$$

Actual class \ Predicted class	Class = A	Class = B	Class = C
Class = A	T_A	F_{B1}	F_{C1}
Class = B	F_{A2}	T_B	F_{C2}
Class = C	F_{A3}	F_{B3}	T_C

Figure 2. A Confusion Matrix of the Three-class Classification Task

For the case of multiclass classification, a confusion matrix is a square of N ×N, where N is the number of classes. The classifier's performance measurement is computed per class. For instances, when N is 3, the confusion matrix can be shown as in Figure 2, and the sensitivity, specificity, and precision values can be computed as follows:

$$\text{Sensitivity (or recall) of class A} = \frac{T_A}{T_A + F_{B1} + F_{C1}}$$

$$\text{Sensitivity (or recall) of class B} = \frac{T_B}{T_B + F_{A2} + F_{C2}}$$

$$\text{Sensitivity (or recall) of class C} = \frac{T_C}{T_C + F_{A3} + F_{B3}}$$

$$\text{Specificity of class A} = \frac{T_B + T_C}{T_B + F_{A2} + T_C + F_{A3}}$$

$$\text{Specificity of class B} = \frac{T_A + T_C}{T_A + F_{B1} + T_C + F_{B3}}$$

$$\text{Specificity of class C} = \frac{T_A + T_B}{T_A + F_{C1} + T_B + F_{C2}}$$

$$\text{Precision of class A} = \frac{T_A}{T_A + F_{A2} + F_{A3}}$$

$$\text{Precision of class B} = \frac{T_B}{T_B + F_{B1} + F_{B3}}$$

$$\text{Precision of class C} = \frac{T_C}{T_C + F_{C1} + F_{C2}}$$

3. Preliminary Results on Primary Tumor Prediction

Human body is made up of many types of cells. Living cells grow and divide to produce new cells in an orderly and controlled manner. However, cell production process can go wrong by continuing to produce new cells even when they are not needed. The result of such event is a mass of extra tissue called a tumor. A primary tumor refers to a tumor that has been developed at the original site where it first generated [19]. A tumor can be benign, which means it does not a cancerous one, or malignant that causes cancer. The cancerous cells can invade nearby tissue or spread to cause secondary tumor in other parts of the body. When the spread occurs, an effective treatment becomes a difficult task. Detecting tumor at its original site is therefore important to a successful treatment planning. In this research study, we focus on the problem of detecting and correctly classifying specific types of primary tumors.

In machine learning and data mining, classifying primary tumor data [6] is a difficult task due to the multiclass and imbalanced characteristics inherent in the data set. Many learning algorithms in the past have been proposed to solve the binary classification

problem successfully. The problem concerns the finding of a classification model from a given data set to predict either positive, or negative class labels for the new unseen examples. For the data domains with more than two classes, such as text categorization and medical diagnosis, efficient data mining algorithms need some extensions to deal with the multiclass problem. Decision tree induction algorithms [1, 13] use the information theoretic approach to handle data with multiclass, whereas other learning algorithms such as support vector machines employ the serial binary classification techniques including one versus all [5, 14] and some other sophisticated techniques [16], [17, 21]. In this preliminary assessment, we study the application of decision tree induction algorithm to the multiclass classification problem. Decision tree has the advantage of understandability over other forms of classification models and it has been widely used in the biomedical domain [9, 11, 12].

Another challenging characteristics of primary tumor data is the class distribution imbalance. From the 21 different kinds of primary tumors, some majority classes such as lung and stomach tumors occur three to five times more often than the average frequency rate, while the minority classes occur less than one percent. The primary tumor data set used in our study contains six minority classes, that is, duoden and small intestine, salivary glands, bladder, testis, cervix uteri, and vagina. The minority classes are often missed out by most data mining algorithms because of their extremely low occurrence.

In this preliminary study, we apply the two over-sampling techniques to the primary tumor data set, that is, random over-sampling and the synthetic minority over-sampling or SMOTE [4]. Over-fitting problem has been observed by applying cross validation and holdout methods for classifier performance analysis. The main objective of this experiment is to investigate the advantages of applying random over-sampling and SMOTE techniques to the primary tumor data set. This data set is highly imbalanced in terms of the class distribution (given in Table 1). Our preliminary hypothesis is that by biasing the class distribution of the minority data, the tree-based learning algorithm may perform better on recognizing the rare classes. To make a fair comparison, we use a holdout method, instead of the traditional method of 10-fold cross validation, to test the classifier performance.

The first step of our experimentation is to duplicate a data record containing only a single case (that is, the case of duoden and small intestine, testis, and vagina tumors) to contain two records of each class of tumor. This duplication step is for the purpose of splitting the original data set into two parts: a train set and a test set. Each data set contains the same amount of cases in each type of primary tumors. The independent test data set contains 171 data records. The train data set is to be copied into 3 versions. The first version contains 171 data records with the same class distribution as the test data set. It is called the imbalanced data set. The second version of the train data is to be over-sampling the minority classes with the SMOTE technique [4]. The third version of train data is for the random over-sampling.

Table 1. Class Distribution of the Primary Tumor Data Set

Tumor class	Number of cases	Distribution (%)	Rare class
Lung	84	24.8%	
Head and neck	20	5.9%	
Esophagus	9	2.6%	
Thyroid	14	4.1%	
Stomach	39	11.5%	

Duoden and small intestine	1	0.5%	*
Colon	14	4.1%	
Rectum	6	1.7%	
Salivary glands	2	0.5%	*
Pancreas	28	8.2%	
Gallbladder	16	4.7%	
Liver	7	2.1%	
Kidney	24	7.1%	
Bladder	2	0.5%	*
Testis	1	0.5%	*
Prostate	10	2.9%	
Ovary	29	8.5%	
Corpus uteri	6	1.7%	
Cervix uteri	2	0.5%	*
Vagina	1	0.5%	*
Breast	24	7.1%	

We prepare the random over-sampling data set by duplicating data records in each class to be almost the same amount. The maximum number of cases in the majority class is 42, and the minimum number of cases after duplicating is 36. This random over-sampling data set contains 848 data records with the same proportion of class distribution (around 4.2%-4.9%). This data set is thus has a class distribution different from the original data set (the imbalanced data). When we test the accuracy of classifier built from this data set with the 10-fold cross validation method, the true positive rate and precision are extremely high. But these values are much lower when we test the classifier with an independent test set that has different class distribution. This is obviously the over-fitting problem. We therefore compare classifiers obtained from different sampling techniques with the holdout method that can better guarantee over-fitting avoidance. The comparative results in terms of true positive rate (recall or sensitivity), false positive rate, precision, F-measure, and specificity are given in Figures 3-7. The symbolic codes for different primary tumor types are as follows:

A = salivary glands,	B = bladder,	C = testis,
D = duoden and small intestine,	E = vagina,	F = corpus uteri,
G = rectum,	H = cervix uteri,	I = liver,
J = esophagus,	K = prostate,	L = thyroid,
M = colon,	N = gallbladder,	O = head and neck,
P = breast,	Q = kidney,	R = pancreas,
S = ovary,	T = stomach,	U = lung.

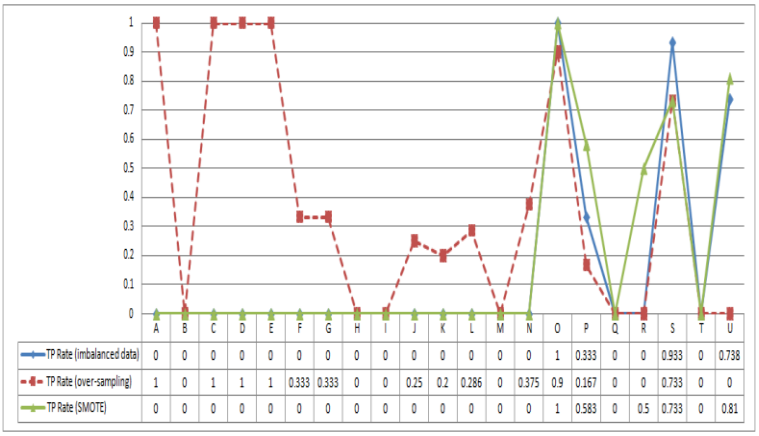


Figure 3. True Positive Rate (recall, sensitivity) Comparison of Primary Tumor Data

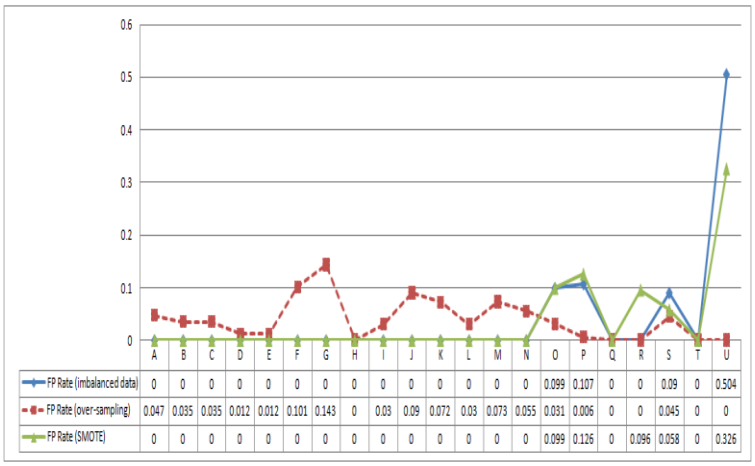


Figure 4. False Positive Rate Comparison of Primary Tumor Data

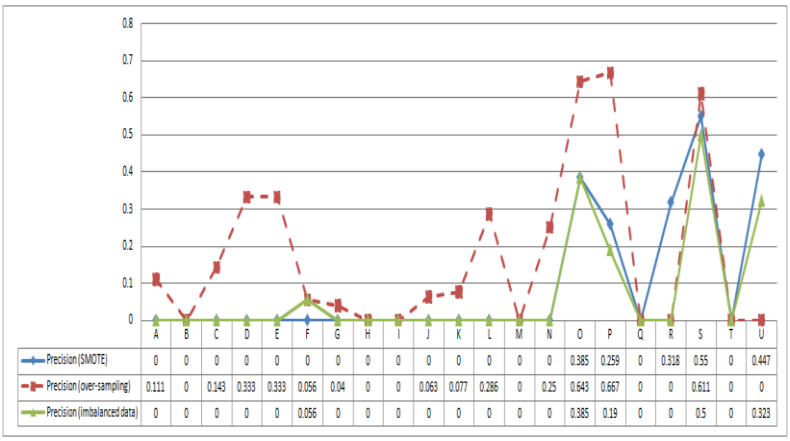
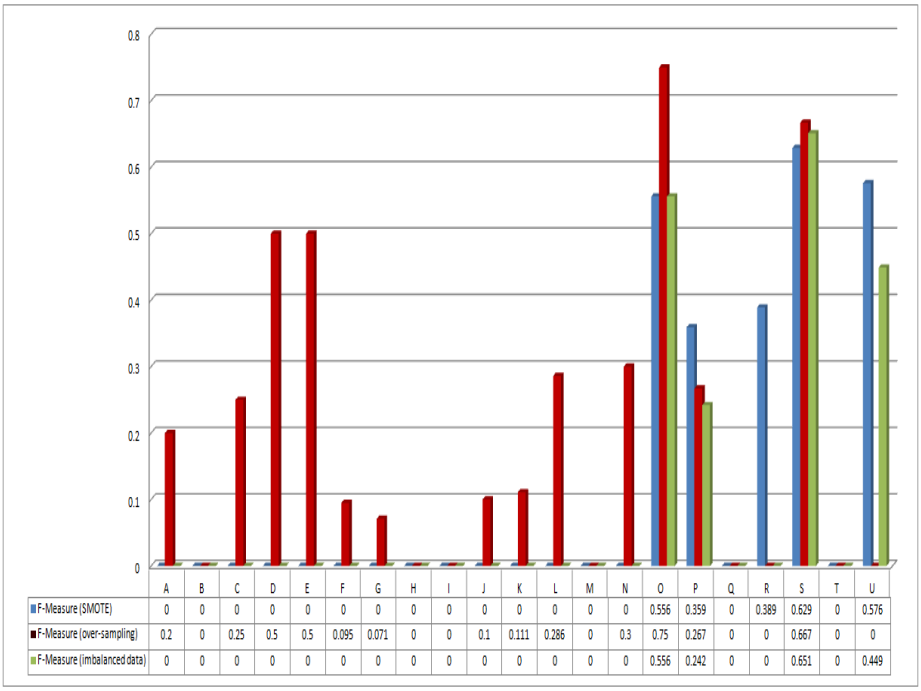


Figure 5. Precision Comparison of Primary Tumor Data



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
F-Measure (SMOTE)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.556	0.359	0	0.389	0.629	0	0.576
F-Measure (over-sampling)	0.2	0	0.25	0.5	0.5	0.095	0.071	0	0	0.1	0.111	0.286	0	0.3	0.75	0.267	0	0	0.667	0	0
F-Measure (imbalanced data)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.556	0.242	0	0	0.651	0	0.449

Figure 6. F-measure Comparison of Primary Tumor Data

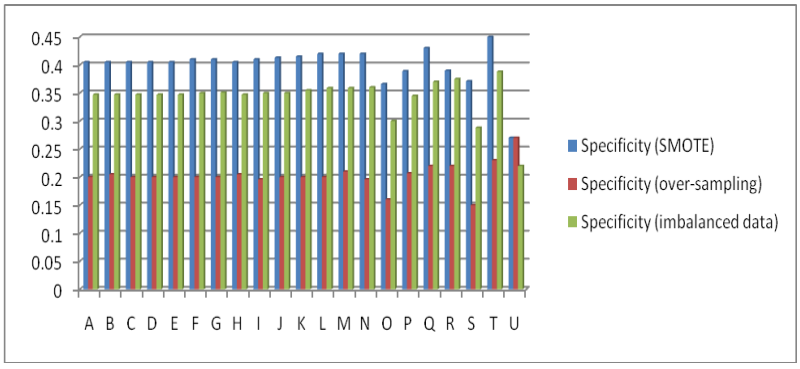


Figure 7. Specificity Comparison of Primary Tumor Data

It can be seen from the experimental results that the random over-sampling and SMOTE techniques can improve the performance of predicting rare classes (class A, B, C, D, and E). Random over-sampling yields a better result in terms of sensitivity and precision, whereas the SMOTE technique gives the best sensitivity performance. We also consider the ROC (receiver operating characteristic) area under curve of each technique. The ROC area is a measurement to compare a tradeoff between true positive and false positive error rates. The desired ROC area is over 0.5, and the higher is the better. The ROC area comparison of the two over-sampling techniques against the imbalanced data, specifically for the five most rare classes, is illustrated in Figure 8.

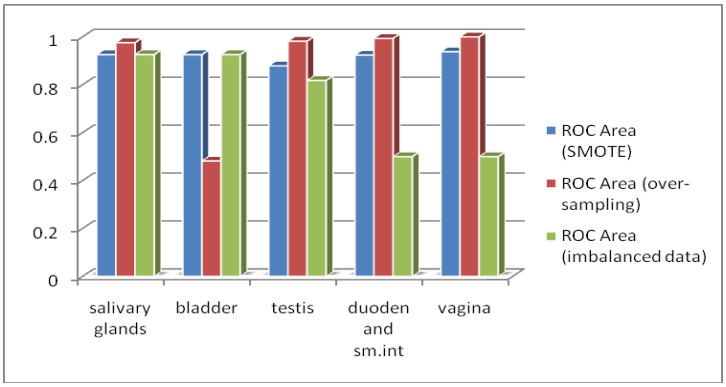


Figure 8. ROC Area Comparison of the Five Rare Classes of Primary Tumor

4. Data replication and heuristic-based feature selection techniques

Although sampling methods are simple and yet efficient for mining rare objects, under-sampling may remove good representatives, while over-sampling may cause the over-fitting problem. In this section, we present the heuristic-based unsupervised feature selection technique to be applied to the training data prior to the application of over-sampling technique, which duplicates the number of rare class instances to the same proportion to the majority class instances. The steps of cluster-based feature selection technique are presented as pseudo-code in Figure 9. The flow of the data preparation steps including the model's predictive performance evaluation step is also graphically presented in Figure 10.

Data Cleaning Phase

(1) Investigate data quality. If the data in any feature appear to be a single value, then remove that feature.

(2) Consider missing values in each feature. If data are missing more than 55%, then remove that feature.

Feature Selection Phase

(3) Apply the following cluster-based feature selection technique:

(3.1) Clustering data into *k* clusters (*k* is usually a number of classes)

(3.2) Compare mean value differences in every feature

(3.3) Ranking features in descending order according to the magnitude of mean differences, and output the ranked features

Over-sampling Phase

(4) Separate data obtained from step 2 into two datasets: train data and test data. Each data set maintains the same proportion of cases.

(5) Increase the number of minority cases in the train data by duplicating cases to be the same amount as the majority cases.

Model Building Phase

(6) Build a prediction model with learning algorithms.

(7) Evaluate model accuracy with the test dataset.

Figure 9. Feature Selection Technique Based on Cluster Comparison

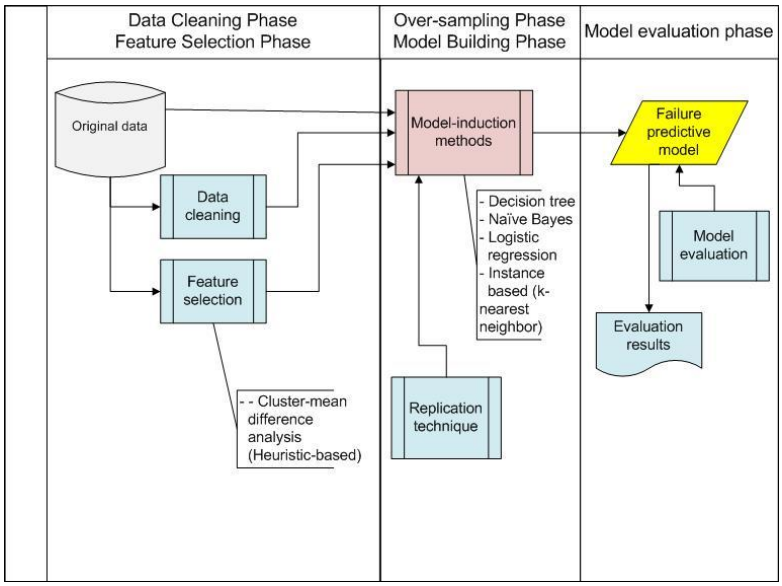


Figure 10. A Diagram of Data Preparation, Model Building, and Model Evaluation Steps

5. Experimental Results on Arrhythmia and Communities-and-crime Data

We then further our experimentation on arrhythmia and communities-and-crime data sets [6]. At this step, we also apply the cluster-based feature selection technique to avoid the over-fitting problem. The unsupervised feature selection technique has been applied to the training data prior to the application of over-sampling method. During the replication step, data instances in rare class are duplicated to the same proportion to the majority class, whereas the test data set still maintain the imbalance characteristic. We test the data replication techniques using both over-sampling and under-sampling methods. The results of predictive performances of the tree-based, k-nearest neighbor, logistic regression, and naïve Bayes models after testing with the hold-out arrhythmia data and the communities-and-crime data set are shown in Figures 11-18, respectively.

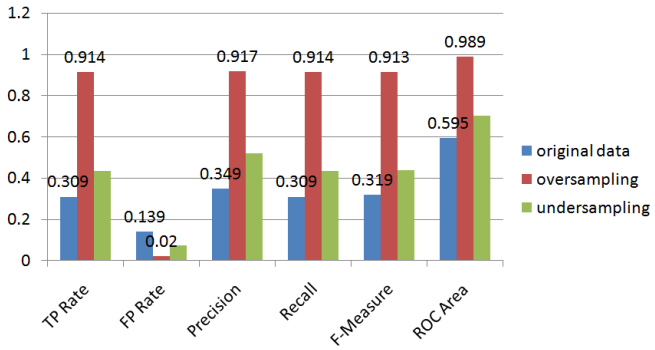


Figure 11. The *tree-based model* Predictive Performance Testing on *arrhythmia* Data that are Prepared with the Heuristic Cluster-based and Replication Techniques (showing results of both over-sampling and under-sampling methods**). The higher is the better in all metrics, except the FP rate.**

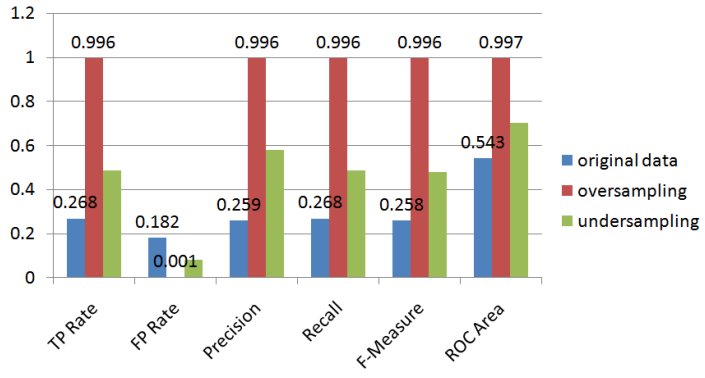


Figure 12. The *k-nearest neighbor* Model Predictive Performance Testing on *arrhythmia* Data that are Prepared with the Heuristic Cluster-based and Replication Techniques

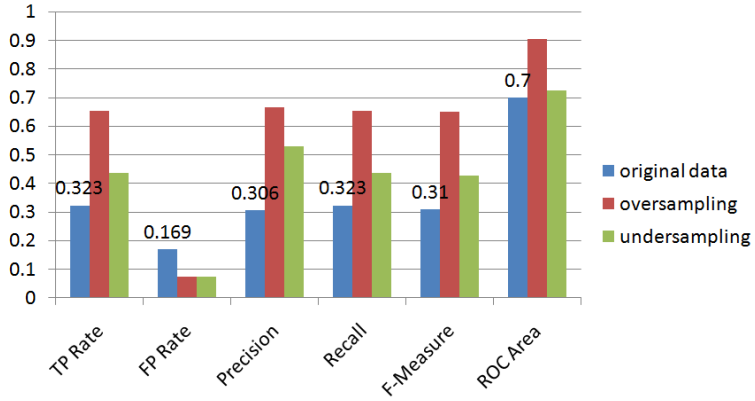


Figure 13. The *logistic regression* Model Predictive Performance Testing on *arrhythmia* Data that are Prepared with the Heuristic Cluster-based and Replication Techniques

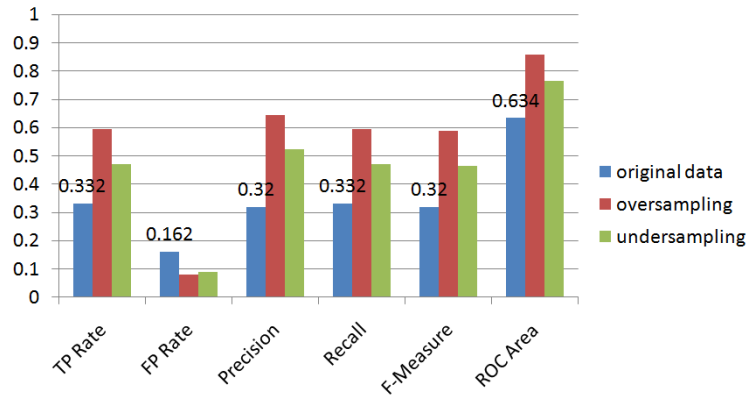


Figure 14. The *naïve Bayes* Model Predictive Performance Testing on *arrhythmia* Data that are Prepared with the Heuristic Cluster-based and Replication Techniques

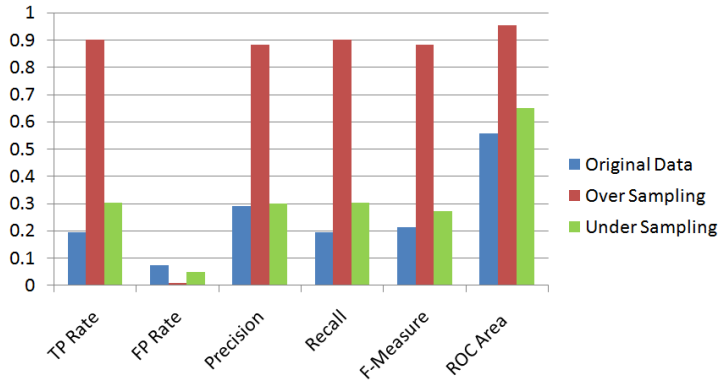


Figure 15. The *tree-based* Model Predictive Performance Testing on *communities-and-crime* Data

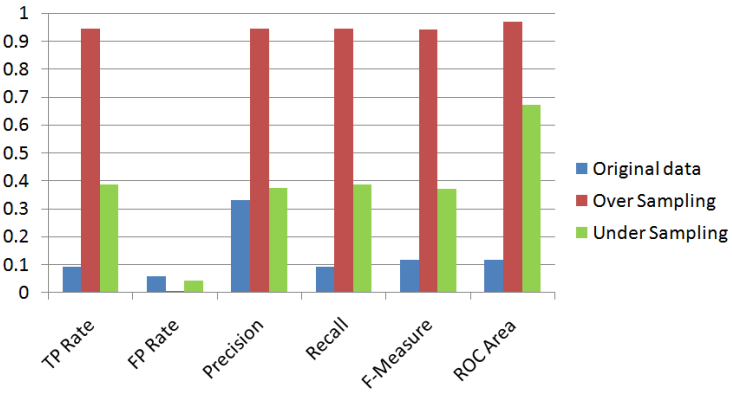


Figure 16. The *k-nearest neighbor* Model Predictive Performance Testing on *communities-and-crime* Data

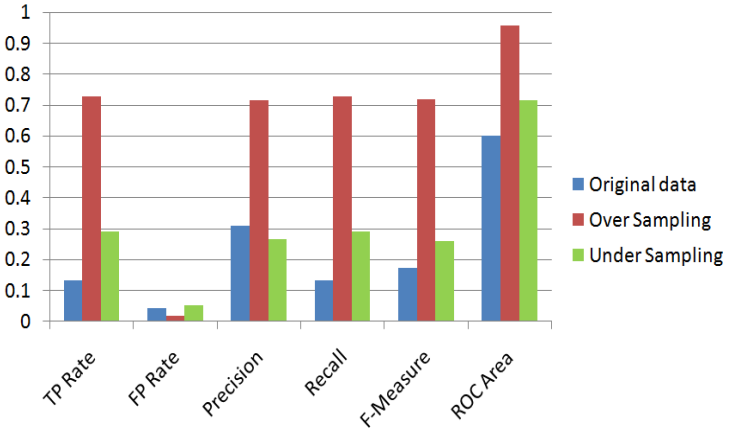


Figure 17. The *logistic regression* Model Predictive Performance Testing on *communities-and-crime* Data

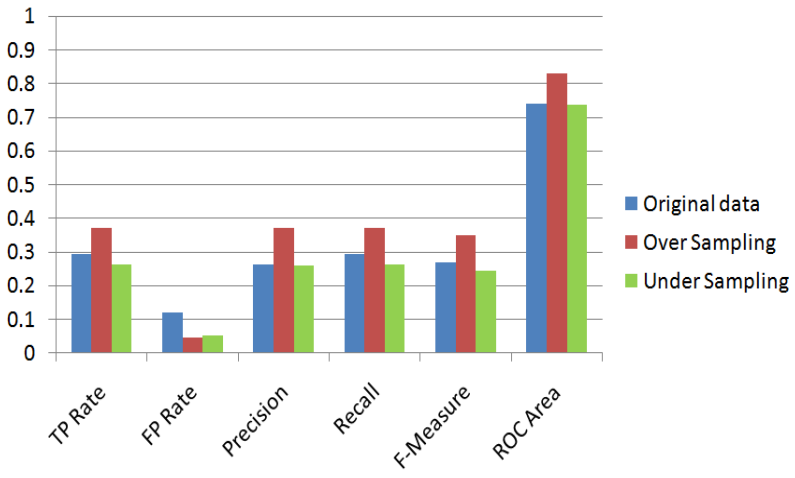


Figure 18. The *naïve Bayes* Model Predictive Performance Testing on *communities-and-crime* Data

6. Conclusion

The problem of predicting correctly rarely occurring cases is important to many real life applications such as network intruder identification, fraudulent credit card transaction detection, as well as vital clinical tests and medical diagnosis. Some data domains in the clinical and medical environment are difficult to build an accurate predictive model because of the inherent rare class situation. In the context of data mining, these rare classes refer to labeled data instances that are infrequently occurred in the database. Mining rarely occurred data is thus a challenging problem in several specific domains. We study the rare class mining problem in the context of life science and socio-economics in which rarely occurred events are of interest.

We firstly investigate the over-sampling techniques to bias a decision tree learning algorithm towards the minority classes. In this study, we comparatively investigate the random over-sampling and the synthetic minority over-sampling (SMOTE) techniques. The experimental results on predicting the primary tumors show a good predictive performance of both over-sampling techniques on predicting minority classes. Random over-sampling performs slightly better than SMOTE on recalling (or sensitivity test) the minority cases. But in terms of sensitivity (or the true negative rate), SMOTE shows the best performance. From this preliminary results, we further extend our study on other kinds of medical data domains and also include the feature selection in our research agenda.

From the promising preliminary results, we then decide to use an over-sampling technique to alleviate the outnumber situation of majority class. Such sampling technique is however prone to introducing the over-fitting problem. We thus propose the remedy by applying the cluster based technique to selectively extract data instances showing discrimination characteristics. The built models from various mining algorithms including the Bayesian learning, nearest neighbors, and logistic regression have been tested with a separate data set and the results show significant improvement on the predicting accuracy. The results confirm our hypothesis that the selective techniques to pick representative data with minimal set of features can improve the

predictive accuracy of the induced model. We plan to move our research direction towards the devise of a learning algorithm suitable for medical and biological domains.

Acknowledgement

This work has been supported by grant from the SUT Research and Development Fund. Series of experimentation on arrhythmia and communities-and-crime data sets had been done by Data Engineering's research assistants including Zagon Budsabong, Fonthip Koonggaew, and Phaichayon Kongchai.

References

[1] Breiman L, Freidman J, Olshen R, Stone C, "Classification and Regression Trees", Wadsworth, **(1984)**.

[2] Burez J, Van den Poel D, "Handling class imbalance in customer churn prediction", Expert Systems with Applications, vol. 36, **(2009)**, pp. 4626--4636.

[3] Chawla N, "Data mining for imbalanced datasets: an overview", In: O. Maimon and L. Rokach, (eds.) Data Mining and Knowledge Discovery Handbook, Springer, **(2005)**, pp. 853-867.

[4] Chawla N, Bowyer K, Hall L, Kegelmeyer W, "SMOTE: Synthetic Minority Over-sampling Technique", J of Artificial Intelligence Research, vol. 16, **(2002)**, pp. 341-378.

[5] Debnath R, Takahide N, Takahashi H, "A decision based one-against-one method for multi-class support vector machine", Pattern Analysis & Applications, vol. 7, no. 2, **(2004)**, pp. 164-175.

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

[7] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH, "The WEKA data mining software: an update", SIGKDD Explorations, vol. 11, no. 1, **(2009)**, pp. 10-18.

[8] Han S, Yuan B, Liu W, "Rare class mining: progress and prospect", In: Proc Chinese Conference on Pattern Recognition, **(2009)**, pp.1-5.

[9] Kretschmann E, Fleischmann W, Apweiler R, "Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT", Bioinformatics, vol. 17, no. 10, **(2001)**, pp. 920-926.

[10] Lalkhen AG, McCluskey A, "Clinical tests: sensitivity and specificity", Continuing Education in Anaesthesia, Critical Care & Pain, vol. 8, no. 6, **(2008)**, pp. 221-223.

[11] Mugambi EM, Hunter A, Oatley G, Kennedy L, "Polynomial-fuzzy decision tree structures for classifying medical data", Knowledge-Based Systems, vol. 17, no. 2-4, **(2004)**, pp. 81-87.

[12] Pandey B, Mishra RB, "Knowledge and intelligent computing system in medicine", Computers in Biology and Medicine, vol. 39, **(2009)**, pp. 215-230.

[13] Quinlan JR, "Induction of decision tree", Machine Learning, vol. 1, **(1986)**, pp. 81-106.

[14] Rifkin R, Klautau A, "In defense of one-vs-all classification", J of Machine Learning Research, vol. 5, **(2004)**, pp. 101-141.

[15] Stefanowski J, Wilk S, "Selective pre-processing of imbalanced data for improving classification performance", In: Proc DaWaK **(2008)**, pp. 283-292.

[16] Tapia E, Ornella L, Bulacio P, Angelone L, "Multiclass classification of microarray data samples with a reduced number of genes", BMC Bioinformatics, vol. 12, no. 59, **(2011)**.

[17] Thabtah FA, Cowling P, Peng Y, "Multiple labels associative classification", Knowledge and Information Systems, vol. 9, no. 1, **(2006)**, pp. 109-129.

[18] Van Hulse J, Khoshgoftaar T, "Knowledge discovery from imbalanced and noisy data", Data & Knowledge Engineering, vol. 68, **(2009)**, pp. 1513-1542.

[19] Webster's New World™ Medical Dictionary, 3rd edition. Wiley Publishing, **(2008)**.

[20] Weiss GM, "Mining with rarity: a unifying framework", SIGKDD Explorations, vol. 6, no. 1, **(2004)**, pp. 7-9.

[21] Yeung KY, Bumgarner RE, "Multiclass classification of microarray data with repeated measurements: application to cancer", Genome Biology, vol. 4, no. 12, R83 **(2004)**.

Authors

Nittaya Kerdprasop is an associate professor at the school of computer engineering, Suranaree University of Technology, Thailand. She received her B.S. from Mahidol University, Thailand, in 1985, M.S. in computer science from the Prince of Songkla University, Thailand, in 1991 and Ph.D. in computer science from Nova Southeastern University, USA, in 1999. She is a member of ACM and IEEE Computer Society. Her research of interest includes Knowledge Discovery in Databases, Artificial Intelligence, Logic Programming, Deductive and Active Databases.

Kittisak Kerdprasop is an associate professor at the school of computer engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in Mathematics from Srinakarinwirot University, Thailand, in 1986, master degree in computer science from the Prince of Songkla University, Thailand, in 1991 and doctoral degree in computer science from Nova Southeastern University, USA., in 1999. His current research includes Data mining, Artificial Intelligence, Functional and Logic Programming Languages, Computational Statistics.