Improving Rare Case Prediction with Replication Technique

Nittaya Kerdprasop, Fonthip Koongaew, Zagon Budsabong, Phaichayon Kongchai, and Kittisak Kerdprasop

**Data Engineering Research Unit, School of Computer Engineering,
Suranaree University of Technology, 111 University Avenue,
Nakhon Ratchasima 30000 Thailand**

Abstract

The ability to predict correctly rarely occurring cases is important to the success of applying data mining method to many real life applications. In the context of data mining, rare cases refer to labeled data instances that are infrequently occurred in the database. Discovering infrequent patterns are of interest in some specific domains such as genetic mutant identification, fraud credit card detection, network intruder prevention. But most learning algorithms are biased toward the majority cases such that the minority cases are considered as noise and thus they are ignored during the model induction steps. This ignorance causes the learning algorithm to generate a model that cannot classify or predict a minority case. We thus study the replication technique based on the over-sampling method to solve this problem. However, a straightforward application of over-sampling method may lead to the over-fitting problem in such a way that the generated model is too specific to the manipulated data. We thus apply the cluster-based technique to selectively filter a training dataset. The experimental results on primary tumor, arrhythmia and communities-and-crime datasets show significant improvement on predicting accuracy, specificity, and sensitivity of the induced models. But the results on multiple features correlation dataset show non-significant improvement; this case requires further investigation.

Keywords: *Rare Case Prediction, Classification Model, Sample Replication, Data Mining, Over-sampling Technique.*

1. Introduction

The discovery of hidden patterns from large databases can uncover knowledge to support the nontrivial task of decision making. Researchers and practitioners in several areas have successfully applied data mining technology to obtain descriptive patterns and predictive models from their database contents. However, data mining application in some specific areas such as biomedical [12], [14], [15], [21], [22], [27] and clinical professions is still in a limited scope due to a severe problem of low predictive accuracy of the model induced from the data samples. Low accuracy of the induced model is due to the multi-features and imbalanced characteristics inherent in some datasets.

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 [1], [17], [23], [24]. 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 [2], [3], [5], [9], [10], [11], [16], [18], [19], [20]. We refer to this problem as rare case 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 [4], [8], [26]. The under-sampling technique may remove good representatives, while over-sampling may cause the over-fitting problem.

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. Model Accuracy Measurement

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 (true positive), FN (false negative), FP (false positive), and TN (true negative). Each measurement can be defined as follow [13]:

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

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

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

FN = the number of positive cases that are misclassified as negative cases.

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 [11] of these metrics are as follows:

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

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

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

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

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

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

Fig. 1 A confusion matrix of the binary classification.

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:

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

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

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

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

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

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

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

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

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

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

Fig. 2 A confusion matrix of the three-class classification.

3. Cluster-Based Feature Selection

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 paper, we propose the unsupervised feature selection technique to be applied to the training data prior to the application of over-sampling technique duplicating the rare class instances to the same proportion to the majority class. The cluster-based feature selection technique is presented in Figure 3.

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.

Fig. 3 Feature selection technique based on cluster comparison.

4. Experimentation and Results

We performed experiment on four datasets: primary tumor, arrhythmia, multiple features correlation, and communities-and-crime. The first three datasets are standard benchmark data publicly available at the UCI repository [6]; the last dataset can be downloaded from http://mlr.cs.umass.edu/ml/datasets.html. Experimentation with data mining methods has been performed with the WEKA software [7].

4.1 Results From Primary Tumor Dataset

A primary tumor is a tumor that has been developed at the original site where it first generated [25]. Among the 21 kinds of primary tumors, 6 of them are rarely occurred. The rare cases of primary tumors are duodenum and small intestine, salivary glands, bladder, testis, cervix uteri, and vagina. They are observed at a 0.5% frequency rate, whereas the most frequent one (which is lung cancer) occurs at a high rate of 24.8%. Class distribution of the primary tumors (that is, lung – 84 cases, head and neck – 20 cases, esophagus – 9 cases, thyroid – 14 cases, stomach – 39 cases, duodenum and small intestine – 1 case, colon – 14 cases, rectum – 6 cases, salivary glands – 2 cases, pancreas – 28 cases, gall bladder – 16 cases, liver – 7 cases, kidney – 24 cases, bladder – 2 cases, testis – 1 case, prostate – 10 cases, ovary – 29 cases, corpus uteri – 6 cases, cervix uteri – 2 cases, vagina 1 case, and breast – 24 cases) is given in Figure 4.

Our preliminary hypothesis is that by biasing the class distribution of the minority data with the replication methods (random over-sampling and synthetic minority over-sampling – SMOTE [3]), the learning algorithm may perform better on recognizing the rare cases. The first step of our experimentation is to duplicate a data record containing only a single case (that is, the case of duodenum and small intestine, testis, and vagina tumors) to contain two records of each class of tumor.

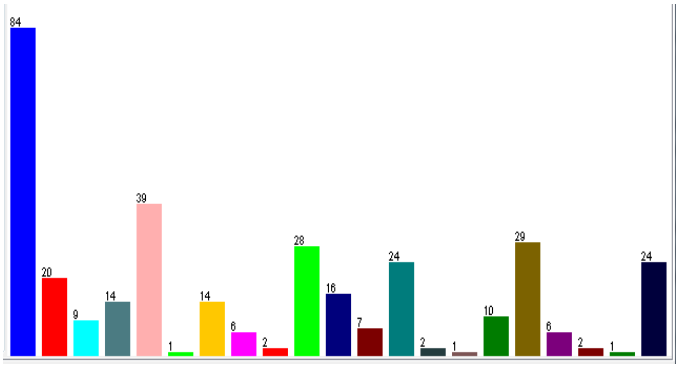


Fig. 4 Class distribution of primary tumor dataset.

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 [3]. The third version of train data is for the random over-sampling.

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 outcomes on precision and recall (as shown in Figure 5) confirms our hypothesis. The symbolic codes for different primary tumor types are as follows:

A = salivary glands,
B = bladder,
C = testis,
D = duodenum 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.

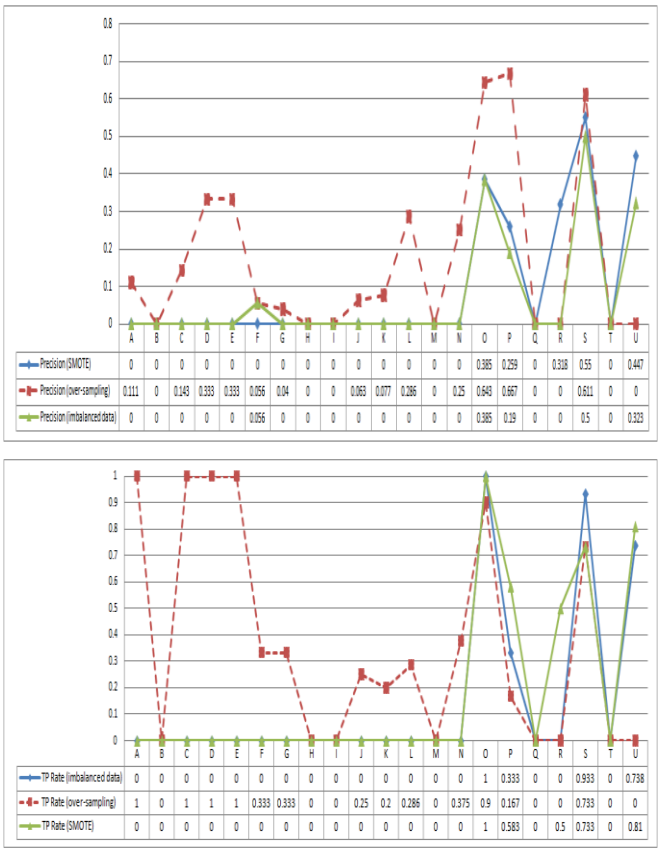


Fig. 5 Precision (above) and recall (below) of the tree-based model primary tumor prediction (the dashed line is the performance of random over-sampling method).

The specificity and F-measure values of the different data preparation methods are shown in Figures 6 and 7, respectively. The improved predictive performance is also shown (in Figure 8) as the ROC (receiver operating characteristic) area under curve, which 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.

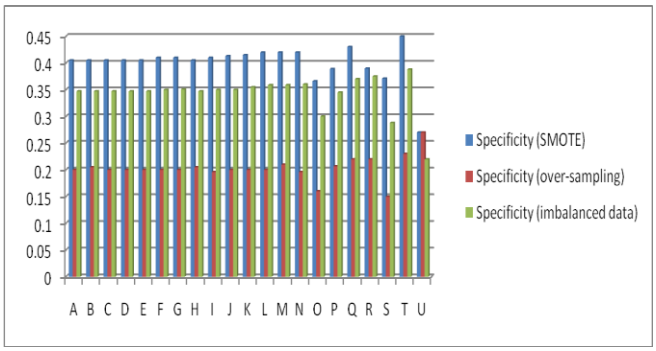


Fig. 6 Specificity comparison.

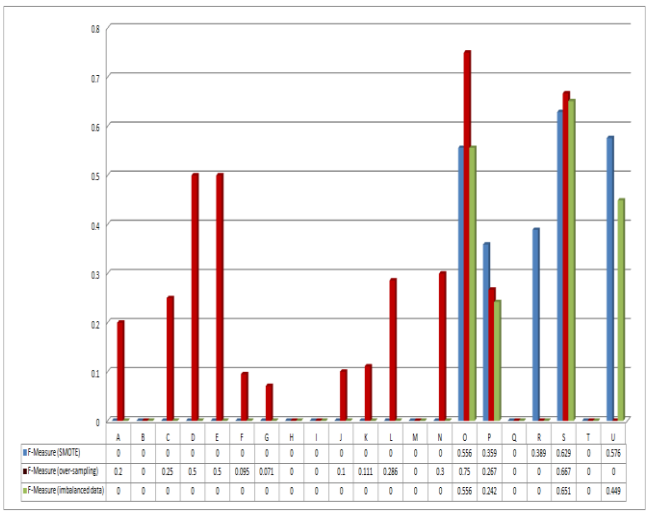


Fig. 7 F-measure comparison.

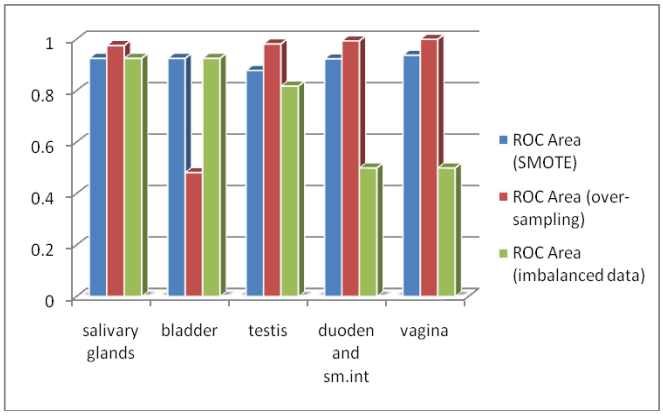
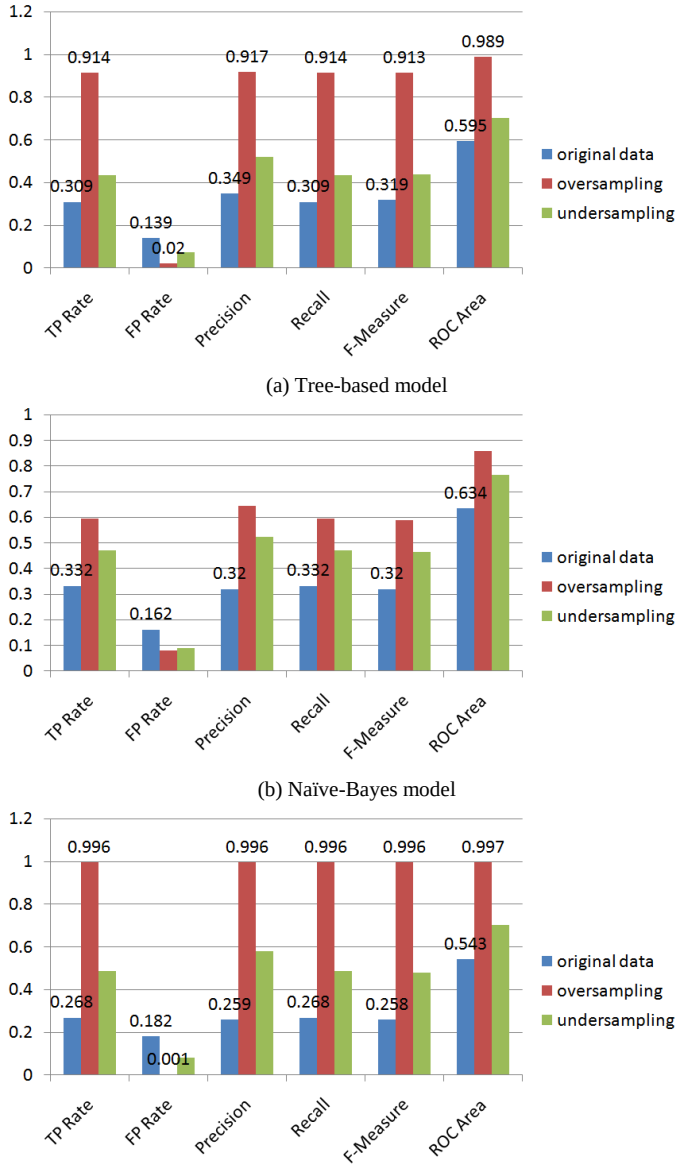


Fig. 8 ROC area comparison of the five rare cases.

4.2 Results From Arrhythmia Dataset

The first experiment that compared sampling techniques has confirmed the potential of over-sampling. We then further our experimentation on arrhythmia dataset. 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 dataset still maintain the imbalance characteristic.

The results of predictive performances of the tree-based, k-nearest neighbor, and naïve Bayes models after testing with the hold-out dataset is shown in Figure 9.



(a) Tree-based model

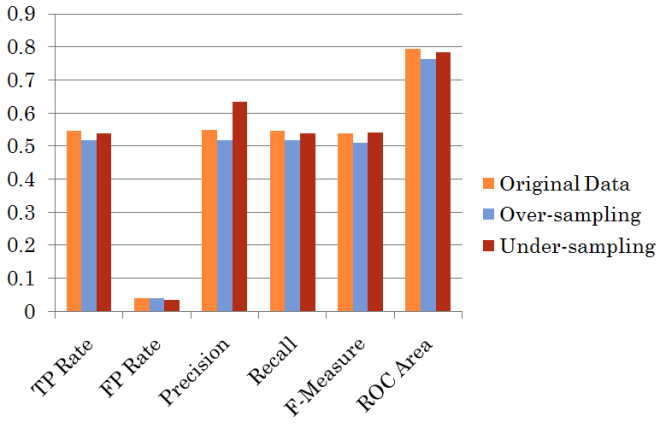
(b) Naïve-Bayes model

(c) k-nearest neighbor model

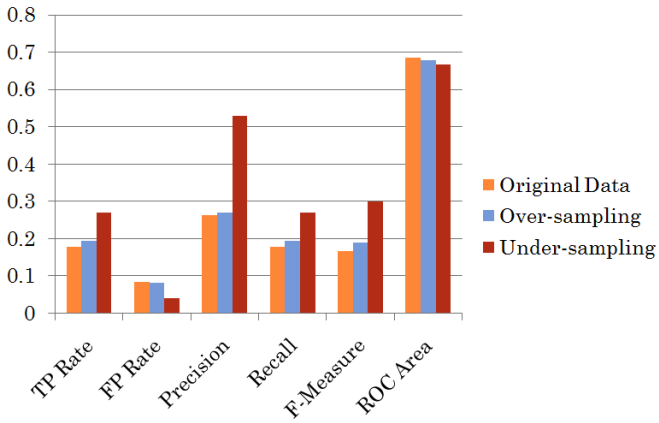
Fig. 9 Predictive performance testing on arrhythmia dataset.

4.3 Results From Multiple Features Correlation Dataset

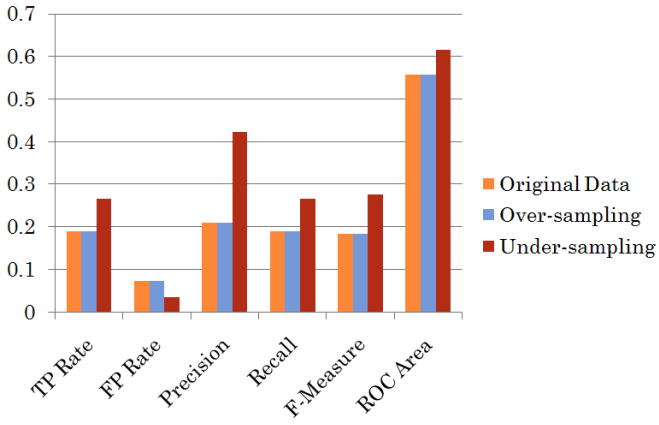
This dataset is digitized binary images consisting of 2000 patterns of handwritten numerals. The dataset used in this experimentation is profile correlation containing 216 attributes. The results of predictive performances of the tree-based, k-nearest neighbor, and naïve Bayes models after testing with the hold-out dataset is shown in Figure 10.



(a) Tree-based model



(b) Naïve-Bayes model



(c) k-nearest neighbor model

Fig. 10 Predictive performance testing on multiple features correlation dataset.

4.4 Results From Communities-and-Crime Dataset

This dataset is part of socio-economic data from the 1990 US census, law enforcement data survey, and crime data from the 1995 FBI report. The target of prediction is the per capita violent crimes. Violent crimes include murder, rape, robbery, and assault. Non-violent crimes are burglaries, larcenies, car thefts, and arsons.

The per capita violent crime is a real value calculated from population (per capita is per 100,0000 population) and the sum of violent crimes. Original dataset is all real values; we have to discretize the target variable into 15 intervals (as shown in Figure 11) so as the classification data mining methods can be performed. The results of predictive performance evaluation of this dataset can be graphically shown in Figure 12.

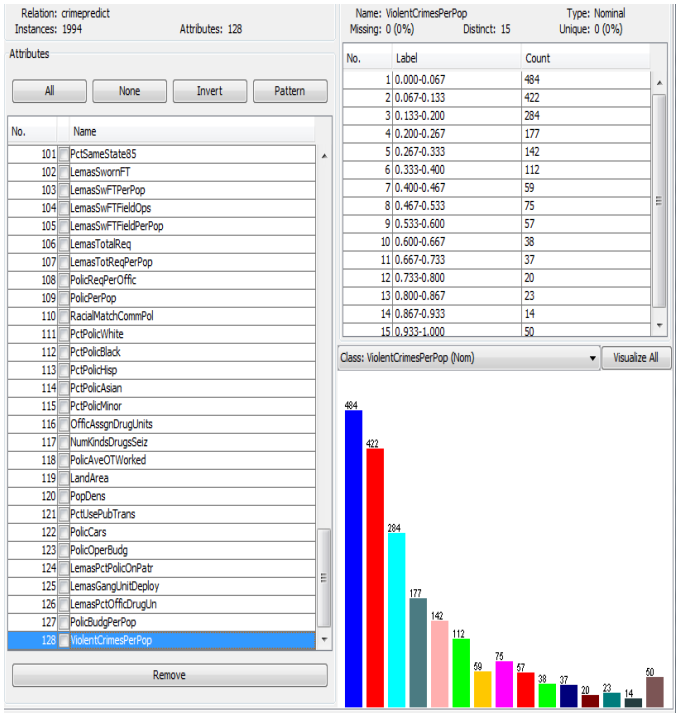
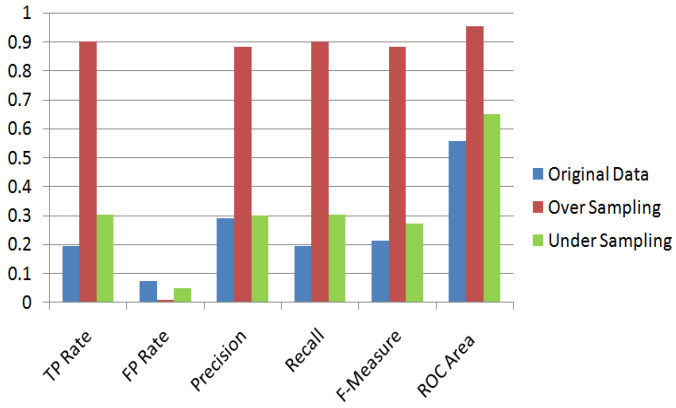
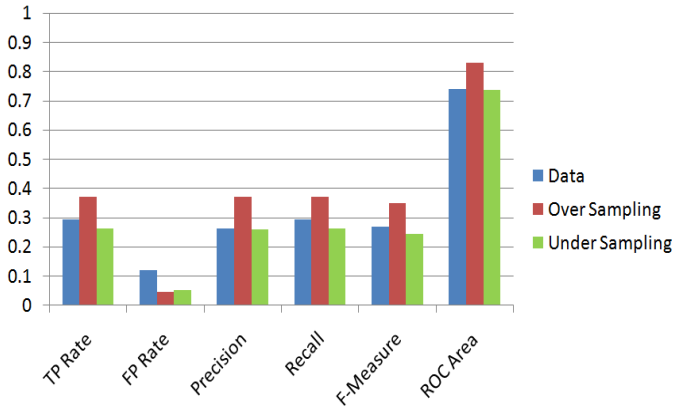


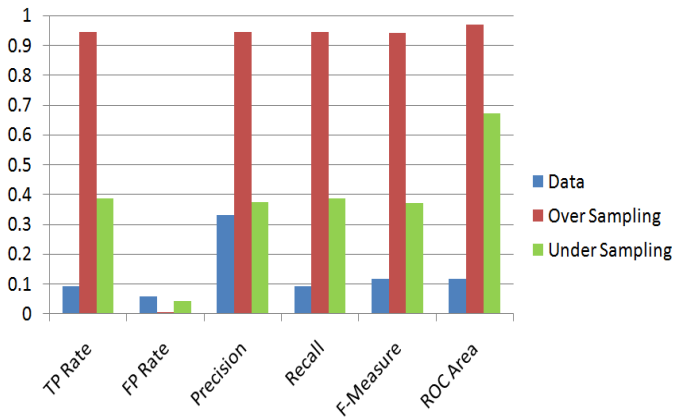
Fig. 11 Discretization of a target variable and the class distribution of communities-and-crime dataset.



(a) Tree-based model



(b) Naïve-Bayes model



(c) k-nearest neighbor model

Fig. 12 Predictive performance testing on communities-and-crime dataset.

5. Conclusion

Rare case prediction is the data mining task aiming at building a model that can correctly identify objects or events rarely occurring in the data set. In many real life applications such as identification of intruders accessing a network system, detecting fraudulent credit card transactions, it is rare events that are of great interest. Unfortunately, traditional mining algorithms fail to predict rare events because the model are inherently built in favor of the majority class to draw common characteristics among data instances. Mining rarely occurred data is thus a challenging problem in some specific domains. We study the rare case mining problem in the context of life science and socio-economics in which rarely occurred events are of interest.

In this paper, we propose 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 have been tested with a separate data set and the results show significant improvement on the predicting accuracy.

Acknowledgment

This research was supported by the SUT Research and Development Fund, Suranaree University of Technology.

References

[1] L. Breiman, J. Freidman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Belmont, California: Wadsworth, 1984.

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

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

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

[5] R. Debnath, N. Takahide, and H. Takahashi, "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] A. Frank and A. Asuncion, *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ ml], Irvine, University

of California, School of Information and Computer Science, 2010.
[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations*, vol. 11, no. 1, 2009, pp. 10-18.
[8] S. Han, B. Yuan, and W. Liu, "Rare class mining: progress and prospect," *Proceedings of Chinese Conference on Pattern Recognition*, 2009, pp.1-5.
[9] I. Jamali, M. Bazmara, and S. Jafari, "Feature selection in imbalance data sets," *International Journal of Computer Science Issues*, vol. 9, no. 2, 2012, pp. 42-45.
[10] P. Jhonpita, S. Sinthupinyo, and T. Chaiyawat, "Ordinal classification method for the evaluation of Thai non-life insurance companies," *International Journal of Computer Science Issues*, vol. 9, no. 2, 2012, pp. 362-366.
[11] K. Kerdprasop and N. Kerdprasop, "A data mining approach to automate fault detection model development in the semiconductor manufacturing process," *International Journal of Mechanics*, vol. 5, issue 4, 2011, pp. 336-344.
[12] E. Kretschmann, W. Fleischmann, and R. Apweiler, "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.
[13] A.G. Lalkhen and A. McCluskey, "Clinical tests: sensitivity and specificity. Continuing Education in Anaesthesia," *Critical Care & Pain*, vol. 8, no. 6, 2008, pp. 221-223.
[14] E.M. Mugambi, A. Hunter, G. Oatley, and L. Kennedy, "Polynomial-fuzzy decision tree structures for classifying medical data," *Knowledge-Based Systems*, vol. 17, no. 2-4, 2004, pp. 81-87.
[15] B. Pandey and R.B. Mishra, "Knowledge and intelligent computing system in medicine," *Computers in Biology and Medicine*, vol. 39, 2009, pp. 215-230.
[16] R. Pant, T.B. Trafalis, and K. Barker, "Support vector machine classification of uncertain and imbalanced data using robust optimization," *Recent Researches in Computer Science – Proceedings of the 15th WSEAS International Conference on Computers*, 2011, pp. 369-374.
[17] J.R. Quinlan, "Induction of decision tree," *Machine Learning*, vol. 1, 1986, pp. 81-106.
[18] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *Journal of Machine Learning Research*, vol. 5, 2004, pp. 101-141.
[19] J. Stefanowski and S. Wilk, "Selective pre-processing of imbalanced data for improving classification performance," *Proceedings of DaWaK*, 2008, pp. 283-292.
[20] H. Sug, "Improving the performance of minor class in decision tree using duplicating instances," *Recent Researches in Artificial Intelligence, Knowledge Engineering and Data Bases – 10th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Databases*, 2011, pp. 234-237.
[21] E. Tapia, L. Ornella, P. Bulacio, and L. Angelone, "Multiclass classification of microarray data samples with a reduced number of genes," *BMC Bioinformatics*, vol. 12, 2011, article 59.
[22] F.A. Thabtah, P. Cowling, and Y. Peng, "Multiple labels associative classification," *Knowledge and Information Systems*, vol. 9, no. 1, 2006, pp. 109-129.
[23] C.-J. Tsai, C.-I. Lee, C.-T. Chen, and W.-P. Yang, "A multivariate decision tree algorithm to mine imbalanced data," *WSEAS Transactions on Information Science and Applications*, vol. 4, issue 1, 2007, pp. 50-58.
[24] J. Van Hulse and T. Khoshgoftaar, "Knowledge discovery from imbalanced and noisy data," *Data & Knowledge Engineering*, vol. 68, 2009, pp. 1513-1542.
[25] *Webster's New World™ Medical Dictionary*, 3rd edition, Wiley Publishing, 2008.
[26] G.M. Weiss, "Mining with rarity: a unifying framework," *SIGKDD Explorations*, vol. 6, no. 1, 2004, pp. 7-9.
[27] K.Y. Yeung and R.E. Bumgarner, "Multiclass classification of microarray data with repeated measurements: application to cancer," *Genome Biology*, vol. 4, no. 12, 2004, R83.

Nittaya Kerdprasop is an associate professor with 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, U.S.A., 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.

Fonthip Koongaew was a computer engineer with the data engineering research unit. She received her bachelor and master degrees in computer engineering from the school of computer engineering, Suranaree University of Technology in 2010 and 2012, respectively. Her research interest is data mining, constraint logic programming, and decision tree induction.

Zagon Budsabong is a master student with the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in computer science from Rajamangala University of Technology Tawan-Ok, Chakrabongse Bhuvanarth Campus, Thailand, in 2010. His research interest is software engineering and intelligent data analysis.

Phaichayon Kongchai is currently a doctoral student with the School of Computer Engineering, Suranaree University of Technology, Thailand. He received his bachelor degree in computer engineering from Suranaree University of Technology, Thailand, in 2010, and master degree in computer engineering in the same institution in 2012. His research topic is related to data mining, constraint logic programming, and artificial intelligence.

Kittisak Kerdprasop is an associate professor and chair of computer engineering school, 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, U.S.A., in 1999. His current research includes Data mining, Artificial Intelligence, Functional and Logic Programming, and Computational Statistics.