Expert Systems with Applications 38 (2011) 4222–4227

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Determining the optimal re-sampling strategy for a classification model with imbalanced data using design of experiments and response surface methodologies

Lee-Ing Tong, Yung-Chia Chang, Shan-Hui Lin *

Department of Industrial Engineering and Management, Chiao Tung University, 1001 Ta Hsueh Road, Hsinchu 300, Taiwan

ARTICLE INFO

Keywords:
Re-sampling strategy
Imbalanced data
Classifier
Machine learning
Design of experiments
Response surface methodologies
The area under ROC curve

ABSTRACT

Imbalanced data are common in many machine learning applications. In an imbalanced data set, the number of instances in at least one class is significantly higher or lower than that in other classes. Consequently, when classification models with imbalanced data are developed, most classifiers are subjected to an unequal number of instances in each class, thus failing to construct an effective model. Balancing sample sizes for various classes using a re-sampling strategy is a conventional means of enhancing the effectiveness of a classification model for imbalanced data. Despite numerous attempts to determine the appropriate re-sampling proportion in each class by using a trial-and-error method in order to construct a classification model with imbalanced data (Barandela, Vadovinos, Sánchez, & Ferri, 2004; He, Han, & Wang, 2005; Japkowicz, 2000; McCarthy, Zabar, & Weiss, 2005), the optimal strategy for each class may be infeasible when using such a method. Therefore, this work proposes a novel analytical procedure to determine the optimal re-sampling strategy based on design of experiments (DOE) and response surface methodologies (RSM). The proposed procedure, S-RSM, can be utilized by any classifier. Also, C4.5 algorithm is adopted for illustration. The classification results are evaluated by using the area under the receiver operating characteristic curve (AUC) as a performance measure. Among the several desirable features of the AUC index include independence of the decision threshold and invariance to *a priori* class probabilities. Furthermore, five real world data sets demonstrate that the higher AUC score of the classification model based on the training data obtained from the S-RSM is than that obtained using oversampling approach or undersampling approach.

1. Introduction

In supervised learning, learning systems construct a classification model for an output variable with several categories based on simulated relations between input and output variables. The output variable is also called the class variable. For instance, credit scoring data can be classified as either good or bad credit. In real world applications, data sets with several classes are generally imbalanced, i.e. each class differs in the number of instances and, occasionally, differs significantly. For instance, in fiduciary loan data, the number of good credit customers significantly exceeds that of bad ones.

For imbalanced data of two categories, the category with more data is called the majority class; the minority class refers to the category with less data. Imbalanced data are common owing to erroneous decision or a rare subject, e.g., a valuable abalone in comparison with a common one. Bankrupt prediction data and credit scoring data are normally imbalanced data. As a dichotomous decision, bankrupt prediction forecasts whether enterprises or individuals are bankrupt. For instance, banks can determine whether an enterprise or an individual has good or bad credit based on a credit scoring model. Results from a credit scoring model can also facilitate banks to decide whether or not to grant loans to a corporation or an individual. In this case, corporations or individuals with bad credit are normally significantly lower than those with good credit. However, bad credit cases may incur substantial revenue losses for a bank.

When a classification model is developed based on instances from imbalanced data, most classifiers are subjected to *a priori* class probability and thus fail to construct an appropriate model (Japkowicz & Stephen, 2002). The *priori* class probability refers to a probability in which an instance belongs to a certain class under general circumstances. As long as data are accumulated properly, the probability that the accumulated data belong to a certain class can be treated as *a priori* class probability. Consider a credit scoring data set, in which bad credit data comprises 1/10 of the entire data set; in addition, the *priori* class probability of bad credit and good credit are 0.1 and 0.9, respectively. Classifiers such as artificial neuron network (ANN) or decision tree (DT) construct models are based on properties from instances to identify classes. Most

* Corresponding author.
E-mail addresses: teapenny.iem94g@nctu.edu.tw, teapenny300@gmail.com (S.-H. Lin).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.087

classifiers prioritize the overall accuracy during model construction. Once training data are extremely imbalanced, most classifiers neglect the properties provided by the minority class and attempt to classify all instances belonging to the majority class. Some classifiers, including the C4.5 decision tree, cannot construct an appropriate classification model with extremely imbalanced data. Under this circumstance, although the overall accuracy of an inappropriate classification model may, for instance, exceed 95% and the accuracy rate of the majority class may, for instance, also exceed 95%, the accuracy rate of the minority class may fall below 30%. In practice, misjudging the minority class generally incurs a significantly higher cost than that of a majority class (Phua, Alahakoon, & Lee, 2004). Thus, the inability of a classifier to identify instances of the minority class accurately would normally incur considerable loss.

Most classifiers in machine learning assess their classification results based on the overall accuracy while assuming equal prior class probabilities. However, the prior class probabilities are unequal in the imbalanced data. Moreover, the performance of classification results is evaluated inaccurately when using the overall accuracy as a performance measure. Rather than using the overall accuracy as a performance measure for a classifier, some studies recommend other measures to mitigate the influence caused by prior class probabilities, such as true positive rate (TPR), true negative rate (TNR), and the geometric mean of TPR and TNR (Barandela, Sánchez, García, & Rangel, 2003; Barandela et al., 2004), the relative misclassification cost (Japkowicz & Stephen, 2002; McCarthy et al., 2005), the receiver operating characteristic (ROC), and the area under the ROC curve (AUC) (Bradley, 1997; Provost, Fawcett, & Kohavi, 1998). Among these measures, AUC is deemed a more effective means in evaluating the classification results than other measures since AUC is not subjected to the prior class probabilities and different decision thresholds in classifiers; in addition, it can be expressed as a single number (Bradley, 1997). Moreover, when the relative misclassification cost is known, the relative misclassification cost can also be used as an assessment measure. This is despite the fact that, in practice, the relative misclassification cost is normally unknown.

Two conventional approaches to constructing a classification model for imbalanced data are (a) to modify current classification algorithms, such as utilizing meta-learning with different algorithms to enhance the ability of the classifier (Phua et al., 2004) and (b) to balance the sample sizes for different classes based on the re-sampling strategy (Provost, 2000). Undersampling and oversampling are two commonly adopted re-sampling methods. When an undersampling approach is adopted, few instances are drawn from the majority class as the training data. For the oversampling approach, instances are duplicated one or more times the amount of the original data in the minority class. The two re-sampling approaches have their merits and limitations. An undersampling approach produces less training data than the original data, possibly increasing the calculation efficiency. However, an undersampling approach discards information involved in unselected instances, possibly reducing the classifying accuracy of a classification model. Similarly, an oversampling approach introduces more data into analysis, thus making it time-consuming. Moreover, an oversampling approach occasionally incurs over-fitting of the classification model (McCarthy et al., 2005). Many studies attempted to determine the appropriate re-sampling proportion in each class of an imbalanced data set based on a trial-and-error method (Barandela et al., 2004; He et al., 2005; Japkowicz, 2000 McCarthy et al., 2005). However, trial-and-error method may not include the appropriate re-sampling proportion in each class of an imbalanced data set. Constructing an effective classification model requires a systematic re-sampling approach to determine the optimal proportions of instances for the majority and minority classes, respectively, in imbalanced data.

Design of experiments (DOE) is extensively adopted in industry to determine the optimal settings of process parameters in order to attain the desired quality of a process/product. Response surface methodology (RSM) optimizes the process/product by constructing proper equations to correlate the input and output variables with each other. RSM has recently been applied to machine learning. Irani, Cheng, Fayyand, and Qian (1993) constructed a polynomial using RSM with only a limited amount of semiconductor manufacturing data and, then, derived an equation to simulate a variety of manufacturing circumstances. Subsequently, a large data set was generated to provide classifiers in order to construct models. Shin, Guo, Choi, and Kim (2007) developed a two-stage robust data mining (RDM) method that consists of two stages for a water treatment plant to determine the optimal settings of process parameters in order to minimize the conductivity of treated water. The first stage involved selecting four critical manufacturing variables, i.e. three controllable and an uncontrollable variable, based on feature selection. The second stage entailed applying RSM to determine the optimal setting of process parameters while considering of an uncontrollable variable.

Therefore, this work attempts to determine the optimal re-sampling strategy of a classifier for two-class imbalanced data using DOE and RSM. The proposed method, Sampling-RSM (S-RSM), can determine the proper re-sampling proportions for the majority class and the number of duplications of the instances in minority class for the training data for a classifier. Additionally, a classification model is developed for a two-class imbalanced data set based on re-sampled data obtained from S-RSM. Moreover, effectiveness of the proposed S-RSM model is evaluated based on the AUC score. The rest of this paper is organized as follows. Section 2 describes the research methodology. Section 3 then introduces the proposed S-RSM model. Next, Section 4 summarizes the experimental results obtained from using the proposed model to classify five real world data. Conclusions are finally drawn in Section 5.

2. Methodology

This section describes pertinent literature. Section 2.1 introduces the evaluation of classification results by using AUC. Sections 2.2 and 2.3 briefly introduce RSM and C4.5 decision tree, respectively.

2.1. AUC score

AUC refers to the area under ROC, which was originally used in signal detection theory and, more recently, in machine learning to determine an appropriate operating point or decision threshold. ROC provides trade-off information between TPR and FPR. In ROC, the x-axis denotes FPR, the y-axis represents TPR, and the area under ROC is AUC. AUC equals the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (Fawcett, 2006). AUC, with a value ranging between 0 and 1, is a single number evaluation of a classifier performance which can simplify the comparison of different classifiers or classification models.

The area under the ROC curve is often determined through trapezoidal integration, which uses the slope and intercept of the fitted ROC to obtain AUC (Bradley, 1997). However, given its use of straight lines to connect all the points in ROC, this approach tends to underestimate the actual AUC formed by smooth concave curves (Bradley, 1997; Hand & Till, 2001). Alternatively, the AUC score can be obtained by the formula which is the same as the Mann–Whitney U test statistic (Hand & Till, 2001; Hanley & McNeil, 1982).

A higher AUC score implies a better performance of the classification. A situation in which AUC equals 1 suggests the impossibil-

ity that the probability of a randomly chosen negative sample to be classified into a positive (p(+|−)) is higher than the probability of any randomly chosen positive sample to be classified into a positive sample (p(+|+)). When AUC is 0.5, the classification results provide no credible information because all samples are classified into the same class.

2.2. Response surface methodology

RSM comprises statistical and mathematical approaches that use DOE to explore how several explanatory variables and one or more response variables are related. RSM largely focuses on obtaining an optimal response based on a set of designed experiments. While RSM models polynomial functions for the functional relationship between a response and independent variables, a response surface visualizes the surface shape (Montgomery, 2005). Importantly, RSM can reduce the number of trials when considering many factors and interactions between factors. Moreover, the continuous search feature RSM is useful in determining how continuous factors and responses are related.

RSM provides design criteria to design proper experimental points for experimenters to select based on their specific requirements. Conventional response surface designs include central composite designs (CCD) (Montgomery, 2005) and Box–Behnken design (Box & Behnken, 1960). In addition to conventional response surface designs, experimenters can alternatively adopt computer-generated designs to design experiments under certain circumstances such as when experimenters can not select a standard design. A computer-generated design can occasionally prove more effective in the number of trails than that of a standard design. As a computer-generated design, D-optimal design may be the most frequently adopted optimality criterion (Jahani, Alizadeh, Pirozifard, & Qudsevali, 2008).

2.3. C4.5

Developed by Quinlan (1993) for inducing classification models, decision tree C4.5 algorithm has been extensively adopted in different fields owing to its ability to generate rules and operate fast. Let D denote a collected data set with N instances, $D = \{(\mathbf{x}_l, y_l)\}_{l=1}^{N}$, l represent the lth instance and $y_l \in \{0, 1\}$ be the corresponding binary variable. Moreover, let $p_0(p_1)$ be the proportion of class 0 (class 1), $p_0 = 1 - p_1$, in sample S. The entropy of S is calculated as

$$\text{Entropy}(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0) \tag{1}$$

Decision tree algorithm continuously splits nodes into subtrees to obtain a lower entropy until system cannot increase the gain, which is the reduced entropy after splitting and can be calculated by Eq. (2):

$$\text{Gain}(S, x_l) = \text{Entropy}(S) - \sum_{v=values(x_l)} \frac{|S_v|}{|S|} \text{Entropy}(S_v) \tag{2}$$

In Eq. (2), S_v denotes the subsample of S and the attribute x_l has one specific value. In different splitting points, decision tree prioritizes the largest gain after splitting. However, Eq. (2) prefers to split at the feature containing many values, e.g., the age attribute or the ID attribute, than fewer ones, e.g., gender. For rectifying this situation, C4.5 uses Gainratio which is defined as Eq. (3):

$$\text{Gainratio}(S, x_l) = \frac{\text{Gain}(S, x_l)}{\text{SplitInformation}(S, x_l)} \tag{3}$$

SplitInformation (S, x_l) is calculated by Eq. (4):

$$(S, x_l) = -\sum_{k \in values(x_l)} \frac{|S_k|}{|S|} \log_2 \frac{|S_k|}{|S|} \tag{4}$$

where S_k denotes a subsample of S and the attribute x_l has a specific value. SplitInformation is the entropy of S with respect to x_l .

3. S-RSM procedure

The proposed S-RSM procedure attempts to determine the optimal proportion of a two-class imbalanced data by using D-optimal design, DOE and RSM in order to develop an effective classification model.

The S-RSM procedure contains the following three steps:

Step 1: Design an experiment.

An experiment is designed to obtain an appropriate re-sampling strategy for the majority class and minority class in a two-class imbalanced data. While the number of instances drawn from the majority class and the number of the instances duplicated from the minority class are designed by using undersampling and oversampling, respectively. The experiment considers two factors. Factor A and factor B represent a/b and d/b, respectively, where a denotes the total number of the re-sampling instances in the majority class; b denotes the total number of instances in the minority class of the training data; and d denotes the number of instances duplicated in the minority class. Both factors are continuous, ranging from 0 to r, where r represents N_L/N_S, $r \geqslant 1$; N_L represents the total number of instances in the majority class; and N_S is the total number of instances in the minority class. This work adopts the D-optimal design with a cubic model. The D-optimal design is generated using Design-Expert 7.0.0 computer software, in which a 25-run design is generated, including five replications at the center. The experimental error is estimated using replications and the adequacy of a fitted model is confirmed. The response variable is the average AUC score of classification model for the majority class and minority class.

Step 2: Conduct the experiment.

(a) Randomly split the data into training data (D_1) and testing data (D_2) using 5-fold cross validation. Do (b) and (c) for each fold.
(b) Sample and duplicate D_1 based on each generated combination in step 1 to obtain a new data composition (D_3).
(c) Utilize C4.5 to construct a classification model using D_3; use the classification model to classify D_2 and access classification results by the AUC score.
(d) Calculate the average AUC score of the 5-fold cross validation and, then, use the averaged AUC score as the response variable.

Step 3: Fit a model and obtain the optimal re-sampling strategy.

The response surface model is obtained to demonstrate the relation between factor A, factor B, and the response variable, i.e. AUC score. The fitted model adequacies are confirmed by the lack-of-fit test, coefficient of determination (R^2) and the adjusted coefficient of determination R^2 (Adj-R^2). Obtain the optimal re-sampling strategy for the majority class and minority class.

4. Illustration

Effectiveness of the proposed method was demonstrated using five of the UCI data sets provided by Machine Learning Repository at the University of California, Irvine (Asuncion & Newman, 2007). The five data sets chosen are Abalone, Balance Scale, Letter Recognition, Mfeat-zer and Satimage. The S-RSM procedure focuses on the re-sampling strategy, which functions with any classifier. For demonstrative purposes, C4.5 decision tree is used as the classifier.

4.1. Data description

This work develops the optimal re-sampling strategy for two-class imbalanced data. Therefore, the original five UCI data sets are transformed to simulate two-class imbalanced data. Table 1 summarizes the contents of the transformed data. The transformed standard is based on Lin, Wu, and Zhou (2006). Column 2 lists the number of features in each data set; columns 3 and 4 list the criteria that form the majority class and minority class, respectively; and column 5 shows the ratio of number of instances in the majority class to that in the minority class.

4.2. S-RSM procedure using abalone data set

By using the abalone data set as an illustration, the effectiveness of the proposed S-RSM procedure is demonstrated as follows.

Step 1: Design an experiment.
The factors of interest range from 1 to 9.8 (r = 3786/ 391). A D-optimal design with 25 combinations is generated by using Design-Expert 7.0.0, as shown in Table 2.
Step 2: Perform the experiment.
Randomly split abalone data set into D_1 and D_2 based on 5-fold cross validation. Each D_1 contains 3342 instances (3029 majority instances and 313 minority instances) and each D_2 contains 835 instances. For every fold, (b) and (c) were implemented. Next, the average AUC scores of the 5-fold cross validation were calculated, as shown in the last column in Table 2.
Step 3: Fit the model and develop the optimal re-sampling strategy.
The relationship between factor A, factor B, and response-AUC score was demonstrated by fitting the following response surface model.

$$\widehat{\mathrm{AUC}} = 0.722837 - 0.04864A + 0.045787B + 0.046286AB - 0.02919A^2 - 0.0093B^2 + 0.0436A^2B - 0.0551AB^2 + 0.008141A^3 - 0.0502B^3$$

Table 3 summarizes the results of the lack-of-fit test of the fitted models with R^2 and Adj-R^2. Owing to the insignificance of the cubic term for the lack-of-fit test, the cubic model is appropriate (p-value = 0.5949). In Table 3, boldface represents the results of the selected cubic model. The values of R^2 and Adj- R^2 for the cubic model are 96.3% and 94.1%, respectively. By utilizing the response surface model, the optimal factor-level combination of factor A and factor B is determined as $(A, B) = (1.45, 1.00)$.

Fig. 1 shows the generation of the response surface and contour plot for the fitted model. According to this figure, factor A is more sensitive to AUC scores than factor B. While factor A increases, AUC increases slowly initially and then decreases dramatically.

Table 2
Experiments and results for abalone data.

Run	Factor A	Factor B	Response; AUC score
1	1.00	6.63	0.757
2	1.00	1.00	0.796
3	5.29	1.00	0.758
4	5.34	5.34	0.729
5	5.29	1.00	0.693
6	7.45	7.50	0.709
7	3.17	8.36	0.746
8	9.68	1.00	0.500
9	1.00	9.68	0.766
10	9.68	1.00	0.500
11	5.34	5.34	0.713
12	1.00	9.68	0.774
13	2.81	3.23	0.739
14	1.00	1.00	0.773
15	5.34	5.34	0.707
16	7.82	3.21	0.655
17	5.33	9.68	0.724
18	9.68	9.68	0.673
19	5.34	6.18	0.726
20	9.68	6.58	0.695
21	5.34	5.34	0.723
22	5.34	5.34	0.737
23	9.68	9.68	0.670
24	3.17	5.79	0.753
25	5.34	3.41	0.681

Notably, the number of sampling instances is (the level of factor) × (the total number of the minority class in D_1 = 313). When the levels combination is $(A, B) = (1.00, 6.63)$, the number of sampling instances is (313, 2075).

Table 3
Lack-of-fit tests of abalone data and model summary statistics.

Source	SS	DF	MS	F-Value	P-Value	R^2	Adj- R^2
Linear	0.037	13	0.003	8.681	0.0014	0.675	0.645
2FI	0.017	12	0.001	4.356	0.0171	0.836	0.813
Quadratic	0.011	10	0.001	3.439	0.0384	0.884	0.854
Cubic	**0.002**	**6**	**0.000**	**0.798**	**0.5949**	**0.963**	**0.941**
Pure Error	0.003	9	0.000				

Boldface represents the results of the selected model and 2FI represents the two-factor interaction.

4.3. Experimental results with five data sets from the S-RSM procedure

The S-RSM procedure is adopted in five real world data sets described in Table 1. The right-hand side of Table 4 summarizes the experimental results. Column 6 of this same table lists the selected model, while Column 7 lists the optimal re-sampling strategies obtained via the S-RSM procedure (called S-RSM strategy). For instance, the S-RSM strategy of abalone data is (1.45, 1.00), which represents the factor-level for the majority class and for the minority class, respectively. Column 8 shows R^2, while column 9 shows Adj-R^2, which represents variability in the response could be explained by the model. According to Table 4, Adj-R^2 exceeded 80% in four of the five data sets, except for the balance scale data set. The highest Adj-$R^2$was the abalone data set, which was 94.11% of

Table 1
Experimental data sets.

Data set	Number of feature	Majority class	Minority class	Ratio
Abalone	8	Ring ≠ 7	Ring = 7	(3786/391) = 9.68
Balance Scale	4	Class ≠ balance	Class = balance	(576/49) = 11.76
Letter Recognition	16	Class ≠ A	Class = A	(19,211/789) = 24.35
Mfeat-zer	47	Class ≠ 10	Class = 10	(1800/200) = 9.00
Satimage	36	Class ≠ 4	Class = 4	(5809/626) = 9.28

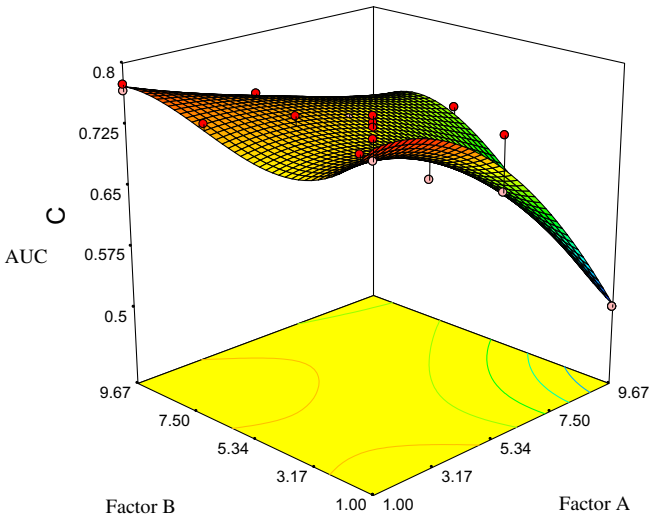

Fig. 1. Response surface of abalone data set.

variability in the response that could be explained. However, the lowest Adj-R^2 was the balance scale data set, which was merely 54.44% of variability in the response that could be explained.

A comparison was made of three other re-sampling methods frequently adopted to deal with imbalanced data, i.e. without sampling (without strategy), undersampling (S strategy) and oversampling (L strategy). Re-sampling strategies are only applied to the training set and not to a testing set. Without a strategy involves directly using training sets to construct classification models; S strategy refers to randomly selecting the majority instances until the amount of majority instances is as much as the minority class; and L strategy refers to duplicating and randomly selecting the minority instances until the amount of minority instances is as much as the majority class. For L strategy, duplicating and randomly selecting are for the integer part and the decimal part of the ratio of the amount of majority instances divided by that of the minority instances.

Next, classification models are constructed using four re-sampling strategies: without a strategy, S strategy, L strategy and S-RSM strategy. Table 4 lists the estimated AUC scores to access the classification models. According to this table, AUC scores of without a strategy were the average scores of 5-fold cross validation. The other three re-sampling strategies were obtained by averaging the 50 AUC scores (5-fold cross validation × each strategy applying re-sampling ten times). Comparing these four strategies in terms of AUC scores reveals that without a strategy performed worse than other strategies, except in the Letter recognition data set, as shown in the left portion of Table 4. Comparing the other three strategies, except for without strategy, reveals that L strategy performed worse than S strategy, i.e. a finding similarly observed in other studies (Barandela et al., 2004; Japkowicz, 2000), and significantly worse than the S-RSM strategy.

Table 5
Multiple comparison factors of the three strategies.

Data	Strategies		
	S-RSM	S	L
Abalone	_1_	_1_	_2_
Balance	_1_	_3_	_2_
Letter	_1_	_1_	_1_
Mf-zer	_1_	_1_	_2_
Satimage	_1_	_2_	_3_

Note: When applying Duncan's multiple comparison, a smaller number represents a better performance.

Next, three re-sampling strategies were compared in terms of AUC scores by applying Duncan's multiple comparison test, with a significance at the 5% level. Table 5 summarizes the results of Duncan's multiple comparison tests in terms of the ranks of strategies' performance, in which a lower number represents a better performance. According to Table 5, S-RSM strategy performed significantly better than the other strategies in four of five data sets, expect for in the Letter data set. For the Letter data set, Duncan's multiple comparison test indicated that the three strategies did not differ. Their similarity may be owing to that classifying the Letter data set is an easy classification task, resulting in the AUC scores exceeding 0.97 in the four strategies, as shown in Table 4.

5. Conclusion

When a classification model for two-class imbalanced data is developed using classifiers, the re-sampling strategy for majority and minority classes is often determined based on a trial-and-error method. The optimal re-sampling strategy determined using the trial-and-error method may not classify the imbalanced data effectively if the re-sampling strategy determined by the trial-and-error method does not include the optimal re-sampling strategy. The conventional S strategy or L strategy determines a specific re-sampling proportion. The classification model is developed based on the specific re-sampling proportion for either the majority class or the minority class. A situation in which the optimal re-sampling proportion is not the specific re-sampling proportion for the majority class or the minority class makes it impossible for the classifiers to develop an effective classification model either.

This work presents a novel analytical procedure to determine the optimal re-sampling strategy using RSM. The S-RSM procedure utilizes the oversampling and undersampling approaches simultaneously to determine the sufficient number of instances drawn from the majority class and duplicated in the minority class, respectively. By using the re-sampling strategy determined by the proposed procedure to compose the training data, a classification model can then be developed using any classifier. Additionally,

Table 4
AUC scores from re-sampling strategies and results of S-RSM strategies.

Data sets	AUC scores				S-RSM strategy			
	Without strategy	S strategy	L strategy	S-RSM strategy	Model	Optimal strategy	R^2	Adj-R^2
Abalone	0.500	0.793	0.680	**0.803**	Cubic	(1.45, 1.00)	0.9632	0.9411
Balance Scale	0.500	0.505	0.542	**0.618**	2FI	(2.22, 11.76)	0.6022	0.5454
Letter Recognition	0.983	0.978	0.977	**0.980**	Quadratic	(17.73, 1.08)	0.8399	0.7978
Mfeat-zer	0.617	**0.812**	0.744	0.807	Cubic	(2.39, 2.05)	0.8762	0.8010
Satimage	0.761	0.823	0.781	**0.833**	Linear	(1.00, 9.28)	0.9130	0.9050

Comparison of four strategies (without strategy, S strategy, L strategy and S-RSM strategy) in AUC scores, where an underline denotes the worst strategy in each data set and boldface is the best strategy.

the S-RSM strategy is not subjected to integral sampling levels. Utilizing S-RSM to deal with imbalanced data may reduce the inefficient time incurred by a trial-and-error method when composing a training set. Moreover, a more effective classification model can be constructed via the S-RSM strategy than by using a training set formed by trial-and-error. Furthermore, the S-RSM procedure is not restricted to certain classifiers and can be used in a diverse array of applications.

Finally, validity of the classification model obtained using the proposed procedure is demonstrated using five data sets. Through statistical testing (Duncan's multiple comparison), analysis results indicate that the S-RSM strategy performs significantly better than S strategy and L strategy, as shown in Table 5, further demonstrating that the S-RSM procedure can enhance the ability of a classifier to identify a class in two-class imbalanced data.

Acknowledgements

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financially supporting this research under Contract No. NSC 95-2221-E-009-187-MY3. Ted Knoy is appreciated for his editorial assistance.

References

Asuncion, A., & Newman, D. J. (2007). *UCI machine learning repository*. Irvine, CA: University of California, School of Information and Computer Science. <http://www.ics.uci.edu/~mlearn/MLRepository.html>

Barandela, R., Sánchez, J. S., García, V., & Rangel, E. (2003). Strategies for learning in class imbalanced problems. *Pattern Recognition, 36*, 849–851.

Barandela, R., Vadovinos, R. M., Sánchez, J. S., & Ferri, F. J. (2004). The imbalanced training sample problem: Under or over sampling. *Lecture Note in Computer Science, 3138*, 806–814.

Box, G. E. P., & Behnken, D. W. (1960). Some new three level designs for the study of quantitative variables. *Technometrics, 2*, 455–475.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition, 30*, 1145–1159.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861–874.

Hand, D. J., & Till, R. J. (2001). A simple generalization of the area under the ROC curve for multiple class classification problems. *Machine Learning, 45*, 171–186.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and the use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143*, 29–36.

He, G., Han, H., & Wang, W. (2005). An Over-sampling expert system for learning from imbalanced data sets. In *Proceedings of the international conference on neural networks and brain* (Vol. 1, pp. 537–541).

Irani, K. B., Cheng, J., Fayyand, U. M., & Qian, Z. (1993). Applying machine learning to semiconductor manufacturing. *IEEE Expert, 8*, 41–47.

Jahani, M., Alizadeh, M., Pirozifard, M., & Qudsevali, A. (2008). Optimization of enzymatic degumming process for rice bran oil using response surface methodology. *LWT – Food Science and Technology, 41*(10), 1892–1898.

Japkowicz, N. (2000). Learning from imbalanced data sets: A comparison of various strategies. In *Proceedings of learning from imbalanced data* (pp. 10–15).

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligent Data Analysis, 6*(5), 429–449.

Lin, X. Y., Wu, J., & Zhou, Z. H. (2006). Exploratory under-sampling for class-imbalance learning. In *Proceedings of international conference on data mining* (pp. 965–969).

McCarthy, K., Zabar, B., & Weiss, G. (2005). Does cost-sensitive learning beat sampling for classifying rare classes? In *Proceedings of the 1st international workshop on utility-based data mining* (pp. 69–77).

Montgomery, D. C. (2005). *Design and analysis of experiment* (6th ed.). New York: John Wiley and Sons.

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: Classification of skewed data. *ACM SIGKDD, 6*, 50–59.

Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proceedings of the 15th international conference on machine learning* (pp. 445–453).

Quinlan, J. R. (1993). *C4.5: Program for machine learning*. San Mateo, CA: Morgan Kaufmann.

Shin, S., Guo, Y., Choi, M., & Kim, C. (2007). Development of a robust data mining method using CBFS and RSM. *Lecture Notes in Computer Science, 4378*, 377–388.