Indian Journal of Science and Technology, Vol 9(42), DOI: 10.17485/ijst/2016/v9i42/100812, November 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Evaluation of Cost Sensitive Learning for Imbalanced Bank Direct Marketing Data

Khor Kok-Chin* and Ng Keng-Hoong

Faculty of Computing and Informatics, Multimedia University, 63100, Cyberjaya, Selangor;
kckhor@mmu.edu.my, khng@mmu.edu.my

Abstract

Objectives: The imbalanced bank direct marketing data set utilized in this study is a two-class data mining problem, where a customer may or may not subscribe a product from a bank. **Methods/Statistical Analysis**: The data set inherited the rare class problem where the classification rate attained for the rare class is low. In this study, we attempted cost sensitive learning to mitigate the problem, and to address that there are various costs involved when misclassification occurs. Three learning algorithms, namely, Naive Bayes (NB), C4.5 and Naive Bayes Tree (NBT) were involved in the cost sensitive learning and their results were empirically evaluated. **Findings:** The results were also compared with two previous studies that utilized the cost insensitive SVM and over-sampling, respectively. Although cost sensitive learning is claimed able to handle imbalanced data sets, but we noticed that the learning is less effective for the bank direct marketing data set in overall. Cost sensitive learning provides a way of "wrapping" learning algorithms that are not designed to handle imbalanced class distributions. Therefore, it may not work well for certain imbalanced data sets. Over-sampling, on the other hand, worked well for the data set. **Improvements/Applications:** Over-sampling helped to generalize the decision region of the rare class clearly and subsequently improved the classification result.

Keywords: Bank Direct Marketing, Cost Sensitive Learning, Imbalanced Data Set, Rare Class Problem, Over-Sampling

1. Introduction

Direct marketing plays an important role for banks to promote their products and services directly to their customers. Advanced communication technology such as Voice over IP has enabled banks to reach their customers directly and extensively while keeping the communication costs low. Consequently, a large amount of customer data can be collected by them. Analysing such data allows banks to identify potential customers that can purchase their products and services. Although it is not easy to analyse large amounts of customer data, but it is now possible with data mining. Data mining is an analysis to find relationships in observational data sets and to provide an informative summarization of data to users[1]. The data sets involved are usually large in dimension as well as size.

Bank direct marketing can be viewed as a two-class data mining problem, where a customer accept or reject a product or service promoted by a bank. Regardless of the communication channels used (telephone, email, SMS, etc.), the positive response from customers are usually rare. Therefore, the data collected are normally imbalanced in its class distribution[2,3]. Applying classification algorithms on such imbalanced data could cause the rare class problem where identifying potential customers becomes difficult.

The rare class problem is a major challenge in data mining research, especially in research domains where the interested class to study and analyse is always very few in a data set[4]. The causes of the rare class problem are summarized as follows[5].

(1) The rare class (customers that subscribe a bank product or service) is relatively few as compared with the other class. In real-life problems, the ratio of the rare class to the other class can be 1 to 100, 1 to 1,000, 1 to 10,000 or even more.

*Author for correspondence

(2) The other class has overwhelmed the rare class, in the way that the other class has overlapped feature values with the rare class. Such condition creates difficulty for many learning algorithms in generalizing a decision region for the rare class. A decision region is a volume of the feature space in a data set where instances of the same class reside in. Consequently, the rare class cannot be identified well by learning algorithms.

There are two basic approaches for solving the rare class problem inherited in imbalanced data sets: (1) data-level and (2) algorithm-level approaches[6]. The data-level approach involves balancing of class distribution in a data set to suit the nature of general learning algorithms that are weak in identifying rare classes. Example of data-level approaches includes under-sampling, over-sampling and the combination of both samplings[7]. Algorithm-level approach, on the other hand, makes learning algorithms suitable for imbalanced data set without manipulating a data set; the approach provides solutions which introduce bias based on a data set, train a classifier for rare classes (classifier specific). Cost sensitive learning is an example of algorithm-level approach to improve the detection rates for rare classes in a dataset. Using cost sensitive learning, error costs of classifiers are not treated equally[8]. Instead, error costs are assigned differently for the classifiers involved. Another example of algorithm-level approach is outlier detection. Outlier detection searches patterns that do not comply with the norm[9].

In our previous study, we attempted a data-level approach, which is over-sampling, to mitigate the rare class problem inherited in a bank direct marketing data set[10]. Using the same data set, we attempted cost sensitive learning of the algorithm-level approach in this study. Results from both studies and a previous study from other authors are then compared and justifications are provided on why a particular approach outperformed the other.

2. The Bank Direct Marketing Data Set

The bank direct marketing data set utilized in this study was provided[11]. It contains customer data collected from 14 direct marketing campaigns organized by a Portuguese Banking Institution. The communication channel used was telephone. The data set can be found in the UCI machine learning repository website. The full data set consists of 16 features and 1 class (y) Table 1. If a customer subscribed the bank service, the class value would be YES and otherwise NO. A further study using a larger bank direct marketing data set was also conducted by the same researchers[12].

Table 2 shows the class distribution of the data set where the rare class (YES) takes only 11.7% of the data set. Although the class distribution is considered not extremely imbalanced as compared to the other research domains such as network intrusion detection[13], it might

Table 1. The bank direct marketing data set

Feature	Description	Data Type
Age	The age of the client.	Numeric
job	The type of job that the client is having.	Nominal
marital	The marital status of the client.	Nominal
education*	The type of education that the client currently in.	Nominal
default*	Whether the client has credit in default?	Nominal
balance	The average yearly balance of the client in Euros.	Numeric
housing*	Whether the client has housing loan?	Nominal
loan*	Whether the client has personal loan?	Nominal
contact*	Type of communication to contact the client.	Nominal
day*	Last contact day of the month.	Numeric
month*	Last contact month of the year.	Nominal
duration*	Last contact duration in seconds.	Numeric
campaign	Number of contacts performed for this client during this campaign.	Numeric
pdays	Number of days that passed by after the client was last contacted from previous campaigns.	Numeric
previous	Number of contacts performed for this client before this campaign.	Numeric
poutcome*	Outcome of the previous marketing campaign.	Nominal
y	Whether the client subscribed a term deposit?	Nominal

* The important features determined by our previous research[10].

still cause the rare class problem. To determine whether or not the rare class problem occurs, the full data set was preliminarily trained and tested with three learning algorithms, namely, NB, C4.5, and NBT.

Using the full data set, these three algorithms were unable to produce high True Positive (TP) rates for the rare class (YES). The results are as shown in Table 3. Such unsatisfactory results can be explained via matrix plots by examining the pairwise relationship between the features of the data set. Generally the matrix plots generated from the data set show the phenomenon as follows.

Figure 1 consists of three matrix plots showing the instances of the full data set using features named "education" and "poutcome", "education" and "poutcome", and "duration" and "month". The decision regions are difficult to generalize for class YES (red crosses) because: (1) class NO (blue crosses) are large in number as compared with class YES and (2) instances of class YES are overlapped by class NO.

3. Cost Sensitive Learning

The algorithm-level approach is commonly used by data mining researchers to solve the rare class problem. Making a classifier cost sensitive is an example of the algorithm-level approach, which is used to improve classification rates for rare classes in an imbalanced data set[14]. Making a classifier cost sensitive has been identified as one of the top 10 challenging problems in data mining.

Table 2. The class distribution of the bank direct marketing data set

Class	No. of instances	Distribution in Percentage
NO	39,922	88.3%
YES	5,289	11.7%
Total	45,211	100.0%

Table 3. The results of the preliminary study of the bank direct marketing data set

Class	TP rate (NO)	TP rate (YES)	ROC
NB	0.927	0.528	0.861
C4.5 (c-0.09)*	0.964	0.463	0.880
NBT	0.961	0.444	0.900

* c value is the confidence factor for optimizing the performance of C4.5.

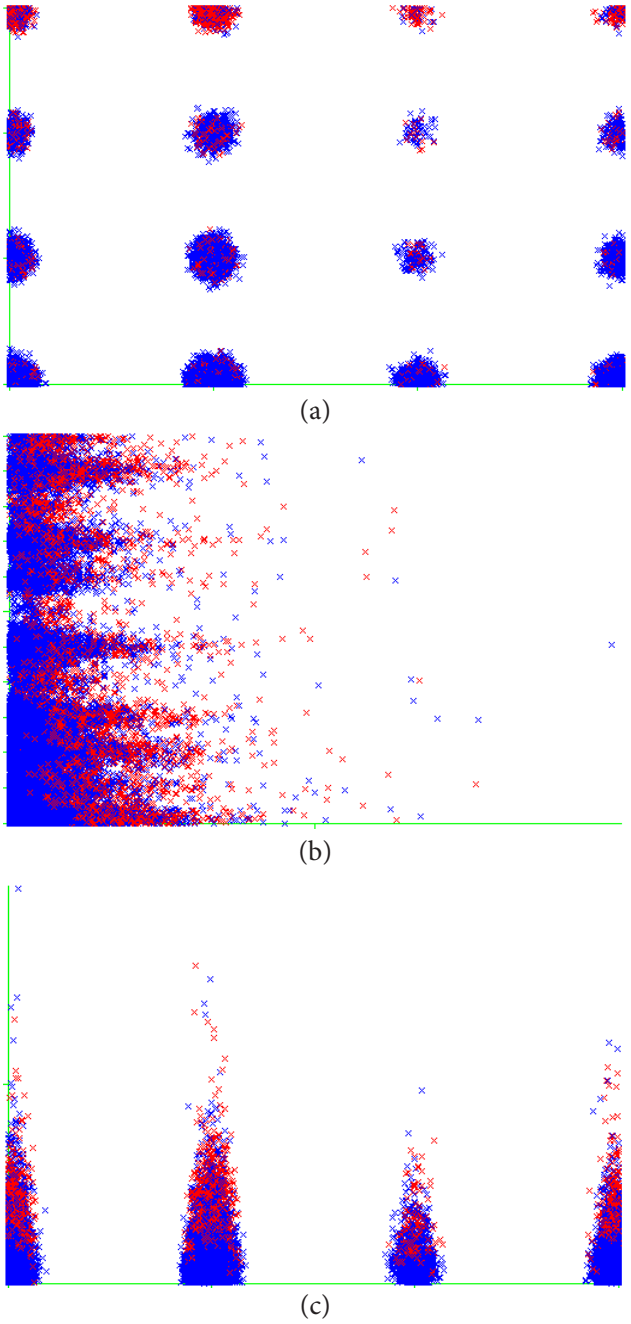


Figure 1. Three matrix plots showing the pairwise relationship between features (a) "education" and "poutcome" (b) "duration" and "month" (c) "education" and "duration" of the full data set.

Traditional learning algorithms assume equality for all classes in a data set and they tend to minimize misclassification rates by not favoring rare classes. Therefore, low classification rate is normally attained for rare classes, particularly when the causes of the rare class problem exist (refer to Introduction). In real-life situations, there

is always an associated cost with every error made (misclassification). Costs can be monetary, waste of time, the severity of a consequence, etc.[15]. The cost of misclassifying a rare class (an interested class) is always greater than the cost of misclassifying a non-rare class[16]. Taking the example of the bank direct marketing, the cost of not approaching a potential subscriber is always higher that calling a customer who is not going to subscribe any product or service. Therefore, different error costs should be considered to build a realistic classifier using any learning algorithm.

There are two ways of making a classifier cost sensitive[17]. Firstly, *cost sensitive classification* which alters the classifier output by changing its probability threshold. The objective is to minimize the cost of misclassification. Secondly, *cost sensitive learning* which builds a cost sensitive classifier by sampling: (1) under-sampling the major classes or (2) over-sampling the rare classes; however, the former may remove important instances of a data set while the latter may cause over-fitting problem. Therefore, a more practical way of *cost sensitive learning* is to reweight instances in a data set by referencing to the relative cost of false positives and false negatives, and then relearn them[18].

4. Methodology

Firstly, we discuss the features used in this study. Secondly, we discuss the cost matrix and evaluation metrics used for cost sensitive learning. Only nine features of the data set were used in this evaluation .This feature subset was selected based on our previous study with the reason to reduce the large number of hypotheses generated using all features. A hypothesis is a regularity that predicts classes which can be found on a given dataset. Suppose a dataset has N features and a class, both binary types, then the number of hypotheses is 2^{2^N} . Feature selection is, therefore, necessary to select only important and relevant features in this data set. The selection of the nine features is described as follows. In our previous study, we utilized a number of feature selection evaluators with different search strategies. Classifier Subset Evaluator which incorporated with NB, scatter search and 10-fold cross validation had resulted this feature subset with the best classification result as compared with the other feature selection evaluators.

We focused on only *cost sensitive learning* (by reweighting instances) in this study. Table 4 and Table 5 show the confusion matrix and the cost matrix for the bank direct marketing. In Table 5, there are four notations used to represent the error costs involved, namely, C_{nn}, C_{ny}, C_{yn} and C_{yy}. Take the example of C_{yn}, it is the cost of misclassifying an actual class YES to Class NO. We assume, in general, predicting a class correctly does not cost anything. Instead, they are regarded as "benefits"[19]. Therefore, the error cost for C_{nn} and C_{yy} are both zero. The remaining error costs are: (1) the cost of misclassifying a potential customer to be non-subscriber (C_{yn}) and (2) the cost of misclassifying a non-potential customer to be a potential one (C_{ny}). Obviously, C_{yn} is higher than C_{ny} as the bank will lose a possible long term earning if it does not approach the potential customer. Since we do not have the actual error costs, it will be realistic to train classifiers with different C_{yn} values (from 1 to 5) and then evaluate them. For relative comparison, C_{ny} shall be set to one.

Table 4. The confusion matrix

	Predicted NO	Predicted YES
Actual NO	True Negative (TN)	False Positive (FP)
Actual YES	False Negative (FN)	True Positive (TP)

Table 5. The cost matrix for the bank direct marketing

	Predicted NO	Predicted YES
Actual NO	C_{nn} , 0	C_{ny} , 1
Actual YES	C_{yn}, [1,5]	C_{yy}, 0

To compare cost sensitive classifiers, the total cost metric can be used. The lower the cost, the better a classifier is. The formula for the metric is as follows:

$$Total\ cost = (false\ negative\ x\ C_{yn}) + (false\ positive\ x\ C_{ny})$$

Besides the total cost, Receiver Operating Characteristic (ROC) is another metric to compare the classifiers. ROC shows the trade-offs between TP and FP. An ideal classifier gives 1 for TP and 0 for FP. We also used TP rates in this study; they can be used to monitor the performance of the classifiers on the particular rare class.

5. Results

Three learning algorithms were initially evaluated, namely, NB, C4.5 and NBT in building cost sensitive classifiers. The confidence factor of C4.5 is 0.09 for the performance optimization of the classifier. The results of the cost sensi-

tive classifiers are as shown in Table 6. Total costs in Table 6 were used to plot a graph to compare the cost sensitive classifiers Figure 2.

Initially, the performance of the classifiers on the rare class (YES) was evaluated. When the error cost increased, the TP rate for the rare class increased. It seemed worth to keep on increasing the error cost so that the TP rate for class YES can be increased. But note that when the error cost increased, the TP rate for class NO also decreased.

Subsequently, the classifiers were compared using ROC and total cost. NB gave the lowest ROC values and the highest total costs as compared with C4.5 and NBT, regardless of the error cost ratios used. Therefore, NB was excluded from our consideration. We then compared the remaining two learning algorithms. The total costs of C4.5 had been always slightly lower than NBT. However, NBT gave slightly better performance than C4.5 if both of them were evaluated using ROC. Notwithstanding that the results were inconsistent, the difference between C4.5 and NBT was not significant.

Table 6. The results attained using NB, C4.5 and NBT

Learning algorithm (C_{ny}: C_{yn})	TP Rate (NO)	TP Rate (YES)	ROC	Total Cost
NB (1:1)	0.963	0.423	0.888	4540
NB (1:2)	0.942	0.533	0.888	7250
NB (1:3)	0.929	0.590	0.888	9336
NB (1:4)	0.916	0.628	0.888	11281
NB (1:5)	0.904	0.663	0.888	12756
C4.5(1:1)	0.965	0.467	0.880	4225
C4.5(1:2)	0.923	0.704	0.890	6214
C4.5(1:3)	0.899	0.777	0.872	7577
C4.5(1:4)	0.882	0.805	0.865	8822
C4.5(1:5)	0.869	0.826	0.854	9819
NBT(1:1)	0.963	0.479	0.911	4250
NBT(1:2)	0.925	0.681	0.911	6355
NBT(1:3)	0.908	0.743	0.903	7739
NBT(1:4)	0.895	0.774	0.899	8975
NBT(1:5)	0.883	0.801	0.899	9935

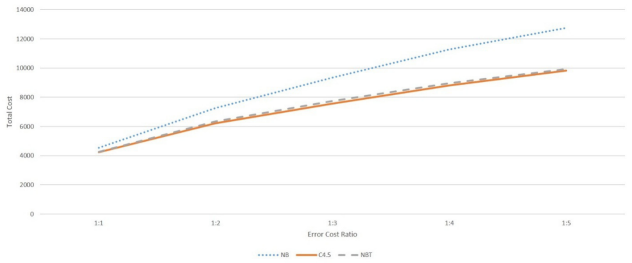


Figure 2. Comparing the learning algorithms using total cost based on various error cost ratios.

The evaluation for cost sensitive classifiers was then continued with bagged NB (NB_bg), bagged C4.5 (C4.5_bg) and bagged NBT (NBT_bg). With bagging, classification models are built using different data set samples by bootstrapping. The classification result is based on the majority votes of the models. We evaluated bagging in this study for the reasons as follows. Firstly, a number of studies showed that the results of cost sensitive learning can be improved by integrating bagging[20-22]. Secondly, bagging is also claimed to be able to handle imbalanced data sets well[23-25]. The results are as shown in Table 7 and Figure 3. Same as the three unbagged algorithms, when the error cost increased, the TP rate for the rare class also increased. On the other hand, the TP rate for the class NO decreased.

NB_bg was excluded from our consideration as it gave the lowest ROC values and the highest total costs among the three bagged algorithms Table 7. The remaining two bagged algorithms were then compared. The total costs of NBT_bg were generally lower than C4.5_bg, except for the error cost ratio 1:3. In addition, ROCs of NBT_bg were higher than C4.5_bg, regardless of the error cost ratios. In

Table 7. The results attained using NB_bg, C4.5_bg, and NBT_bg

Learning algorithm (C_{ny}: C_{yn})	TP Rate (NO)	TP Rate (YES)	ROC	Total Cost
NB_bg (1:1)	0.963	0.423	0.888	4543
NB_bg (1:2)	0.942	0.533	0.888	7250
NB_bg (1:3)	0.929	0.591	0.888	9333
NB_bg (1:4)	0.916	0.627	0.888	11252
NB_bg (1:5)	0.904	0.663	0.888	12746
C4.5_bg (1:1)	0.965	0.462	0.917	4236
C4.5_bg (1:2)	0.932	0.669	0.923	6224
C4.5_bg (1:3)	0.912	0.739	0.922	7651
C4.5_bg (1:4)	0.900	0.773	0.921	8790
C4.5_bg (1:5)	0.888	0.797	0.918	9822
NBT_bg (1:1)	0.967	0.460	0.930	4176
NBT_bg (1:2)	0.933	0.668	0.929	6172
NBT_bg (1:3)	0.915	0.731	0.926	7684
NBT_bg (1:4)	0.903	0.770	0.926	8732
NBT_bg (1:5)	0.894	0.796	0.926	9631

general, NBT_bg performed better than C4.5_bg but the difference between them was also not significant.

The results in both Table 6 and Table 7 were then compared. Excluding the less effective NB and NB_bg, the differences between these four algorithms, namely, C4.5, NBT, C4.5_bg, and NB_bg were not significant.

5.1 Compare with Previous Studies

We also compared our results with two previous studies as shown in Table 8. Without using cost sensitive classifier, the result of the authors who provided the data set, was better than the results obtained in this study. However, we were only able to compare with the authors using ROC. The performance of the classifier on the rare class, total cost involved were not available.

In our previous study, over-sampling was conducted on the full data set using SMOTE and then the data set was reduced by preserving only nine features. Different sampling rates for the rare class were tested, started with 100% and ended with 1000%. According to the study, changing the distribution of negative and positive classes will result different misclassification costs. For instance, changing the class distribution from 1:1 to 1:2 will yield a misclassification cost of 1:2. The reduced data set was then trained and tested with NBT. The best result was obtained by over-sampling the rare class for 1000%; the error cost ratio is 1:1.46 (NO: YES). The result was superior to the study as well as this study. This was because not only the ROC had increased, the TP rate for the rare class had also increased. In addition, the TP rate for No class was also well maintained.

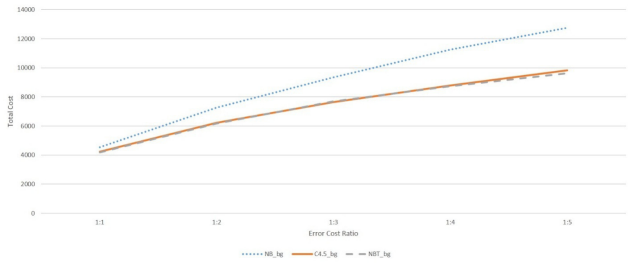


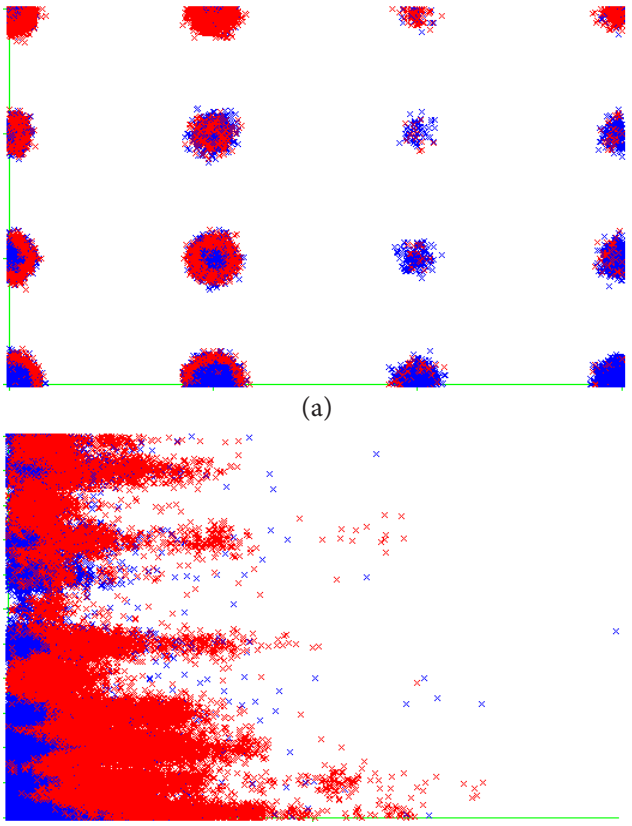
Figure 3. Comparing three bagged learning algorithms using total cost based on various error cost ratios.

Table 8. The results of two previous studies

Method used (error cost ratio)	Support Vector Machine (1:1)	Over-sampling 1000%, NBT (1:1.46)
TP Rate (NO)	N/A	0.969
TP Rate (YES)	N/A	0.929
ROC	0.938	0.987
Total Cost	N/A	5389

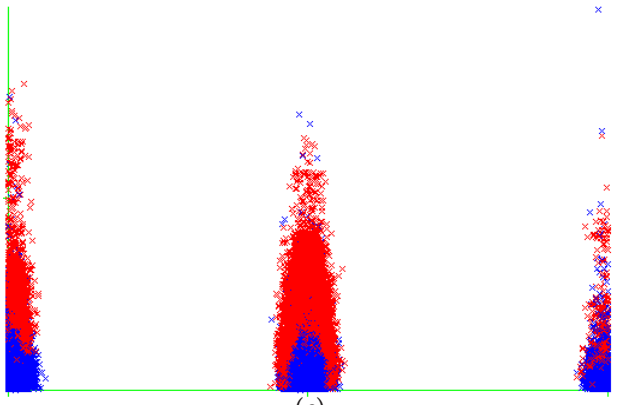
6. Discussion

We discuss if cost sensitive learning is necessary for the data set. Comparing with two previous studies showed that it is unnecessary, especially with the outstanding previous results using over-sampling. Why over-sampling outperformed cost sensitive learning? Firstly, many existing learning algorithms are not designed to handle issues such as imbalanced data sets and to be cost sensitive. Cost sensitive learning just provides a way of "wrapping" the learning algorithms to handle the issues. Therefore, cost sensitive learning may not work well for certain data sets. Secondly, we noticed that the over-fitting problem that could be triggered by over-sampling (as discussed in Cost Sensitive Learning) did not occur in this data set. After over-sampling, the decision region for the rare class can be clearly generalized. Therefore, the learning algorithm involved can identify the rare class easily. The matrix plots in Figure 4 show that class YES is no longer overwhelmed by class NO.



(a)

(b)



(c)

Figure 4. Three matrix plots showing the pairwise relationship between features (a) "education" and "poutcome" (b) "duration" and "month" (c) "education" and "duration" of the reduced data set after over-sampling.

7. Conclusion

In this study, we applied cost sensitive learning using three bagged and un-bagged learning algorithms with different error cost ratios to the imbalanced bank direct marketing data set. We evaluated the classifiers using the results obtained in this study and then compared the results with the two previous studies.

To conclude, cost sensitive learning (an algorithm-level approach) is less effective for the bank direct marketing data set. Over-sampling (a data-level approach) should be the preference instead.

We would like to consider the followings in our future research. Firstly, to evaluate cost sensitive classification and to find out how would learning algorithms respond when the probability threshold is adjusted according to different error cost ratios. Secondly, to consider boosting in evaluation.

8. References

1. Hand DJ, Mannila H, Smyth P. MIT press: Principles of Data Mining. 2001 Aug; p. 1-578.
2. Ling CX, Li C. Data Mining for Direct Marketing: Problems and Solutions. KDD. 1998 Aug; 98:73-9.
3. Wong KW, Zhou S, Yang Q, Yeung JM. Mining Customer Value: From Association Rules to Direct Marketing. Data Mining and Knowledge Discovery. 2005 Jul; 11(1):57-79.
4. Yang Q, Wu X. 10 Challenging Problems in Data Mining Research. International Journal of Information Technology & Decision Making. 2006 Dec; 5(04):597-04.
5. Chawla NV, Japkowicz N, Kotcz A. Editorial: special issue on learning from imbalanced data sets. ACM Sigkdd Explorations Newsletter. 2004 Jun; 6(1):1-6.
6. Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering. 2006 Dec; 30(1):25-36.
7. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002 Jan; 16(1):321-57.
8. Fan W, Lee W, Stolfo SJ, Miller M. Springer Berlin Heidelberg: A multiple model cost-sensitive approach for intrusion detection. European conference on machine learning. 2000 May; p. 142-54.
9. Chandola V, Banerjee A, Kumar V. Outlier detection: A survey. ACM Computing Surveys. 2009 Jul; 41(3):1-58.
10. Khalid SN, Khor KC, Ng KH, Ting CY. Effective Classification for Unbalanced Bank Direct Marketing Data with Over-sampling. KMICE. 2014 Aug; p. 16-21.
11. Moro S, Laureano R, Cortez P. Using data mining for bank direct marketing: An application of the crisp-dm methodology. Proceedings of European Simulation and Modelling Conference. 2011 Dec; p. 1-5.
12. Moro S, Cortez P, Rita P. A data-driven approach to predict the success of bank telemarketing. Decision Support Systems. 2014 Jun; 62:22-31.
13. Khor KC, Ting CY, Phon-Amnuaisuk S. Springer International Publishing: The effectiveness of sampling methods for the imbalanced network intrusion detection data set. Recent Advances on Soft Computing and Data Mining. 2014 Jun; p. 613-22.
14. Zhou B, Liu Q. Springer Berlin Heidelberg: A comparison study of cost-sensitive classifier evaluations. International Conference on Brain Informatics. 2012 Dec; p. 360-71.
15. Elkan C. The foundations of cost-sensitive learning. International joint conference on artificial intelligence. 2001; 2:973-78.
16. Weiss GM, McCarthy K, Zabar B. DMIN.CSREA Press: Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs. 2007; p. 35-41.
17. Zhao H. Instance weighting versus threshold adjusting for cost-sensitive classification, Knowledge and Information Systems. 2008 Jun; 15(3):321-34.
18. Hall M, Witten I, Frank E. Kaufmann, Burlington: Data mining: Practical machine learning tools and techniques. 2011 Jan.
19. Ling CX, Sheng VS. Springer US: Cost-sensitive learning. Encyclopedia of Machine Learning. 2008; p. 231-35.
20. Sheng VS, Ling CX. Springer Berlin Heidelberg: Roulette sampling for cost-sensitive learning. European Conference on Machine Learning. 2007 Sep; p. 724-31.

21. Sittidech P, Nai-arun N, Nabney IT. Bagging Model with Cost Sensitive Analysis on Diabetes Data. Information Technology Journal. 2015 Jan-Jun; 11(1):1-9.
22. Wang B, Pineau J. Online ensemble learning for imbalanced data streams. 2013 Oct. arXiv preprint arXiv: 1310.8004.
23. Khoshgoftaar TM, Van Hulse J, Napolitano A. Comparing boosting and bagging techniques with noisy and imbalanced data. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans. 2011 May; 41(3):552-68.
24. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2012 Jul; 42(4):463-84.
25. Sumana BV, Santhanam T. Optimizing the Prediction of Bagging and Boosting. Indian Journal of Science and Technology. 2015 Dec; 8(35):1-13.