tion Strategy. *Clinical Therapeutics*, 12(3) , p.p.23-30.

2. Su Xiu Xu, Qiang Lu, Zhuoxin Li (2012) Optimal modular production strategies under market uncertainty: A real options perspective. *International Journal of Production Economics*, 16(7) , p.p.34-41.
3. Bartolomeu Fernandes, Jorge Cunha, Paula Ferreira (2011) The use of real options approach in energy sector investments. *Renewable and Sustainable Energy Reviews*, 19(12) , p.p.71-80.
4. Shun-Chung Lee, Li-Hsing Shih (2010) Renewable energy policy evaluation using real option model – The case of Taiwan . *Energy Economics*, 12(6) , p.p.21-32.
5. E.A. Martínez-Ceseña, J. Mutale (2011) Application of an advanced real options approach for renewable energy generation projects planning. *Renewable and Sustainable Energy Reviews*, 5(2), p.p.59-69.
6. Lei Zhu, Ying Fan (2011) A real options–based CCS investment evaluation model: Case study of China’s power generation sector. *Applied Energy*, 6(11) , p.p.11-20.

Research on the Database Marketing in the Big Data Environment Based on Ensemble Learning

Yu Wang, Chengqun Yu

School of Economics and Business Administration, Chongqing University,
Chongqing, 400030, China

Corresponding author is Yu Wang

Abstract

With the increasement of market competition and marketing costs, more and more companies use marketing database model to analyze and identify potential customers interested in marketing activities or products which may cause the problem of bad performance in data processing and data redundancy. In order to solve this problem, targeting customers is considered in database marketing as a classification and prediction problem in data mining under big data enviroment. Due to the variety and class imbalance of customers, a database marketing model based on supervised clustering and ensemble learning is proposed. The empirical study indicates that the proposed approach is able to improve the performance of database marketing.

Keywords: DATABASE MARKETING; CLASSIFICATION AND PREDICTION; SUPERVISED CLUSTERING; ENSEMBLE LEARNING

1. Introduction

Marketing is an important means for enterprises to develop the market. With the increase of market competition and marketing costs, more and more companies use marketing database model to analyze and identify potential customers interested in marketing activities or products, and use e-mail, SMS, telephone and other means of customer depth mining and relationship maintenance with customers to establish a pair of interactive communication. Compared with the traditional marketing mass, database marketing makes consumers with different needs and characteristics to get the marketing information they need. At the same time, database marketing can make the enterprise focus on target customers, improve their satisfaction and loyalty, reduce marketing costs, enhance the enterprise's market competitiveness.

A key problem of database marketing is to correctly locate target clients. In the example of Baesens etc [1] the accurate target client locating can add 500,000 Euro extra earnings for every extra percentage. Knott [2] etc point out that for a retail service bank, 0.7% extra percentage correct target client locating can help to improve 20% revenues for each customer. Judging from the perspective of data mining, the target client locating can be beckoned as classification problem, namely to predict the percentage for customers buying or to buy product based on their consumption characteristics. However, the diversity and unbalanced category of consumer groups restrict traditional classification predication technique. To begin with, among consumer groups, the number of target consumers is far lower the number of non-target clients, in other word, a class imbalance problem. The traditional classification predication technique aims to minimize the risk and it will be hard to effectively deal with class imbalance [3-4]; what's more, consumer groups are extensive and diverse, so single classification predication model cannot accurately reflect various consumption model as well as characteristics and over-fitting problem for learning model may appear [5-6].

In order to solve this problem, 2e consider targeting customers in database marketing as a classification and prediction problem in data mining, i.e., to predict whether a customer would purchase a product or the probability of purchasing based on his/her characteristics. Due to the variety and class imbalance of customers, a database marketing model based on supervised clustering and ensemble learning is proposed. The empirical study indicates that the proposed approach is able to improve the performance of database marketing.

2. Related Works

The research of data classification problem can be divided into two categories: data level and algorithm level. The research of data level mainly uses two kinds of sampling techniques, under sampling (under-sampling) and over sampling (over-sampling). In the problem of reducing the sample size of the majority of the class, the problem of [7] is overcome. The problem is solved by increasing the sample size of a small number of samples. Chawla [9] based on the idea of the SMOTE algorithm, the algorithm to each of the sample points of the class as the center of its K nearest neighbor sample, in which the sample and its neighbors randomly generated between the new sample points. SMOTE [10] proposed a new method of improving the re sampling method. The combination of over sampling and under sampling is used to select the nearest neighbor and synthesize the sample with different strategies to eliminate the influence of the classification. Rahman et al. [11] design a semi supervised sampling method. The method is based on clustering technique. The empirical results show that the method can effectively improve the classification effect of the imbalanced data sets. Alejo [12] proposed a new dynamic sampling method, which combines the SMOTE method and the sequential backward propagation algorithm to increase the number of samples, and then overcome the impact of the category imbalance on the classification forecast.

The research of the algorithm is mainly aimed at the characteristics of data classification. The common methods include [13] (learning) cost-sensitive (ensemble) and [14] (learning). The main idea of ensemble learning is to improve the whole learning effect by integrating multiple different basic learning devices. In recent years, ensemble learning method has been studied in this field with its stronger generalization ability and better learning effect. [15]. The Bagging algorithm proposed by Breiman [16] is trained by the multiple sampling subset of the training set, so as to get multiple learners. Experiments show that the Bagging algorithm can effectively improve the accuracy of learning. Chawla et al. [17] algorithm and SMOTE algorithm combined with Boosting algorithm, so as to increase the number of samples, which can improve the learning algorithm to identify a small number of classes. Maclin [18] uses decision tree and artificial neural network to compare the effect of two kinds of ensemble learning methods, Boosting and Bagging. The results show that ensemble learning is always better than the single method. Sun et al. [19] proposed a novel ensemble learning method. This method first divides the sample set into a set of multiple categories, and then integrates them. The experimental results show that the method can solve the two classification problems. Zhai Yun et al. [20] proposed a resampling method based on different sampling rate, and designed a new type of ensemble classifier based on this method.

Even though database marketing research has achieved important progress, an existing problem is to effectively optimize the difference and individual performance among various learning models so as to improve the accuracy of database marketing. The main idea of the learning model generated in existing ensemble studying research is to repeatedly and randomly take samples in sample space or feature space which yet cannot guarantee the big difference among various learning models. At the same time, the training subset created by random sampling cannot accurately depict the diversity and difference among consumer groups and as a result the accuracy of establishing learning model based on training subset will decrease. In order to solve above problems, this paper comes up with ensemble learning model based on supervised clustering which firstly adopts K-Means to gather the non-target consumers and divide them into several consumer subgroups with big differences. Then it combines subgroups with few respondent consumers so as to have several training subsets to train artificial neural network sample learning model in several subgroups to carry out integration and overcome the data imbalance problem meanwhile improve the learning performance of every single learner.

3. The Ensemble Learning Based on Supervised Clustering

The ensemble learning based on supervised clustering firstly gather majority class samples, and then combines aggregate of data with minority class sample to get several subsets. Then, artificial neural network can be trained in several subsets to carry out integration. Commonly, clustering is a unsupervised learning process [21], namely to divide data set into several aggregates of data with differences without any prior knowledge. The clustering in this paper is to cluster the data with certain signs in the premise of knowing data class signs and therefore, it is also known as supervised clustering. A representative approach in clustering is K-Means and this paper adopts it because of its merits [22-23] for example it is proper to deal with large scale data with high efficiency.

3.1. The K-means Algorithm

K-Means algorithm inputs data set S and aggregate of data number K to repeatedly search and calculate so as to output the optimal measure function convergence aggregate of data K[24]. *K*-Means algorithm can compact the internal parts of various aggregates of data and separate various aggregates of data and its main steps show as follows: (1) randomly extract K as the original aggregate of data center in data set S; (2) calculate the Euclidean distance between every node and the center, then divide this node into the cluster with the shortest distance with the center; (3) calculate the measure function of existing clustering results to relocate the center of aggregate of data based on this; (4) repeat step two and three to measure function convergence so as to have the final clustering result.

3.2. The Model

The ensemble learning based on supervised clustering can be divided into three stages of data pre-processing, supervised clustering and ensemble learning.

The first stage is data pre-processing. The dimension property of consumer is different (for example, if the property of income is over 1000, the property of age is dozens). In clustering analysis dimension discrepancy will cause a big influence for those properties with big differences while for those with small differences, the influence can be ignored. Therefore, before carrying out clustering analysis, we have to carry out data pre-processing so as to make the dimension property consistent. This paper adopts min-max normalized methods to conduct linear conversion on data and the formula shows as follows:

$$\overline{x}_{ij} = \frac{x_{ij} - \min_j}{\max_j - \min_j}, (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m) \quad (1)$$

In this formula, n means the table cardinality and m refers to data attribute number (not including class label), x_{ij} refers to the original value of i record's j property, $\overline{x}_{ij}$ refers to the value of i record's j property after standardized implementation, $\min_j$ is the minimum value of j property while $\max_j$ is the maximum value of j property . After standardizing formula (1) we can make the value range of all properties become $[0,1]$ so as to avoid the influence caused by different dimension properties.

The second stage is supervised clustering. Firstly calculate the proportion of majority class and minority class, set K is equal to the proportion in the *K*-Means algorithm and then we carry out clustering analysis on the majority class and cluster the majority class into K aggregates of data. Based on this, we recombine each aggregate of data with minority class to form K sub set sample with relevant balance.

Suppose the training sample set of N consumer information is $G = \{y_i, \mathbf{x}_i\}_{i=1}^{N}$. The individual information property of each consumer (such as age, sex and family members) shows in the form of $\mathbf{x}_i \in R^p$, among which p refers to the property number of customers. Suppose G_{pos} and G_{neg} ($G_{pos} \cup G_{neg} = G$) respectively refers to the customer set who response and do not response to marketing activities and in general. In order to explore the consumer subgroup with different consumption model and characteristics and at the same time overcome the problem of class imbalance we set the K value in K-Means algorithm is $|G_{neg}|/|G_{pos}|$ and adopts this value to carry out clustering analysis on the majority class (which does not affect consumers). Based on this, we combine every aggregate of data and minority class (which can affect consumers) to form K subgroups which are relevantly balanced. The algorithm flow chart for supervised clustering is shown as follows:

The third stage is to establish basic learning model and carry out ensemble learning. BP neural network has good learning ability and fault tolerant ability in classification predication and this paper adopts it as the basic learning algorithm.

In supervised clustering stage, we train the BP neural network of every sample set $subG^k, (k = 1, 2, \cdots, K)$ so as to have K basic learning model $Learner^1, Learner^2, \cdots, Learner^K$. Different basic learning models have different accuracy for different data samples and how to choose the ensemble learning method for basic learning model is a key issue. Comparing from choosing the optimal model among all basic learning models, a better approach is to choose different learning modes for different data samples (commonly known as dynamic integration) [27]. This paper takes weighted voting based on sample neighborhood learning accuracy as the integration method which belongs to dynamic integration.

Concerning the unknown data record $\mathbf{x}$, the first step is to find out the close sample NN_x^k of $subG^k$ based on KNN algorithm so as to get the predict output of $Learner^1, Learner^2, \cdots, Learner^K$ and $\mathbf{x}$ as well as its close sample. P_x^k is the predict percentage on $\mathbf{x}$ carried out by basic learning model $Learner^k$ while φ_x^k is the accuracy of $Learner^k$ in NN_x^k. When integrate the predict output of K basic models we can finally get the integrated forecasting output FP_x and the formula shows as follows:

$$FP_x = \sum_{k=1}^{K} P_x^K \frac{\varphi_x^K}{\sum_{k=1}^{K} \varphi_x^K} \qquad (2)$$

$$(k = 1, 2, ..., K)$$

The output of ensemble learning in whole testing set is $FP = (FP_1, FP_2, ...FP_x)$.

Based on the above three stages, the overall flow chart for database marketing model based on supervised clustering and ensemble learning shows in figure 2.

4. Results and Discussion

This paper adopts the data of predict contest of COIL (Computational Intelligence and Learning) in 2000 as the empirical research. This data includes the sample data of 9822 European families buying the car insurance. What's more, this paper divides this data set into training set and testing set, among which training set includes 5822 data which is used to establish the ensemble learning model put forward in this paper while another 4000 data is used to assess the effect of this model. Among the 5822 training data set, there are only 348 minor class sets, occupying 5.97% of the training set. Among the 4000 testing data, there are 238 minor class sets, occupying 5.95%. Therefore, we can conclude that the data set has obvious class imbalance problem which can be beckoned as the data source of this paper.

Every data in the data set does not only show whether consumers respond to class label or not but also includes 85 properties reflecting relevant information of social demography characteristics as well as warranty policy showing in table 1.

The common assessment standard in classification problem is simple accuracy:

$$Simple\ Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \qquad (3)$$

The meaning of TP, TN, FP and FN shows in table 2.

Simple precision can only measure the overall performance of common classifiers which cannot effectively reflect the classification performance of unbalanced data set. In fact, if the imbalance degree is high we can still have a overall high rate of accuracy while carrying out random prediction. That is to say, the value of TN in the table is high and the classification effect on minority class is ignored while the assessment standard is ineffective. In fact, while facing the problem of imbalance class, people tend to predict accuracy of minority class. Therefore, concerning the imbalance problem in database marketing, this paper adopts Life Curve and Hit Rate as the assessment standard, which have been widely adopted in database marketing.

Hit Rate is one of the important standards to assess the data classification with imbalance class problem and its formula is shown as follows:

$$Hit\ Rate = \frac{TP_i}{depth_i \cdot M} \qquad (4)$$

```
Input：Set S of the Sample with N datas $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$
Step1: Set $r = |G_{neg}| / |G_{pos}|$, $K = \lceil r \rceil$;
Step 2: K-means clustering is performed on the data sample set $G_{neg}$,
and K data clusters are obtained :
        $subG_{neg}^1, subG_{neg}^2, \cdots, subG_{neg}^K$;
Step 3：K sample set of samples are obtained by the combination of
$subG_{neg}^k, (k = 1, 2, \cdots, K)$ and $G_{pos}$:
        $subG^1, subG^2, \cdots, subG^K$
Output
```

Figure 1. Supervised clustering process

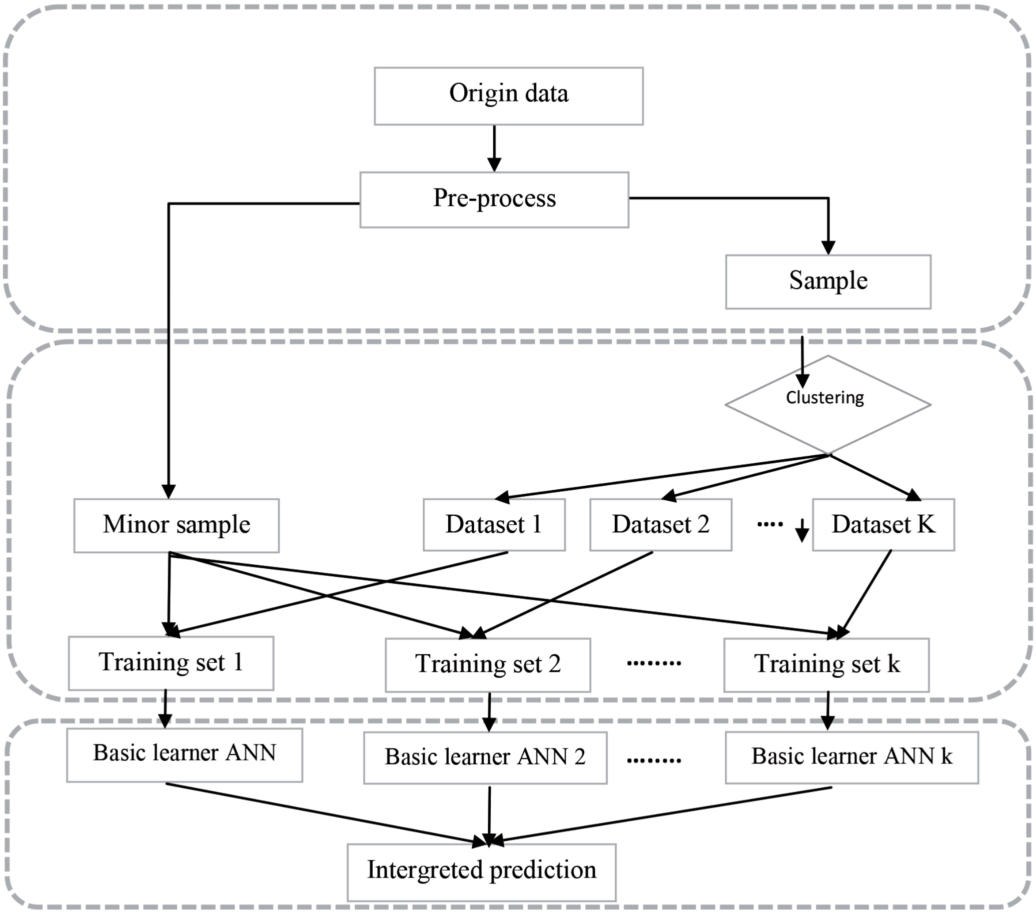

Figure 2. Database marketing based on ensemble learning

These assessment standard responses to the hit rate on customers of advertisement. In formula (5), $depth_i$ is the depth, which refers to the percentage of target samples to total samples. Commonly five quintiles are selected such as 5%, 10%, 15%,…, 100%. TP_i refers to the corresponding quantity affected by $depth_i$ while M is the number of total samples. Obviously, the higher hit rate is, the better the learning method will be.

Lift Curve adopts Cumulative hit rate with corresponding data under different depths as the evaluation standard and the formula shows as follows:

$$Cumulative\ hit\ rate_i = \frac{TP_i}{N} \tag{5}$$

In the formula, TP_i refers to the corresponding quantity affected by $depth_i$,N refers to the total number of corresponding quantity. In the coordinate system, lateral axis means depth $depth_i$ while direct-axis refers to $Cumulative\ hit\ rate_i$ which can elaborate the percentage of corresponding quantity to total sample under different depth learning. Obviously, the quicker speed of curve rising, namely the smaller depth is, the higher rate we shall get and the learning method is better. That is to say, we are able to have more corresponding customers with less advertisement.

Table 1. Dataset Attribute Description

ID	Description
1	The number of houses for a household
2	Average family size
3	Average age of family members
4-13	Psychographic segmentation: the success of hedonism, the average family status, independent business, good life
14-17	Faith in the Catholic Church, Protestant, the other and the religion
18-21	Married, single, cohabitation and other relations
22-23	Have children?
24-26	Education level, high, medium, low.
27	Social status
28-32	Occupation for entrepreneurs, farmers, middle managers, skilled workers and unskilled workers
33-37	Residents classified: A, B1, B2, C and D
38-39	Home purchase or rent
40-42	The number of cars: 1, 2 or 0
43-44	Domestic or private medical services
45-50	Residents income less than 30000, 3-4.5 million, 4.5-7.5 million, 7.5-12.3 million, higher than 123000 (USD)
51	Residents purchasing power
52-72	The contribution of various types of insurance: third party companies, third party agriculture, automobiles, trucks, motorcycles, etc.
73-85	Family holding the same type of policy contribution ratio

Table 2 Confusion Matrix

	Divided into response class	Other
Actual response class	TP	FN
Other	FP	TN

This paper comes up with a supervised clustering ensemble learning model to compare with other four models, which are single ANN, ANN based on SMOTE algorithm, ANN based on FN algorithm and GA/ANN algorithm. The assessment standards are hit rate and life curve showing in table 3 and figure 3. The target of database marketing is to acquire higher Hit Rate with smaller depth and it is not meaningful to have excellent hit rate results with higher depth. Therefore, this paper only carries out comparison between 5%-50% depth.

The result of table 3 shows that the ensemble learning method based on supervised clustering can have higher hit rate with the depth from 5%-35% comparing to other four methods especially from the depth 5%-25%, the hit rate is 32.00%, 23.25%, 18.67%, 15.75 and 14.30% which have obvious prediction advantages over other approaches. To conclude, the ensemble learning method based on supervised clustering is able to have higher Hit Rate with smaller depth which can improve the database marketing efficiency.

Figure 3 compares the lift curve based on different approaches. Singe ANN approach has the slowest lift speed which demonstrates that it has poor ability to deal with the imbalance class problem in database marketing while ANN based on SMOTE, ANN based on FN and GA/ANN approaches can create better effect. The model put forward in this paper has obvious advantages from 5%-35% depth and is worse than

Table 3. Hit rates of different learning methods

	Depth (%)									
Model	5	10	15	20	25	30	35	40	45	50
Proposed method	32.00	23.25	18.67	15.75	14.30	12.50	11.00	10.50	9.94	9.25
SMOTE/ANN	20.00	17.25	14.33	11.63	10.00	9.67	9.21	8.94	8.67	8.30
FN /ANN	24.50	15.50	13.00	10.75	9.90	9.33	8.43	7.88	7.40	7.00
GA/ANN	23.04	20.06	16.40	14.42	13.13	12.04	10.97	10.54	10.00	9.64
ANN	8.5	9	8.17	7.38	7.5	7	6.79	6.94	6.67	6.7

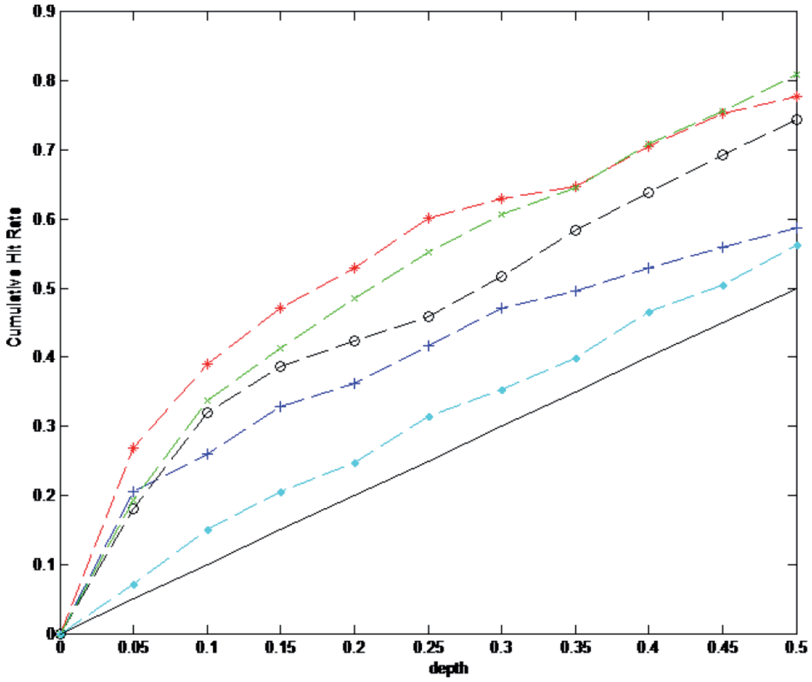

Figure 3. Lift curves of different learning methods

GA/ANN from 40%-50% depth. In reality, there is a dazzling array of customers in database marketing and therefore we pay more attention to the performance in smaller depth. In a summary, the above result indicates that the model in this paper can have higher hit rate in smaller depths which can be effectively affect the database marketing.

Conclusions

This paper comes up with a model which can deal with the imbalance class problem in database marketing based on supervised clustering and ensemble learning. This model believes that inside the data there are consumer aggregate of data with similar characteristics, so we adopt *K*-Means to carry out supervised clustering to conduct integration and classification with artificial neural network based on this. After carrying out experimental analysis and comparison among several learning methods, we can verify that the model in this paper is able to effectively deal with the imbalance class problem in database marketing and improve the accuracy. The following researches can be carried out from two aspects: firstly, extract the data characteristics to improve performance; secondly, carry out optimization as well as integration with other algorithms.

Acknowledgements

This work is supported by the Key Project Of National Natural Science Foundation, China (No. 71001112), Classification techniques considering feature space heterogeneity and their applications in business intelligence;

Key Project Of National Natural Science Foundation, China(No. 71471022), Ensemble learning models and methods for mining large-scale data streams

References

1. Baesens, B., Viaene, S., Van den Poel, D., Vanthienen, J., & Dedene, G. Bayesian neural network learning for repeat purchase modelling in direct marketing. European Journal of Operational Research, 2002, 138(1), 191-211.
2. Knott, A., Hayes, A., & Neslin, S. A. Next-product-to-buy models for cross selling applications. Journal of Interactive Marketing, 2002, 16(3), 59-75.
3. Weiss G M, Provost F J. Learning when training data are costly: the effect of class distribution on tree induction. Journal of Artificial Intelligence Research, 2003, 19: 315-354.
4. Kang P, Cho S, Mac Lachlan D L. Improved response modeling based on clustering, under-sampling, and ensemble. Expert Systems with Applications, 2012, 39(8): 6738-6753.
5. Ha K, Cho S, Mac Lachlan D. Response models based on bagging neural networks. Journal of Interactive Marketing, 2005, 19(1): 17-30.
6. Bose I, Chen X. Quantitative models for direct marketing: A review from systems perspective. European Journal of Operational Research, 2009, 195(1): 1-16.
7. Wilson D R, Martinez T R. Reduction techniques for instance-based learning algorithms. Machine learning, 2000, 38(3): 257-286.]

8. Guo H, Viktor H L. Learning from imbalanced data sets with boosting and data generation: the DataBoost-IM approach. ACM SIGKDD Explorations Newsletter, 2004, 6(1): 30-39.
9. Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2011, 16(1):321-357.
10. Xue Wei. An Improved SMOTE Algorithm for Re-Sampling Imbalanced Data Sets. Statistical Research, 2012(6):95-98
11. Rahman M M, Davis D N. Semi Supervised Under-Sampling: A Solution to the Class Imbalance Problem for Classification and Feature Selection. Transactions on Engineering Technologies. Springer Netherlands, 2014: 611-625.
12. Alejo R, García V, Pacheco-Sánchez J H. An Efficient Over-sampling Approach Based on Mean Square Error Back-propagation for Dealing with the Multi-class Imbalance Problem. Neural Processing Letters, 2014: 1-15.
13. Schapire R E. The strength of weak learnability. Machine learning, 1990, 5(2): 197-227.
14. Breiman L, Friedman J, Stone C J, et al. Classification and regression trees. CRC press, 1984.
15. Debray T. Classification in imbalanced datasets. Faculty of Humanities and Sciences, Maastricht University, 2009.
16. Breiman L. Bagging predictors. Machine learning, 1996, 24(2): 123-140.
17. Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting. Knowledge Discovery in Databases: PKDD 2003. Springer Berlin Heidelberg, 2003: 107-119.
18. Maclin R, Opitz D. Popular ensemble methods: An empirical study. Journal Of Artificial Intelligence Research, 2011,11: 169-198,
19. Sun Z, Song Q, Zhu X, et al. A novel ensemble method for classifying imbalanced data. Pattern Recognition, 2014, 48(5):1623-1637.
20. Zhai Yun, Yang Bing-ru, Qu Wu. Study on source of classification in imbalanced datasets based on new ensemble classifier. Systems Engineering and Electronics, 2011, 33(1): 196-0201
21. Deng Xiaoyi, Jin Chun, Higuchi Yoshiyuki, Han Jim. KSP: A Hybrid Clustering Algorithm for Customer Segmentation in Mobile E-commerce. Journal of Management Science, 2011, 24(4): 54-61
22. Huang Z. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining, Huang Z .DMKD. 1997.
23. MacQueen J. Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967, 1(14): 281-297.