Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

The Study on Feature Selection in Customer Churn Prediction Modeling

Yin Wu
School of Economic and Management
Beijing University of Posts and Telecommunications
Beijing, China
wingmaggie@gmail.com

Jiayin Qi
School of Economic and Management
Beijing University of Posts and Telecommunications
Beijing, China
Department of Information Systems and Operations Management
University of Washington Business School
Seattle, WA 98195-3200
qijiayin@gmail.com

Chen Wang
IBM China Research Laboratory
IBM
Beijing, P.R. China
wangcwc@cn.ibm.com

***Abstract*—When the customer churn prediction model is built, a large number of features bring heavy burdens to the model and even decrease the accuracy. This paper is aimed to review the feature selection, to compare the algorithms from different fields and to design a framework of feature selection for customer churn prediction. Based on the framework, the author experiment on the structured module with some telecom operator's marketing data to verify the efficiency of the feature selection framework.**

***Keywords*—Customer Churn Prediction, Feature Selection, Framework Design, Algorithm Experiment**

I. INTRODUCTION

Before the customer churn prediction being modeled, it is vital to choose the features as the input of the model that the company can calculate the probability of churning and take some measures to retain customers. [1] [2] The features consist of demographic data and billing data, the number of which always reaches hundreds. However, some of the features are not highly related to the churn. Therefore, the large feature set will not only increase the complexity of the model, but also lead to the machine running slowly and even declining in accuracy.

In terms of retaining the accuracy of churn prediction, the process of reducing the number of the features is called feature selection. Feature selection has been studied in different fields such as machine learning, pattern recognition and text categorization. It is reasonable to take those algorithms and methodologies as the foundation of feature selection in the customer churn prediction. Furthermore, the customers' feature can be divided into structured data type and text type, both of which require specific algorithms to process. Hence, the paper designs a framework of feature selection, considering the context of customer churn prediction, and then conducts empirical research using some telecom operator's marketing data.

II. REVIEW OF FEATURE SELECTION

Feature selection has been researched in the areas of machine learning, pattern recognition and text categorization for a long time and has relatively mature algorithms. Based on the relevance of customer churn prediction and those fields, it is essential to study present works as the first step of finding a proper approach and framework for feature selection in customer churn prediction.

A. Feature Selection in Machine Learning

The purpose of feature selection is searching for the smallest variable set which satisfies some criterion [3]. Feature selection algorithm can be categorized into two groups, which are filter and wrapper models, based on their dependency on learning algorithms [4] [5] [6].

Filter model exploits the criterion irrelevant to the learning algorithm to deal with data, which means to take feature selection as a part of the preprocessing procedure without regards to the affect of machine learning. Filter model is general and easily carried out, while the shortage of that is the probability of decreasing the learning precision and less objectivity of evaluation criterions. FOCUS, B&B and ABB are the familiar algorithms.

Wrapper model utilizes the precision of a predetermined learning algorithm to evaluate the feature subset, which involves feature selection in the learning algorithm. Wrapper model has the advantage of ensuring the following machine learning with good performance, but at the expense of computations especially with the large feature set. The frequent algorithms include GA, LVW, SA and so on.

B. Feature Selection in Text Categorization

The primary problem in the text categorization is the high dimension of features. Since the traditional classification techniques can not deal with the characteristics of mass features, the dimensionality reduction techniques [7] were proposed. The dimensionality reduction techniques can be

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

divided into two categories. One is feature selection, the other is feature reconstruction. Feature selection chooses the specific assessment function to score terms successively, sorts them ascending and select the tops of them. This method does not change the nature of the original terms space. Assessment functions such as document frequency, information gain and mutual information are commonly used Feature reconstruction transforms the original feature set into a new one with constructing an assessment function to project measuring space into the feature space and extract the highest scored features. The dominating methods contain term clustering and latent semantic index.

C. *Feature Selection in Pattern Recognition*

In the area of pattern recognition, two concepts of feature extraction and feature selection are proposed, both of which are finding the most representative features of the full set [8], but with different methods. Feature extraction maps high-dimensional space to the low-dimensional space as a result of the secondary features. They are some combination, usually a linear combination of original features. Feature selection selects the most representative features to reduce the dimension of sample space. The research falls two parts. One focuses on quantitatively measuring how the features affect classification with methods such as the distance between categories within the category and the probability distribution. The other is aimed to find a better subset in the satisfactory time. Related algorithms include the optimal search algorithm, suboptimal search algorithm and so on.

There are also new algorithms emerging in pattern recognition such as simulated annealing algorithm [9] and genetic algorithm. In practical applications, feature extraction and feature selection are not distinct from each other, but can be combined together to reduce dimensionality more than once.

D. *Feature Selection in Statistics*

Feature selection in statistics can be categorized into three dimensions. The first dimension is to retain all the information of variables, transform the original individual variables, and generate comprehensive indicators. In the statistics, massive variables are observed, analyzed and modeled. Simply reducing the variables easily loses massive information and even draws the wrong conclusion. Therefore, the individual variables are replaced by the comprehensive indicators by factor analysis, correspondence analysis and optimal scaling. The second one is to select features suited to the statistics model. For example, multivariate regression model sets up the test to examine the multicollinearity. This type of feature selection can ensure the accuracy and efficiency of the model. The third dimension is using the individual test to select features, which can not only calculate the correlation between each feature and the target variable such as PCA and also compute relevance between features as correlation coefficient.

E. *Development Trends of Feature Selection*

One of tendencies is the combination and improvement of various algorithms. For example, minimum solution tree [10], FBB [11] and BBPP [12] are all improved from Branch & Bound algorithm; floating search algorithm is the combination of SFS and SBS and the improvement of L-R algorithm. It is also popular in text categorization as the improved mutual information [13] and ROC algorithm [14].

Furthermore, some new techniques in the knowledge discovery are introduced to enrich the feature selection. Rough set [15] [16] and support vector machine [17] successfully selected features in recent studies.

F. *The Comparison and Contrast of Feature Selection*

Feature selections from four fields share similarities but they are applied to different scopes. It can be concluded in Table Ⅰ.

TABLE I. THE COMPARISON OF ALGORITHMS OF FOUR FIELD

Fields	**Objectives**	**Advantage**	**Disadvantage**	**Application**
Machine Learning	Structured Data	Study frame relatively mature; Massive algorithms	Hard to compare; Messy algorithms	Data preprocess in data mining
Text categorization	Text Data	Targeted	The effect need improving	Website Category Personalized Recommendation
Pattern Recogntion	Image & Signal	Good performance on handling high dimensionality	Complex algorithms	Image segmentation Signal processing
Statistics	Mainly structured data, also text data	Simple Good performance	Some algorithms are difficult to promote	Remove multicollinearity in regression model

First of all, feature selection in machine learning has a sophisticated research framework and is mainly used to eliminate redundant information in structured data. Feature selection in text categorization is based on the document and term frequency with the application in webpage classification. Pattern recognition applies the complex algorithms to the image segment and signal processing instead of ordinary data. Simply copying the algorithm will only make prediction model complex, so it is better to combine it with machine learning approaches. Feature selection in statistics primarily handles structured data but it is also introduced in text categorization as a supplement.

III. FRAMEWORK DESIGN OF FEATURE SELECTION

A. *The Status of Feature Selection and the Current Solutions*

Although the study of feature selection is involved in many fields and has developed many algorithms, three reasons lead the research to the bottleneck as follows.

Firstly, each algorithm is only effective to a certain data type with less scalability. The efficiency and accuracy will decline significantly if it is applied to the unsuitable data type. Secondly, feature selection is an NP-hard problem which means there is not a best algorithm. For example, complete search algorithm could obtain the optimal solution at the expense of much time; heuristic algorithm will risk losing the optimal results but consume fewer resources. Thus the combination of algorithms has become an urging issue. Lastly, different algorithms are relatively independent and separate, unable to conduct the comparison and selection. This situation offers the company low possibility of free choosing algorithms, leading to the less flexibility of the whole selection process.

Having noted the existing problem of feature selection in the study, domestic and foreign experts put more emphasis on the integration of the algorithms in applications than the development and improvement of algorithms. Huan Liu and Lei Yu [18] proposed a unifying platform as an intermediate step to integrate existing feature selection algorithms both in classification and clustering, taking advantage of individual algorithms. Jia Shi, SHARDROM Johnson and Zhang Wu [19] proposed a feature selection library based on strategy-pattern (FSLS) to encapsulate many popular feature selection algorithms under unified interfaces, while different strategies of one algorithm could be exchanged conveniently.

Drawing on the achievements of previous studies, this paper presents a feature selection framework for customer churn prediction model, combining a variety of feature selection algorithm to achieve the purpose of practical application.

B. The Principle of Feature Selection Framework Design

The principle of the framework is proposed as follows.

- Different data dealt separately

For different data types, use different feature selection algorithm targeted.

- Easy operating and algorithms encapsulated

After classifying algorithms in the library, connect and choose the appropriate algorithm in accordance with the needs. This operation can achieve transparence to the user without concerning how algorithms realize.

- Divide into phases with clear borders

As the principle from simple to complex, start with a simple algorithm to remove a collection of irrelevant features and then use more complicated algorithms to make further selections.

- Dynamic management of algorithms library

The algorithm library is not static, but updated according to the performance of algorithms, which performed badly would be deleted after a period of time.

C. Framework of Feature Selection

Based on the principle, the frame should be divided into two modules, that is structured module and text module, to separate different data types. In addition, as the principle from easy to difficult, each module is divided into two phrases to select features.

- Structured Module

The first phrase refers to the core idea of Filter model, evaluating features without considering learning algorithms. All algorithms conforming to the characteristics could enter the library, operating at either individual feature or feature subset. The second phrase uses the Wrapper model to evaluate features by learning accuracies.

- Text Module

The first phrase primarily scores terms by evaluating functions and selects the terms which exceed the threshold to the next phrase. Except the frequently used algorithms in the text categorization, the statistics approaches could be included in the library. In the second phrase, construct an assessment function to project original variables into the feature space and extract the highest scored features.

The feature selection framework of customer churn prediction can be described as Figure 1.

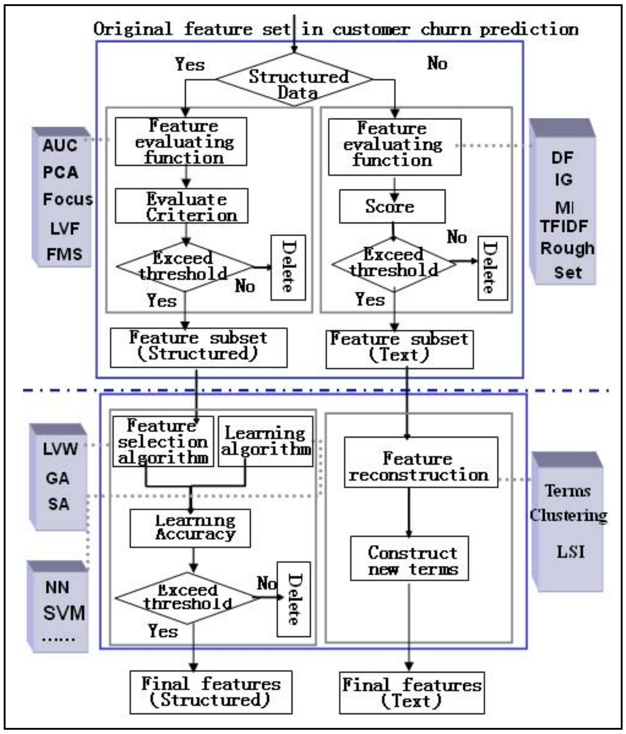

Figure 1. Feature Selection Framework of Customer Churn Prediction)

The process of the feature selection is described as follows.

Step 1: Input the original features of the customer churn prediction, judge these features whether belong to structured data or text data, and then select them to enter the corresponding module.

Step2a: If some features belong to the structured module, start with the first-phrase feature selection. Select the feature subset or individual feature using the evaluating function, deleting features beyond the stopping criterion and obtain the feature subset in the first phrase.

Step 3a: The selected features after the first phrase are chosen by the wrapper model, trained by the commonly used learning algorithms. The indicator such as accuracy and recall is used as the evaluating criterion which determines features whether stay in the final feature set. The final features are the optimal solution through the framework.

Step 2b: If some features enter the text module, the terms are scored by the feature assessment function, and saved ones whose scores are higher than the threshold. The feature selection process stops when it meets a stopping criterion.

Step 3b: The relatively low-dimensional terms after the first phrase are entering the second stage of reconstructing, which re-project them to comprehensive new items. The final features are the optimal solution through the framework.

IV. EXPERIMENT ON STRUCTURED MODULE

Referring to the feature selection framework, the author experiment on the structured module with some telecom operator's marketing data. Combined literatures with present applications, the experiments were executed in the first phase with PCA and SFFS algorithm, and the second phase with LVW and GA algorithm.

Through statistical sampling, 300 customers records are selected for experiments, with the proportion of 2 to 1 between exist and churned customers. 206 features include basic information, other information extracted from billings which consists of absolute, relative and volatility indicators.

A. Algorithms on the First Phrase

1) Principle Component Analysis(PCA)

Principle component analysis simplifies the analysis of the complex relationship interrelated variables. In the research of customer churn prediction, PCA reduces the high-dimensional space to generate comprehensive features at a minimum loss of information. The process of PCA is shown as Figure 2.

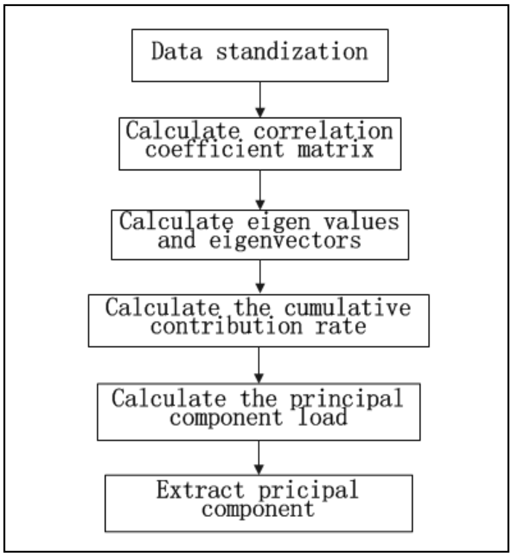
Data standization
Calculate correlation coefficient matrix
Calculate eigen values and eigenvectors
Calculate the cumulative contribution rate
Calculate the principal component load
Extract pricipal component

Figure 2. Principle Component Analysis Algorithm

Setted the cumulative contribution rate to 98 percent as the threshold, 34 principal components are gained. Furthermore, the components which are complicated and non-correlated could be named separately as new meaningful variables.

2) Sequential Forward Floating Searching

Floating searching method which combined SFS and SBS algorithms was first proposed by P.Pudil., Novovieova and J.Kittle in 1994[20]. SFFS procedure is applied after each forward step a number of backward steps as long as the resulting subsets are better than the previously evaluated ones.

SFFS must be provided for the number of forward selection and the significant function to determine the search direction. The algorithm is described as Figure 3.

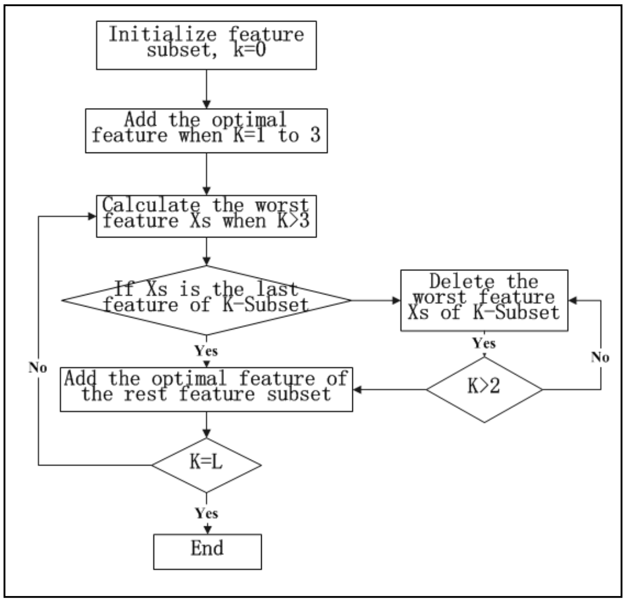
Initialize feature subset, k=0
Add the optimal feature when K=1 to 3
Calculate the worst feature Xs when K>3
If Xs is the last feature of K-Subset
Delete the worst feature Xs of K-Subset
Yes
Yes
No
No
Add the optimal feature of the rest feature subset
K>2
K=L
Yes
End

Figure 3. Sequential Forward Floating Search Algorithm

After the records being imported and processed, 50 features were selected from the original 206 features.

B. Algorithms on the Second Phrase

1) LVW Algorithm

LVW was proposed by Huan Liu and Rudy Setiono [21]. The search direction and strategy are random so as to bring great uncertainty. If it repeats enough times, the optimal solution must be gained. The stopping criterion is the predefined loops. The LVW algorithm is displayed as Figure 4.

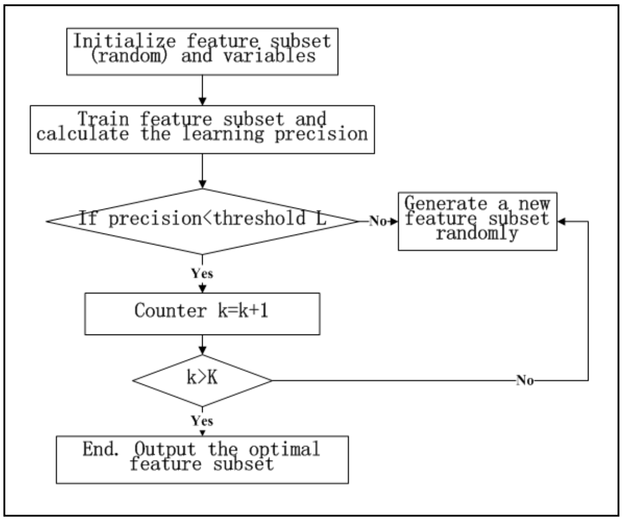
Initialize feature subset (random) and variables
Train feature subset and calculate the learning precision
If precision<threshold L
No
Generate a new feature subset randomly
Yes
Counter k=k+1
k>K
No
Yes
End. Output the optimal feature subset

Figure 4. LVW *Algorithm*

After records being processed, 44 features were selected out of the original 50 features which were outputted from SFFS.

2) Genetic Algorithm

Different from traditional search algorithm, genetic algorithm implements genetic manipulation to achieve individual and group restructuring process by the iterative search method based on the fitness function, which were applied in many areas [22] [23]. The main operations of GA include selection, crossover and mutation. The GA is shown as Figure 5.

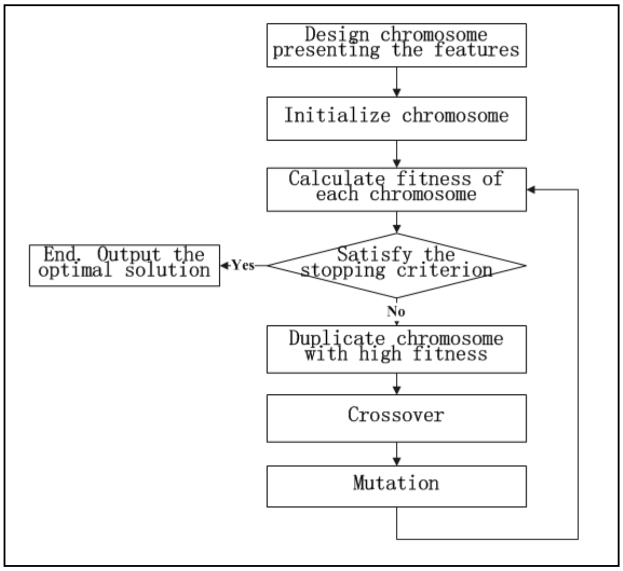
Design chromosome presenting the features
Initialize chromosome
Calculate fitness of each chromosome
Satisfy the stopping criterion
Yes
End. Output the optimal solution
No
Duplicate chromosome with high fitness
Crossover
Mutation

Figure 5. Genetic Algorithm

Through the genetic algorithm, 26 features were selected out 50 features which were outputted from SFFS.

V. Experiment Result and Comparison

Modeling the customer churn prediction with different feature subsets after feature selection, evaluate and compare the result of model as the further study of feature selection algorithms comparison.

A. Model with Small Samples

First of all, four feature subsets are entered the model at the size of 300, and compared with evaluating criterions which include root ASE and misclassification rate.

TABLE II. Root ASE and Misclassification in Small Sample

Algorithm	Validation Set: Root ASE	Validation Set: Misclassification Rate
PCA	0.289716	0.08889
SFFS	0.372479	0.16667
GA	0.350836	0.12222
LVW	0.360833	0.14444

As shown in Table Ⅱ, PCA performed best referring to the root ASE of validation sets with 0.289716 and the misclassification rate with 0.08889 among four algorithms, followed by GA whose root ASE of validation set was 0.350836 and the misclassification rate of validation set with 0.12222. LVW and SFFS ranked three and four respectively.

B. Model with Small Samples

Considered the speed of running programs, only 300 samples were selected in feature selection. However, decision trees modeled with 300 samples could not represent the advantage of data mining and the result of evaluating would be inaccurate. Hence, expending the size of samples to 12556 as the input maybe bring out a new insight with evaluating criterions which include root ASE and misclassification rate.

TABLE III. Root ASE and Misclassification in LARGE Sample

Algorithm	Validation Set: Root ASE	Validation Set: Misclassification
SFFS	0.154079	0.03
GA	0.148027	0.02734
LVW	0.153092	0.02947

As displayed in Table Ⅲ, GA performed best referring to the root ASE of validation set with 0.148027 and the misclassification rate of validation set with 0.02734 among four algorithms, followed by LVW whose root ASE of validation set was 0.153092 and the misclassification rate of validation set with 0.02734. SFFS ranked last.

TABLE IV. Comparison with Small and Large Samples

Algorithm	30% Response (Small)	30% Response (Large)	30% Lift Value (Small)	30% Lift Value (Large)
SFFS	69%	77%	2.21	2.59
LVW	71%	78%	2.38	2.6
GA	78%	79%	2.58	2.61

As described in Table Ⅳ, it is delighted to find all the indicators increasing after the scale of samples expanded. For example, the feature subset selected by SFFS algorithm performed better with the 30% response rapidly climbing from 69% to 77% and the lift value ascending from 2.21 to 2.59.

C. *Experiment Summary*

Compared different feature subsets selected by four algorithms both in small and large samples, the accuracy of customer churn prediction model with the two-phrase selection performed better than just once, indicating two-phrase selection is necessary. Furthermore, with expansion of scales of samples, the effect of modeling after two-phrase selection was promoted rapidly, which showed practical significances.

VI. Conclusion

In this paper, starting with the comparison of feature selection from four different fields, the authors proposed a framework to satisfy different data types and carried out experiments on structured module to validate the efficiency of the two-phrase feature selection. Future work could be researched on the text module and framework improvement.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Project No.: 70701005), the Specialized Research Fund for the Doctoral Program of Higher Education of the People's Republic of China (Project No.: 20070013014), and the Open Research Fund between Beijing University of Posts and Telecommunications and IBM.

References

[1] Jiayin Qi, Li Zhang, et al. ADTreesLogit model for customer churn prediction. Annual of operation research, 2009, 168(1): 247-265.

[2] Yangming Zhang, Jiayin Qi, Huaying Shu, Jiantong Cao. A hybrid KNN-LR classifier and its application in customer churn prediction. 2007 IEEE International Conference on Systems, Man and Cybernetics , Oct. 7-10, 2007,3265-3269.

[3] Jin Zhang, D. "Research on Rough Set Theory Based Data Mining Algorithm.," Unpublished doctoral dissertation, Northwestern Polyteehnieal University, China. December, 2005.

[4] R.Kohavi and B.Frasea.. "Useful Feature Subsets and Rough set reducts," International Workshop on Rough Sets and Soft Computing (RSSC), pp. 310-317, 1994.

[5] Hong-Xing Li, Li D Xu, Jia-Yin Wang and Zhi-Wen Mo, Feature space theory in data mining: transformations between extensions and intensions in knowledge representation, Expert Systems, 20(2), 2003, 60-71.

[6] Hong-Xing Li and Li D Xu, Feature space theory—a mathematical foundation for data mining, Knowledge-Based Systems, 14(5-6), 2001, 253-257.

[7] Jianjun Sun, Yin Chen and ect. Information Retrival Technology, Science Publisher: China, 2004.

[8] Zhaoqi Bian , Xuegong Zhang and ect, 2rd ed., Tshinghua University Publisher: China. January 2000, pp. 176-210.

[9] Keqi Wang, Hui Wang, XuebingBai. Feature Selection Based on Simulated Annealing Algorithm and the Recognition Rate of the Nearest Neighbor Classifier. Techniques of Automation and Application. Vol.26, No.1, Jun.2006. pp: 27-29.

[10] B.Yu, B.Yuan.A More Efficient Branch and Bound Algorithm for Feature Seleetion. Pattem Recognition. Vol26.1993, pp. 883-889.

[11] P.Somol, P.Pudil, F.J.Ferri and etc. Fast Braneh&Bound Algorithm for Feature Selection. Invited Paper for the 4th World Multiconference on Systemies, Cybernetics and Informaties, Proeeedings. Orlando, Florida, 2000,pp. 646-651.

[12] P.Somol, P.Pudil, J.Grim. Braneh&Bound Algorithm with Partial Predietion for Use with Recursive and Non-Reeursive Criterion Forms. Rio de Janeiro.Int.Conf On Advanees In Pattem Recognition. 2001.

[13] P.Somol, P.Pudil and etc. Adaptive Floating Search Methods in Feature Selection. Pattern Recognition Letters. 20(11-13): December 1999, pp. 1157-1163.

[14] Zhi Pei , Zhigang Li and etc. Based on a Kind of Method to Categorize Text of Improve the Mutual Information. Journal of Inner Mongolia University for Nationalities. Vol.22, No.4, Aug.2007, pp. 377-380.

[15] Fangtao Li, Tao Guan, Xian Zhang. An Aggressive Feature Selection Method based on Rough Set Theory. Innovative Computing, Information and Control, 2007. ICICIC '07. Second International Conference. Sept. 2007, pp. 176 -179.

[16] Hengshan Wu, Feng Yan, Xiaobing Pei, Li Liu. A Feature Selection Algorithm Based on Rough Set Theory and Its Applications. Computer Engineering and Applications. Vol.42, No.16, Aug.2006, pp. 175-176, 221.

[17] Taeshik Shon, Yongdue Kim, Cheolwon Lee. A Machine Learning Framework for Network Anomaly Detection Using SVM and GA. Information Assurance Workshop, IAW '05. Proceedings from the Sixth Annual IEEE SMC. June 2005, pp. 176 -183.

[18] Huan Liu, Lei Yu. Toward Integrating Feature Selection Algorithms for Classification and Clustering. IEEE Translation on Knowledge and Data Engineer. Vol. 17 (No. 4). APRIL 2005, pp. 491-502.

[19] Jia Shi, SHARDROM Johnson, Zhang Wu. Design of feature selection library based on strategy - pattern (FSLS). Computer Engineering and Applications, 2007, 43(1) , pp. 181- 184,197.

[20] P. Pudil, F.J. Ferri, J. Novovicova. Floating Search Methods for Feature Selection with Nammonotonic Criterion Functions. Pattern Recognition Letters. Vol. 15. November 1994, pp. 1119-1125.

[21] Huan Liu, Rudy Setiono. Feature Selection and Classification-a Probabilistic Wrapper Approach. In. T. Tanaka, S. Ohsuga, and M. Ali. Proc. Ninth Int'l Conf. Industrial and Eng. Applications of AI and ES. eds. 1996, pp. 419-424.

[22] Chaudhry, Sohail S. , Varano, Michael W. , Xu, Lida. Systems Research, Genetic Algorithms, and Information Systems, Systems Research and Behavioral Science, 17, 149-162,2000.

[23] Yanxia Jiang, Lida Xu, Huacheng Wang, Hui Wang. Influencing factors for predicting financial performance based on genetic algorithms, Systems Research and Behavioral Science, published online: Mar 9 2009.