Enhancing Decision Tree Performance in Credit Risk Classification and Prediction

Raymond Sunardi Oetama
Information System Department, University of Multimedia Nusantara, Tangerang, Indonesia
Raymond@umn.ac.id

Diterima 15 Mei 2015
Disetujui 10 Juni 2015

Abstract - This study is focused on enhancing Decision Tree on its capabilities in classification as well as prediction. The capability of decision tree algorithm in classification outperforms its capability in prediction. The classification quality will be enhanced when it works with resampling techniques such as Adaboost.

***Index Terms*—Data Mining, Decision Tree, Resampling, Credit Analyst.**

I. INTRODUCTION

Credit loans are one of main business of banking industry. The good performance of credit loan business will keep profitability and stability level of a bank [1]. Before a customer has an approval of the credit loan, an assessment process is needed to make sure the customer has fulfilled all required documents. Afterwards, the customers' financial background and history shown on the documents will be analyzed.

Credit risk is the most critical challenge for bank management [1]. There are two final decisions at the end, which are to approve or to reject the loan proposal. The decision may be correct or incorrect. If the decision is correct, the loan proposals from good customers are approved (true postives) and proposals from bad customers are rejected (true negatives). On the other hand, if the decision is false, loan proposals from good customers are rejected (false negatives) or loan proposals from some bad customers are approved (false positives). Instead of doing the analysis manually, today banks can utilize a computer application named credit scoring. By applying a good performance credit scoring application, the bank managers will be supported with some suggestions which are the results of the credit loan analysis processes. By using such information, the quality of the decision will become more accurate.

However, credit scoring is not an excellent tool. In fact, still many customers are not able to pay back their loans. Consequently, these problems cause a significant decrease of profitability of a bank. Without any curative or preventive solutions from the bank, the amount of unpaid loan will fly till the bank is collapse. On the other hand, the manual approval cannot handle the entire loan proposals as a big number of loan proposals come both online and offline sources. Consequently, the bank still needs a help from credit scoring application to speed up the approval processes. Therefore, the quality of the current credit scoring application must be enhanced or at least it must be kept stable at a certain level of quality.

Credit scoring is a group of models that are built using a single or combination of techniques to support lenders for approving credit to customers [2] The objective of a credit scoring model is to analyze the capacity of a customer to repay back the amount of money that has been borrowed [3]. According to Akkoc [4], credit scoring can be seen as a system and the output is the loan proposals will be classified as "good" since they have a high probability to repay the loans from the bank and "bad" if there is a small probability that the borrowed money will be paid back by the customers. The applicants' characteristics are utilized as the input to build the credit scoring system.

A big number of studies have utilized the application of a a single based algorithm to build the credit scoring models. Some popular algorithms are ANN (Artificial Neural Network) [5]. The credit scoring model is shown as a black box to distinguish good customers and bad customers by capturing non linearity in financial data. Other based algorithms are logistic regression and decision tree [6]. The strong ability of

the Logistic regression technique is its ability to decide the binary outputs such as good as 1 and bad as 0. Whilst the strong point of a decision tree is to measure the probability of decisions, and to give weights to the input variables of the credit scoring models. The other studies shows that it is also important to apply some database approach techniques [7] such as clustering, resampled techniques (bagging or boosting).

In fact, there is no a single based algorithm which is the best for all datasets. As a result, many studies try to compare some based algorithms to find which algorithm is the best for their datasets [8]. Moreover, many algorithms have been modified. So there are many versions for any based algorithm developed from other studies. Therefore, rather than trying to find the best algorithm, this study is started with picking up a popular based algorithm for decision making which is decision tree [9], and afterwards, focused on how to improve the performance of this algorithm in classifying and predicting good and bad customers.

II. EXPERIMENT SETTING

A. Datasets and Algorithms

An experiment is conducted to build a credit scoring model through Weka version 3.6.12 [10] using UCI German Credit Data [11]. German Credit data contains 1000 instances including 700 good customers and 300 Bad customers. Each record has 20 attributes which are Status of existing checking account, Duration in month, Credit history, Purpose, Credit amount, Savings account/bonds, Present employment since, Installment rate in percentage of disposable income, Personal status and sex, Other debtors/guarantors, Present residence since, Property, Age in years, Other installment plans, Housing, Number of existing credits at this bank, Job, Number of people being liable to provide maintenance for, Telephone, and foreign worker.

The only algorithm that is used here is Decision Tree version C4.5 in Weka software is named as J48. There are three options here which are J48 works alone, the combination of J48 and Bagging, and the combination of J48 and Boosting. Bagging and Boosting are applied to grow trees in another ways to see which combinations are the best to build the model. Samples are taken from the data training and then will be used to grow trees by using J48. In Bagging techniques, data is sampled by using bootstrap sampling techniques whilst in Boosting techniques data is sampled by using Adaboost sampling techniques [12].

B. Analysis

The analysis is divided into two purposes, which are the performance of decision tree to classify the characteristics of good customers as well as bad customers. The entire data will be utilized as data training. Secondly, the performance of decision tree to predict who are good customers and who are bad customers. For this purpose, data is split into two parts. The decision tree will be trained on the first part of data to build the prediction model. The first part of data represent the current customers that have already taken the loan. Afterwards, the model will be tested on the other parts of data as the new group of customers.

A various techniques of splitting data are applied. Firstly, data is divided into 10 folds randomly. Each fold contains 10% of data. The first 10% data will be taken as data testing and the rest will be used as data training. Afterwards, the next fold will be chosen as data testing and the rest will be used as data training and to be continued till the last fold has been utilized. Secondly, data is randomly split into 10% data for data training and 90% data for data testing. Afterward, data is split randomly into 20%-80% and to be continued till the split is 90%-10% randomly.

Accuracy is very popular in the machine learning research to figure out the performance of the model. It shows how good the decision tree performs in classifying good and bad customers as well as how good it predicts correctly good customers as good customers and bad customers as bad customers. This rate is made from the confused matrix which is the total instances of good customers that is classified correctly as good customers plus the total instances of bad customers that is classified correctly as bad customers divided by the entire instances. The accuracy rate will be applied for both purposes.

III. DISCUSSION

The different approaches will result different performances even though the algorithm and the data sets are the same. J48 is examined through three ways, which are working alone, accompanied with some resampling techniques. Data approaches are also examined by using some splitting techniques.

A. Pure J48 Performance

As can be seen on figure 1, in general, the decision tree performances to classify good and bad customer

characteristics are better than to predict good and bad customers. Its performances of making classification are between 80% to 100% whilst its performance to make prediction decreases to around 60% to 80%.

It is harder to predict good and bad customers because there are some characteristics occur in data testing but not shown on data training. As a result, some prediction results are incorrect. In average, the classification performance shows 20% to 40% better than predicting good and bad customers.

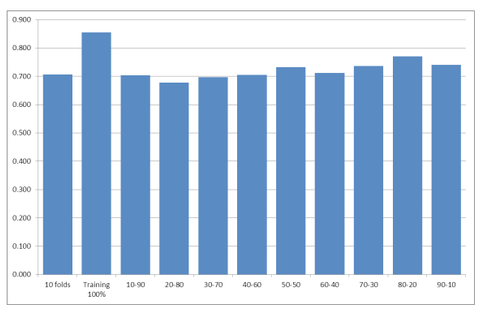

Figure 1: The comparison of J48 Performance using some database approaches (in Accuracy Rate).

B. Resampling Techniques

Afterwards, J48 will combined with some resampling methods such as Bagging and Boosting. The results can be seen on figure 3 and 4.

The maximum performance is shown by using 100% data training with Adaboost technique (). This means the entire good and bad customers' characteristics are classified correctly. However, it will be the best only for this time, because some new combination of characters is not readily expected.

Given a set of data $(a_1,b_1),\ldots,(a_m,b_m)$ where $a_i \in A$, $b_i \in \{-1,+1\}$

Start with $D_1(i)=1/k$ for i-1,…,k.

For t=1,…,T:

Train weak learner using Distribution D_t.

Get weak hypothesis $h_t: A \rightarrow \{-1,+1\}$

Goal: select h_t with small weighted error:

$\varepsilon_t = Pr_{t \sim Dt}[h_t(a_i) \neq b_i)]$

Take $\beta_t = \frac{1}{2} \ln(\frac{1-\varepsilon_t}{\varepsilon_t})$

Calculate for i=1,…,m: $D_{t+1}(i) = \frac{D_t(i)\exp(-\beta_t b_i h_t(a_i))}{Z_t}$

where Z_t is a normalization factor (selected so that D_{t+1} will be a distribution)

Output: $H(a) = \text{sign}(\sum_{t=1}^{T} \beta_t h_t(a))$

Figure 2: The Adaboost Algorithm

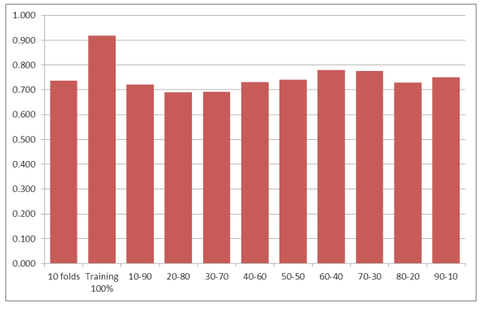

Figure 3: The comparison of J48 and Bagging Performance using some database approaches (in Accuracy Rate).

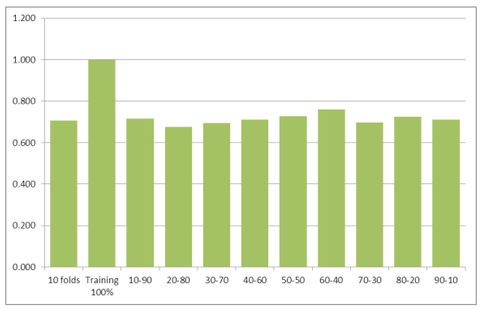

Figure 4: The comparison of J48 and Boosting Performance using some database approaches (in Accuracy Rate).

In comparison of the performances of resampling techniques, the decision tree performances mostly will be improved when bagging is applied, whilst the performance is fluctuated when Adaboost is utilized..

C. Splitting Techniques

In comparison of splitting techniques, the model which is built without splitting will outperform other models from both split and 10 fold techniques. However, the performance is slightly enhanced by applying Bagging Sampling Techniques because the quality of sample is improved by resampling randomly for several times, till the best result is occurred. The best performance is shown when Boosting is applied as the quality of sample is improved by repair its weights to create a better and better result till the best result occurs.

IV. CONCLUSION

To summarize, the decision tree performs better in identifying the characteristics of good and bad customers rather than estimating good and bad customers. To identify the characteristics of good and bad customers, data is not split. The Decision tree can reach 100% accuracy. This best performance occurs when decision tree is combined with Adaboost resampling technique on 100% training data.

Secondly, to estimate good and bad customers, data is split. The accuracy of decision tree to predict good customers is between 60% and 80%. It will remain a maximum 40% of new customers who will be predicted incorrectly.

Overall, the performance of decision tree will be improved when it is combined with other approaches. By resampling data with bagging techniques, overall its performance will be consistanly improved at most split data types. Whilst resampling data using Boosting technique will perfom perfectly when data is not split but it shows unstable performances on splitting data.

V. FUTURE WORK

There are two different directions here. First, the changes of customers' characteristics can be ignored as long as the entire characters of good and bad customers can be identified. If it is true, a clustering technique will be involved to separate all characteristics of good customers and bad customers. Secondly, if anticipation to the changes of customer behavior in the future is more important, then another algorithm such as random forest will be utilized to create more accurate models.

REFERENCES

[1] H. A. Bekhet and S. F. K. Eletter, "Credit risk assessment model for Jordanian commercial banks: Neuralscoring approach," *Review of Development Finance,* pp. 20-28, 2014.

[2] A. Heiat, "Comparing performance of data mining models for computer," *J. Int. Fin. Econ.,* vol. 12, no. 1, p. 78–83, 2012.

[3] A. Emel, M. Oral, Reisman and Y. R. A., "A credit scoring approach for the commercial banking sector.," *Socioecon. Plann. Sci.,* vol. 37, p. 103–123, 2003.

[4] S. Akkoc, "An empirical comparison of conventional techniques, neural networks and the three stage hybrid adaptive neuro fuzzy inference system (ANFIS) model for credit scoring analysis: the case of Turkish credit card data," *Eur. J. Operat. Res.*, vol. 222, pp. 168-178, 2012.

[5] S. Eletter, "Using data mining for an intelligent marketing campaign," . *Glob. Bus. Econ. Anthol,* vol. 2, pp. 276-282, 2012.

[6] H. Koh, W. Tan, H. and C. Goh, "A two-step method to construct credit scoring models with data mining techniques," *Int. J. Bus. Inform,* vol. 1, pp. 96-118, 2006.

[7] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, Wiley-Interscience, 2004.

[8] A. Ko, R. Sabourin, A. J. Britto and A. Britto, "From Dynamic Classifier Selection to Dynamic Ensemble Selection," *Pattern Recognition,* vol. 41, pp. 1718-1731, 2008.

[9] J. R. Quinlan, "Induction of decision trees," *Machine Learning,* vol. 1, no. 1, pp. 81-106, 1986.

[10] "Weka 3: Data Mining with Open Source," The University of Waikato , [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/. [Accessed 23 March 2015].

[11] D. Newman, S. Hettich, C. Blake and C. Merz, "UCI Repository of Machine Learning Databases," [Online]. [Accessed 30 March 2015].

[12] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm" In Proceedings of the 13th international conference on machine learning, San Francisco, USA, pp. 148-156

Rancang Bangun Piranti Lunak Pengelola Parameter Akuisisi Data Terowongan Angin Kecepatan Rendah Indonesia

Ivransa Zuhdi Pane
Unit Pelaksana Teknis Laboratorium Aero Gas-dinamika dan Getaran,
Badan Pengkajian dan Penerapan Teknologi, Tangerang Selatan, Indonesia
izpane@gmail.com

Diterima 03 Juni 2015
Disetujui 10 Juni 2015

***Abstract*—Data acquisition is an important part of a series of activities in a wind tunnel test and determine the validity of aerodynamic characteristics information of the test object. One of the factors which affect the success of the data acquisition process is the control of the data acquisition parameters prior to the execution of the wind tunnel test. A large number of data acquisition parameters, and the configuration complexities of the data acquisition parameters, which are still managed manually, urged the development of a software which is expected to facilitate the management of these parameters in a way that is friendly to use and integratable into the existing data acquisition system. Engineering of data acquisition parameters management software was then carried out through the analysis, design and implementation stages in an iterated manner, starting with a simple prototype toward the establisment of operational product.**

***Index Terms*—software engineering, data acquisition, wind tunnel test**

I. PENDAHULUAN

Pengujian terowongan angin merupakan rangkaian kegiatan pengukuran, akuisisi, pengolahan dan presentasi data yang dilaksanakan dengan tujuan untuk mengetahui karakteristik aerodinamika dari objek uji. Pengujian ini umumnya dilakukan di suatu fasilitas terowongan angin dengan hembusan angin melewati obyek uji yang ditempatkan di seksi uji. Obyek uji merupakan pemodelan terskala dari obyek sesungguhnya (umumnya disebut juga model uji), dan dapat berupa obyek aeronotik, seperti pesawat terbang atau bagiannya (sayap dan badan pesawat), maupun obyek non-aeronotik, seperti gedung atau jembatan. Terowongan Angin Kecepatan Rendah Indonesia (TAKRI) merupakan penyelenggara dan penyedia jasa pengujian terowongan angin di Indonesia, yang dikelola oleh Unit Pelaksana Teknis Laboratorium Aero Gas-dinamika dan Getaran (UPT LAGG). Sejak didirikan, unit kerja di bawah naungan Badan Pengkajian dan Penerapan Teknologi (BPPT) ini telah melaksanakan aktivitas pengujian terowongan angin selama lebih dari 25 tahun terhadap pengguna jasa dari dalam maupun luar negeri, dan senantiasa mengembangkan seluruh komponen pembentuk sistem pengujiannya mengikuti tren teknologi terkini guna mewujudkan hasil pengujian yang memuaskan.

Salah satu komponen penting yang secara berkesinambungan dikembangkan adalah piranti lunak akuisisi data. Seperti ditunjukkan dalam Gambar 1, piranti lunak ini merupakan bagian dari sub-sistem akuisisi data dan memiliki fungsi utama sebagai pengendali perangkat keras sub-sistem akuisisi data dalam mengakuisisi data mentah hasil pengukuran oleh instrumentasi ukur dan menyalurkannya ke sub-sistem pengolahan data, dimana data mentah dikonversi menjadi informasi karakteristik aerodinamika dari objek uji dan selanjutnya ditampilkan dalam bentuk tabular atau grafik melalui sub-sistem presentasi