Machine Learning Apply for Financial Credit Approval to Filter Selected Customer in Domain Specific Bank

Uraiwan Inyaem[1*] and Sirina Chuaytem[1]

[1]Department of Mathematics and Computer Science, Faculty of Science and Technology,

Rajamangala University of Technology Thanyaburi, Khlong Luang, Pathum Thani 12110, Thailand

*E-mail: uraiwan.inyaem@rmutt.ac.th

Abstract

The objective of cooperative society is to support finance for members. A cooperative provides privilege of products and financial credit for members. The approval of the limit of financial credit is very complex condition. A recommending system can help cooperative officials to collect customer's profile and financial status to predict the limit of member's credit. The objective of the proposed research is to filter selected customer group for financial credit approval using data mining techniques. The dataset in the research with case study has 500 records used for preprocess transferring format into a categorization status value (CSV) file using WEKA program. The dataset is divided into 2 sets. First set is a training set used for creating model and the second set is a test set for evaluation model. For research methodology, the preprocessed data is a process using techniques of data cleaning to prepare data into suitable form before testing. The process of the research uses a data analysis technique which is Cross-Industry Standard Process for Data Mining (CRISP-DM). The experimental result is shown that accuracy value is 96.5517% with ADTree algorithm. The comparison between performance of algorithm are found that ADTree Algorithm and LMT Algorithm in test option, training set, shows maximum F-Measure 97.7% Precision 100% and Recall

95.5%. The proposed research is shown that the system can support cooperative officials to recommend the limit of financial credit for cooperative members in high quality.

1. Introduction

A Saving and Credit Cooperative Ltd. is the financial institution which supports members for economizing, saving money. A cooperative also has loan for a member by asking under cooperative approval. The objective of the cooperative is to solve members' economy and society. Many citizens can use credit card loans. At present, commercial banks set the rules, procedures and conditions of business operations for credit card as related to limits credit and the interest of commercial bank. The bank system has considered the applicant's qualifications which to approve the limit credit depend on a debt settlement capability and also the information system is provided for set, review policies and plans regarding credit card services as well. Generally, credit card in Thai banks have private card, business card/ corporate card, and joint venture. These cards are issued on applicant's require from the government agencies, state enterprises or businesses. The cooperative collects data by lecture note method. The service data such as loaning money, repayment, and purchase goods and so on are stored in high volume and complex data into database. The most customer's service data are gathered in a pattern of document file causing much time for searching. So, the customer relation service, for example asking for approval credit and checking customer's credit status are time consuming search as well. Bringing modern technologies to work is used with bank domain. Nowadays, data mining is famous technique for the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules [1] because that collect data warehouse support usefulness of the cooperative. The data mining has three area such as classification, clustering, and association rule work area. In data mining area, the knowledge discover from big data is classification process, the clustering process is method of data analysis and rely on similarity, and the association rule relates discovery with constraint rule depend on corresponding support, confidence and lift that is used in market basket analysis. From literature reviews, data mining is applied to use in many sectors and also in financial credit approval. In medical sector, Dina and et al. about studying combined breast cancer biomarkers using machine learning techniques. The research focused on breast cancer treatment using five Machine Learning Techniques (Decision Trees, Random Forest, K-Nearest Neighbor, Support Vector Machine and Gaussian) joining with different tests. Moreover, novel biomarkers are used to study too [2]. Agriculture sector, Niketa

and Leisa proposed a research about applying data mining to forecast rice in humid subtropical climatic zone of India. The Researchers compare technique between Decision Tree classifier and K-Nearest Neighbors classifier measure performance about the least failure of classification. The result shows that Decision Tree classifier technique is high performance for this research [3]. Advertisement sector, Hamed and et al. analyzed about advertisement websites classification. They analyzed the comparison between six different algorithms: Decision Stump, Hoeffding tree, Logistic Model Tree (LMT), Random Forest, Random tree, and REP tree. The objective was to find the best classification algorithm with the measurement of the highest precision and accuracy rate [4]. Military sector, Taghi and Robert proposed a model for classification using tree algorithm technique to analyze military software [5]. In this paper, data mining is applying in financial sector with machine learning as a famous algorithm (Decision Tree). The research focuses on filter customer to approve credit ceiling.

The methodology of the research is to investigate the best patterns and classify customer groups with machine learning algorithms. The best algorithm is used for classification in customer dataset. Customer data will be separated into two groups as small credit and large credit. The best pattern is implemented in graphic user interface tools for example, mobile application, and web application. The application can support cooperative officials to use customer's behavior description to decide loaning approval in high efficiency.

2. Materials and Experiment

Data mining strategies in this research are a type of learning called supervised modelling. The technique of this supervised modelling is a classification that divides customers into two groups as small credit and large credit. The framework of the research is Figure 1. The customer fill personal data in register form and then customer data detail is collected into electronic file. The system is saved to database in CSV format. In part of data mining process, we start to find pattern model.
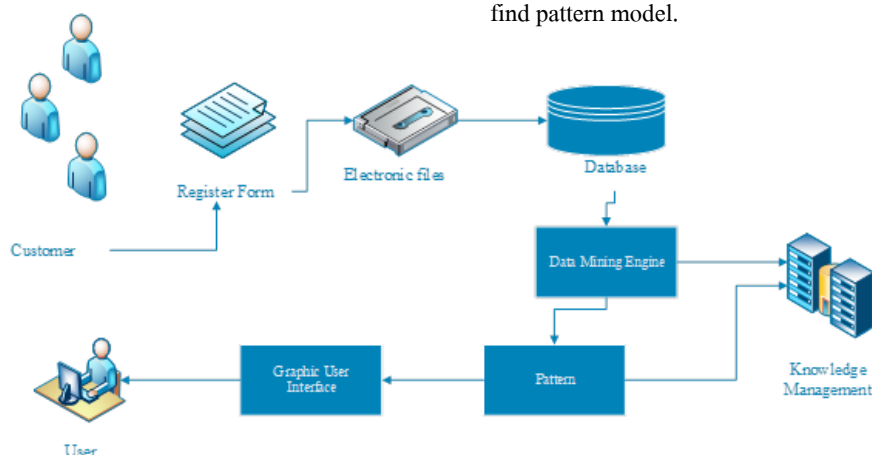

Figure 1. Shows the framework process

The process divide into two steps as follows: 1) training set data and 2) test data. The output of training set data, we got the pattern model for prediction data, and then data mining engine connect to knowledge management for prediction. The finally, the best pattern model is deploy in the real life system via graphic user interface such as mobile application, web application respectively. This research refers the model of Cross-Industry Standard Process for Data Mining: CRISP-DM [6-8] It is a process of standard data analysis as a well-known blueprint. Steps of the process are:

1) Business Understanding: Many people have problems with documents unclear to apply a credit card. Actually, applying a credit card applicants are fully qualified making surely the opportunity to be highly approved. The bank of Thailand set the qualification criteria clearly. Applicants must be aged 20 + to apply for the applicant's credit card, salary is more than 15,000 Thai baht, and working life is more than 6 months [9]. At the same time, financial institutions have different criteria for credit approval consideration.

Nowadays, bank of Thailand is issued the credit card approved depend on applicant's salary. By the ways, the salary level less than 30,000 Thai baht, the limit credit card approved equal 1.5 of applicant's salary while the salary level between 30,001-50,000 Thai baht, the limit credit card approved equal 3 of applicant's salary and the salary level more than 50,001 Thai baht, the limit credit card approved equal 5 of applicant's salary [10]. The research analyses concepts and factors of users' behavior. The factors are age, net income, and financial status respectively. The factors are very important to decide customer's credit ceiling for a cooperative's member.

2) Data preparation: Applicant's data is gathered 500 records in database. The beginning with the collection of applicant's data from any paper. The detail of data consist of 11 attributes as shows in Table 1 such as 1) Income source; salary or business. 2) Customer's age is not more than 70 years old. 3) Income calculation; salary or bonus. 4) Financial status; alone or couple apply.

Table 1 the attribute detail in dataset

No	Attribute Name	Data Type	Domain
1	Income	Numeric	Salary, business
2	Age	Numeric	Number of Age
3	Income calculation	Nominal	Salary or bonus.
4	Financial status	Nominal	Alone, couple
5	Sex	Nominal	Male, Female
6	Car	Nominal	Yes, No
7	Children	Numeric	Number of children
8	Account	Nominal	Yes, No
9	Career	Nominal	Government, Company, Unemployed
10	Asset	Numeric	Number of asset
11	Credit Approval (classes)	Nominal	small credit or large credit

5) Sex; male or female. 6) Car or vehicle; yes or no. 7) Children number. 8) Account; yes or no 9) Career path; government, company, unemployed. 10) Asset number and 11) Class label is credit approval with small or large credit. After collecting data, the data will be prepared with data cleaning process. Raw data cannot be used to test for the research because dataset is stored into incomplete forms. The values in some attributes are empty, some are not interesting to use, or some are effective from calculator or are compound from other attributes. In addition, some attributes gather data into different forms; for example, data about birthday or first working date are in the form of "01/15/1975" or "15-10-1975". The objective of preprocess data is to prepare data for accuracy, increasing performance and speed of data mining. The step of preprocess data, first step is data cleaning to fulfil missing data for the dataset that are inconsistent. So, the researcher has to assign many different forms for attribute "date" into one form. In case of missing data that can lead to misunderstanding, the research uses value "mean" for the attributes that are number and value "unknown" for the attributes that are letter (global constant to fill in the missing value). The disadvantage of Data Cleaning is doing manually that takes much time.

       3) Modelling: Decision Tree is the algorithm used for modelling. It is a linear structure.

The relation in the structure is hierarchy which is the most important node (Parent Node) is on top and the least important node (Child Node) is on bottom. Root Node is a node that has no Parent Node but it has Child Node. Leaf Node is nodes at the bottom. For data mining area, a decision tree is a supervised learning algorithm. The algorithm works for both discrete and continuous variables. It splits the dataset into subsets on the basis of the most significant attribute in the dataset [11] as shown in Figure 2. The decision tree structure can extract into association rule. The following rules can be show as follow: [12].

RULE 1

IF applicant is age less than 50 years old

AND applicant career is government

THEN applicant is approved for small credit

RULE 2

IF applicant is age less than 50 years old

AND applicant career is unemployed

AND applicant hasn't account

THEN applicant is approved for small credit

THEN large credit approved should be created.

RULE 3

IF applicant is age more than 50 years old

AND applicant has asset number more than 10

AND applicant has more than two children

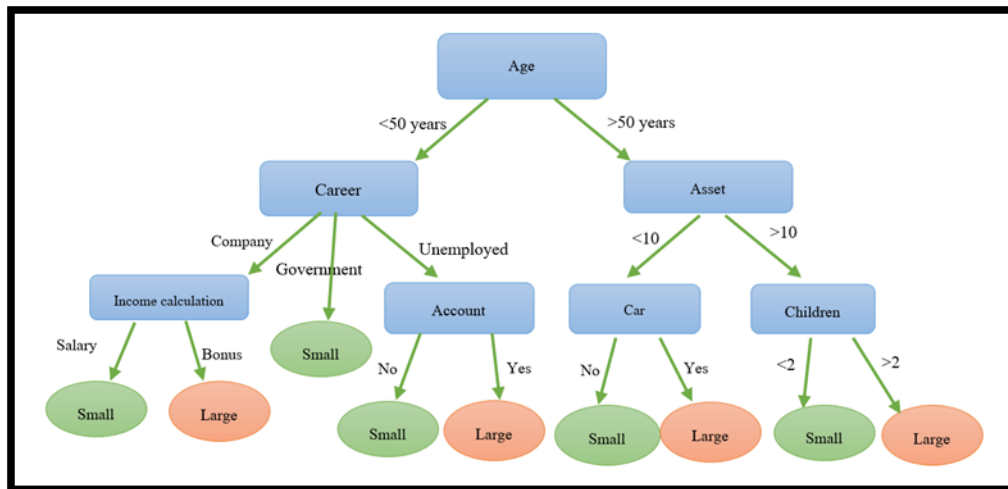*THEN applicant is approved for large credit*

Figure 2. Shows the decision tree structure in data mining area

The above shows how the decision tree in Figure 2 can be used to retrieve some of the knowledge concerning the functioning of limit credit approval. The resulting rules are exactly the same as those that were developed using the analytical model. In this research, WEKA tool is used to classify data by using Decision Tree algorithm. [13] The explanation of each algorithm in WEKA tool are:

a) ADTree algorithm for generating an alternating decision tree [14]. An alternation of decision nodes are the components of an ADTree. This tree is used to specify a predicate condition, and prediction nodes. Usually, ADTree contain a single number. The principle of classification for ADTree using all paths for all decision nodes are true and then summarize any prediction node that are traversed [11].

b) FT comes from Functional trees algorithm [15]. The algorithm is classification trees. It can be logistic regression functions at the inner nodes and/or leafs. In the research, the algorithm is used to classify with binary class target variables, nominal attributes and missing values respectively.

c) J48 algorithm is used for generating a pruned or unpruned C4.5 decision tree. J48 is Java open source implementation of the C4.5 algorithm in the WEKA tool for data mining [16-17].

d) LMT comes from Logistic Model Tree. The algorithm is used for building logistic model trees. It is classification trees with logistic regression functions at the leaf. In the research, the algorithms can use with binary class target variables, nominal attributes and missing values [18]. The dataset in the research with case study is 500 records transferring format into a categorization status value (CSV) files using WEKA program. The dataset is various training sets used to create a model and a test set for evaluation model.

Table 2 is show comparison between accuracy of classification algorithms

Algorithm Name	Test Options	Size of the tree	Time taken to build model (seconds)	Correctly classified Instances (%)	Incorrectly classified Instances (%)
ADTree	Training set	16	0.00	96.5517	3.4483
FT	Training set	1	0.11	75.8621	24.1379
J48	Training set	3	0.08	89.6552	10.3448
LMT	Training set	1	0.12	96.5517	3.4483
ADTree	Cross-validation folds 10	16	0.02	86.2069	13.7931
FT	Cross-validation folds 10	1	0.01	75.8621	24.1379
J48	Cross-validation folds 10	3	0.00	89.6552	10.3448
LMT	Cross-validation folds 10	1	0.09	86.2069	13.7931
ADTree	Percentage split 66%	16	0.00	80.0000	20.0000
FT	Percentage split 66%	1	0.01	90.0000	10.0000
J48	Percentage split 66%	3	0.00	90.0000	10.0000
LMT	Percentage split 66%	1	0.05	90.0000	10.0000

Table 2 shows four algorithm name such as ADTree, FT, J48 and LMT. The test options are training set, cross-validation folds 10 and percentage split (66%). The tree size of ADTree is 16, J48 size is 3, FT and LMT size are 1 level. Time taken to build model in seconds form is difference in each algorithm. The value of correctly and incorrectly classified instance is percentage form. Test options of training set, ADTree algorithm is high correctly classified with 96.5%, size of tree is 16 level and time take to build model is minimal period. Next, test options of cross-validation folds 10, J48 algorithm is high correctly classified with 89.6%, size of tree is 3 level and time take to build model is minimal period. Finally, the option of percentage split, J48 algorithm is high correctly classified with 90%, size of tree is 3 levels and time take to build model is minimal period respectively.

3. Results and Discussion

3.1 Performance measures

The experiments are evaluated by comparing the summaries generated by human experts for the same test set of previously unseen texts. The comparison is performed using an automated scoring program that rates each system according to Precision, Recall and F-Measures as [8] Precision measures the reliability of the information extracted as indicated in Equation (1). Recall measures the amount of the relevant information that the natural processing language system correctly extracts from the test dataset as indicated in the Equation (2) below, F-Measure is two times of product between Precision and Recall, and divide by sum of Precision and Recall as indicated in the Equation (3).

$$\text{Precision} = \frac{\text{\#correct slot fillers in output templates}}{\text{slot fillers in output templates}} \tag{1}$$

$$\text{Recall} = \frac{\text{\#correct slot fillers in output templates}}{\text{slot fillers in answer keys}} \tag{2}$$

$$\text{F-Measure} = \frac{(2*\text{Precision}*\text{Recall})}{(\text{Precision} + \text{Recall})} \tag{3}$$

3.2 The experimental results and discussion

The result shows two parts, accuracy and classification. The comparison between accuracy of classification algorithms are shown in Table 2. The comparison between the performance of algorithm are shown in Figure 2-4 From the result, the classification with ADTree algorithm gives maximum accuracy 96.5517%, test option is training set, and size of Decision Tree is 16. J48 algorithm gives maximum accuracy 89.6552% when test option is cross-validation folder 10. Finally, FT, J48, LMT gives maximum accuracy 90% when test option is Percentage split 66%. The comparison between performances of algorithm shows that ADTree algorithm and LMT algorithm when test option is training set gives maximum F-Measure 97.7 % Precision 100% and Recall 95.5%. J48 algorithm when test option is cross-validation folder 10 gives maximum F-Measure 93.6 % Precision 88% and Recall 100%. The FT algorithm and LMT algorithm when test option is percentage split 66% gives maximum F-Measure 94.7 % Precision 90% Recall 100% as follow.

Figure 3. Comparison between F-Measure, Precision and Recall of classification algorithm of training set option

In case of test option is training set that shows in Figure 3, the result shows that ADTree and LMT algorithm give maximum F-Measure 97.7 % Precision 100% and Recall 95.5%. Next sequence, the J48 algorithm give F-Measure 93.6% Precision 88% and Recall 100% and last sequence, the FT algorithm give F-Measure 86.3% Precision 75.9% and Recall 100%.
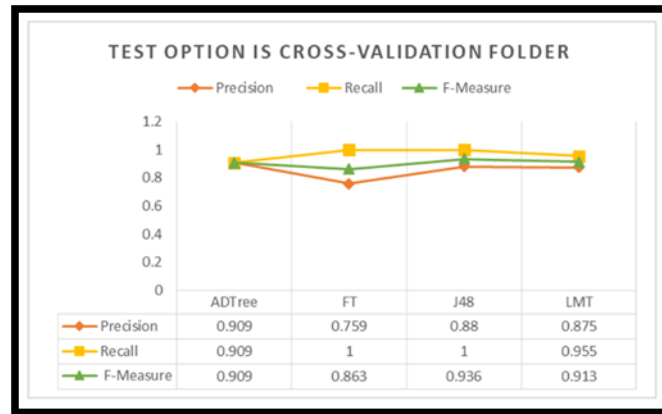

Figure 4. Comparison between F-Measure, Precision and Recall of classification algorithm of cross-validation folder option

In case of test option is cross-validation folder option (10 folders) that shows in Figure 4, the result shows that J48 algorithm give maximum F-Measure 93.6% Precision 88% and Recall 100%. Next, the LMT algorithm give F-Measure 91.3%. Precision 87.5% and Recall 95.5% . The subsequence, the ADTree algorithm give F- Measure, Precision, and Recall equal 90.9% and last sequence, the FT algorithm give F-Measure 86.3% Precision 75.9%and Recall 100%.

Figure 5. Comparison between F-Measure, Precision and Recall of classification algorithm of percentage split 66% option

In case of test option is percentage split that shows in Figure 5, the result shows that FT and LMT algorithm give maximum F-Measure 94.7 % Precision 90% and Recall 100% . Next sequence, the J48 algorithm give the F-Measure 94.1 % Precision 100% and Recall 88.9% and last sequence, the ADTree algorithm give F-Measure, Precision and Recall equal 88.9%.

4. Conclusions

This research uses data mining technique to find the best method for customer's classification. The objective is to approve financial credit for customers. The experimental result for the accuracy value is 96.5517% with ADTree algorithm. The comparison between performances of algorithm shows that ADTree algorithm and LMT algorithm when test option is training set gives maximum F-Measure 97.7 % Precision 100% and Recall 95.5% . The bank of Thailand has the rules for assingning a limit credit approved for the customer. The results from this research, the model pattern supports decision-making which reduce any errors arise from limits credit approval. Executive director use to management tools for credit approval that quick and convenient operation and also more accuracy. The model patter can develop into graphical user interface such as web application, mobile application and also application that is a guide for modern credit approval. The proposed research is shown that the system can support staff members in a cooperative to recommend credit ceiling of members in high quality in real life.

Research in the future, researcher plans to test in big data with complex data mining algorithm. It has to be done for more accuracy to develop the result of the research. The data processing from data mining Technique can be formed be relation rules using for suitable decisions. Each recommendation is based on a customer's transaction and his/her money spent in a cooperative.

5. Acknowledgements

6. References

[1]　Gordon SL, Michael J. AB. Data Mining Techniques for Marketing, Sales and Customer Relationship Management, 3rd ed. Wiley Publishing: 2001.

[2]　Dina TS, Amir A, Olfat S. Studying Combined Breast Cancer biomarkers using Machine Learning Techniques. SAMI. 2016; 247-251.

[3]　Niketa G, Leisa A. Applying Data Mining Techniques to Predict Yield of Rice in Humid Subtropical Climatic Zone of India. INDIACom.2016; 1901-1906.

[4]   Hamed J, Seyed JM, Ali H. Evaluation and Analysis of Popular Decision Tree Algorithms for Annoying Advertisement Website Classification. CSNT.2015; 1025-1029.

[5]   Taghi MK, Robert H. A Tree-Based Classification Model for Analysis of a Military Software System. High-Assurance Systems Engineering Workshop. 1996; 244-251.

[6]   Eakasit P. An Introduction to Data Mining Techniques. 1st ed. Bangkok: Asia Digital Press Co.,Ltd; 2014.

[7]   Colin S. The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13-22, 2000.

[8]   Jiawei H, Micheline K, Jian P. Data Mining: Concepts and Techniques, 3rd ed. MA: Morgan Kaufmann Publishers; 2012.

[9]   Moneyhub. [online] Source: https://moneyhub.in.th

[10] Limited Credit Approval. [online] source https://masii.co.th/

[11] Decision Tree Algorithm. [online] source https://en.wikipedia.org

[12] E. Szczerbicki, Management of Complexity and Information Flow in Agile Manufacturing: The 21st Century Competitive Strategy, 20

[13] Machine Learning Group at the University of Waikato, New Zeland. WEKA.

[14] Freund Y, Mason L. The alternating decision tree learning algorithm. In Proceeding of the Sixteenth International Conference on Machine Learning, Bled, Slovenia.1999; 124-133.

[15] Niels L, Mark H, Eibe F. Functional Trees. Logistic Model Trees. 2005.

[16] Ross Q. C4. 5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann Publishers;1993.

[17] Gaganjot K, Amit C, Improved J48 Classification Algorithm for the Prediction of Diabetes, International Journal of Computer Applications 98(22), July 2014. 0975 – 8887.

[18] Niels L, Mark H, Eibe F. Logistic Model Trees. Machine Learning. 2005; 95 (1-2) : 161-205.