Smart dunning to improve collection ratio in internet service provider using C4.5 algorithm

A H Thohari[1], W S Anita[1]

[1] Department of Informatics Engineering, Politeknik Negeri Batam, Jalan Ahmad Yani, Batam, Indonesia

E-mail: hamim@polibatam.ac.id

Abstract. Collection Ratio (CR) is the ratio between the total payments of all customers to the total invoice in the current month. CR is also act as performance indicator of the payment collection division for internet service companies. One method that currently used to improve CR is dunning, namely providing customer with information about the bills with various communication methods including visiting the customer's address. Large number of customers and limited number of collectors are the major obstacle in the dunning process. We propose a classification method to predict potential delinquent customers, so that the expected accuracy of dunning increases, which in turn increases the company's collection ratio. We use the decision tree method with the C4.5 algorithm on the historical data from one internet service provider customers in the Riau Islands province, Indonesia. The classification process produces a decision tree with 5,885 leaves and 6,765 tree size. The decision tree then evaluated with 10-folds cross validation that resulting in 78.54 percent accuracy, 0.738 precision, and 0.785 recall. The decision tree has been applied to the dunning process of the company.

1. Introduction

Internet Service Provider is a company or entity that provides internet services or other related services. This company provides internet telephone and television subscription services for the public. Services provided to the public are paid periodically every month.

Collection Ratio (CR) is a comparison between total customer payments to the number of arrears that exist. Collection ratio is one indicator of the performance of the Payment Collection and Finance division. Table 1 shows the percentage of collection of internet service provider companies in the region 1 of Riau Islands in the January to March 2018 period.

Table 1. Collection Ratio January – March 2018.

Month	Internet	Phone
January	93.09%	95.81%
February	91.92%	95.09%
March	92.05%	95.69%

Table 1 shows that the percentage of CR is not stable, even in February it experienced a quite dramatic decline. The company's efforts to increase CR are dunning, which is sending billing information so that customers pay on time. Dunning is also act as media fir customers submit complaints.

Dunning is carried out by collection agents with several media such as SMS, e-mail, fax, WhatsApp, Telegram, and direct visits to the customer's address.

The large and increasing number of customers and the limited number of collection agents make the dunning process not optimal. Customer identification process is needed to determine the potential of monthly bill arrears based on attributes or customer profiles; thus, the dunning process is more targeted and can increase the company's CR.

Data mining has been widely used for maintaining relationship between customer and service provider especially in telco industry, such as targeted marketing, customer loyalty, churn rate, and calculating customer lifetime value [1-4]. Decision tree is the most widely used method for classify customer for telecom industry [5], it delivers accurate model for predicting customer churn using customer demographic data [2]. Other research that using classification to improve invoice to cash collection using supervised learning suggest that C4.5 and PART is the most accurate classification algorithm to use for their case [6]. Decision tree model also used to extract important parameters for identifying customer value, credit and loyalty [7]. We propose the application of a classification method with a decision tree using the C4.5 algorithm to predict potential customers in arrears. The case study for this research is internet service provider company in the Riau Islands region, Indonesia.

Rest of this paper will present the theoretical foundation in research, continuing the process of selecting training data and pre-processing data, as well as calculations to generate the decision tree. The last part is discussing the results, conclusions and suggestions for further research.

2. Theory

Decision tree defined as structured model that can be used to convert data into tree structure that depict a decision rule [8]. There are several algorithms that can be used to make a decision tree, one of which is the C4.5 algorithm. This algorithm is the result of the development of the ID3 algorithm created by J. Ross Quinlan, a researcher in the field of artificial intelligence in the late 1970s until the early 1980s [9]. The C4.5 algorithm, ID3 adopts a greedy approach where the process starts from the root node to the leaf node which is done recursively. There are 4 main stages in C4.5 algorithm to build a decision tree [10]: attribute selection for the root node, generating tree branch, divide the cases into branches, iteration process for each branch until all the cases in the branch have the same class

In selecting an attribute, it will be used as a node, both root and internal root are based on the highest gain value of the existing attributes. The formula for calculating gain is portrayed in Equation (1). Formula for calculating Entropy is described in Equation (2).

$$Gain(S,A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \tag{1}$$

Where:

S	: Number of datasets
A	: Attribute
n	: Number of partitions in attribute A
$\lvert Si \rvert$	: Number of cases in partition i
$\lvert S \rvert$	: Number of cases in S

$$Entropy(S) = \sum_{i=1}^{n} -p_i * log_2\, p_i \tag{2}$$

Where:

S	: Number of datasets
n	: Number of partition S
p_i	: Proportion of Si to S

3. Implementation and results

3.1. Dataset

The data to be used in this study are C3MR data (cash collection current month ratio) from January to March 2018. Customer data is taken on the 21st of each month, where on the 20th the due date of the bills is settled. The data taken restricted to the Riau Islands Province.

The data consist of several variables or attributes, such as item number service, account number, name, Segment, age, period, entry date, total payment, instalments, collecting agents, which will be refined and processed using the C4.5 algorithm. The numbers of data to be used are 5.513 tuple from payments in February 2018.

3.2. Data understanding and preprocessing

After collecting data, next step is data understanding and pre-processing, where the data that has been collected are prepared for processing. Data pre-processing is divided into 2 stages, first filtering / data cleaning so that it is easy to observe and the second stage is eliminating noisy. The first step we have to do is select the attributes / variables that will be used for the C4.5 algorithm. The attributes that are used in the calculations are as follows:

- Payment: The label of the prediction whether the customer pay the bill on time or late.
- Segment: The customer segment, such as personal use, business, education, or government. This attribute has 3 code, DCS for residential customer, DBS for business customer, DES for corporate customer.
- Subscription period: The subscription period of each customer that counted in month, we then transform this attribute to two category namely new customer for customer that have less than 12 months subscription and loyal customer for the customer that have more than 12 months subscription.
- Period: Payment period.
- Entry Date: The date that payment is made and the payment data entered into system.
- Usage: Internet or quota usage of a user, categorized as used and not used.
- Instalment: Whether the customer pay with instalment or not.
- Collecting Agent: Payment point where customer made payment.

The attributes that are eliminated are service number, account number, customers number. These three attributes are not used because it shows the customer's identity and the need for this research only determines the pattern of customer payments. Other attributes that are eliminated are total amount paid because the values in this attribute are too large.

3.3. Entropy and gain calculation

By using the total entropy formula in Equation (2) that has been mentioned earlier, we can calculate the entropy value by calculating the number of customer entry dates "TERTIB" (on time payment) and "TELAT" (late payment) from all existing cases.

$$Entropy(Total) = (-\frac{1643}{5513} * log\,2\,(\frac{1643}{5513})) + (-\frac{3870}{5513} * log\,2\,(\frac{3870}{5513}))\text{=0.559740} \quad (3)$$

In Equation (3), the entropy Total is the total value of customers who have “TERTIB” (1.643 data) “TELAT” (3.870 data) while 5.513 is the total dataset.

After all the entropy and information gain are calculated, result of the calculation are compiled in Table 2. Based on the calculation result in Table 2, it can be seen that segment attribute have highest information gain at 0.217755, and usage have the lowest gain at 0.020547. We use the segment attribute that have highest gain as root for decision tree. We iterate the calculation until all attribute have same class.

Table 2. Entropy and gain calculation result.

Variable	Value	Total Case	Payment		Entropy	Information Gain
			On Time	Late		
Total		5513	3870	1643	0.55974	
Segment	Dcs	4409	3108	1307	0.365924	0.217755
	Dbs	921	701	220	0.487021	
	Des	183	67	116	0.430552	
Subscription Period	New Customer	579	341	238	0.666578	0.086171
	Loyal Customer	4934	3529	1405	0.749595	
Period	February	5513	5513	0	0	0.026066
Usage	Used	5407	3786	1621	0.482066	0.020547
	Not Used	106	84	22	0.918296	
Installment	Yes	2	0	2	0	0.025879
	No	5511	3870	1641	0.863121	
Collecting Agent	Provider	3612	2490	1122	0.50665	0.004903
	Other	1901	1380	521	0.791858	

3.4. *Entropy and gain calculation*

Result of the calculation and the iteration then processed into decision tree using data mining software. The preview of the result is portrayed in Figure 1. Size of tree is the number of nodes produced in a tree while the number of leaves is all nodes that does not have successor. Based on Figure 1, the size of tree is 6,765 and the number of leaves is 5,885. The result of calculation then validated using 10 folds cross validation to check the accuracy of classification. Validation process show that the classification model has 78.54% accuracy with 0.738 precision and recall 0.785.

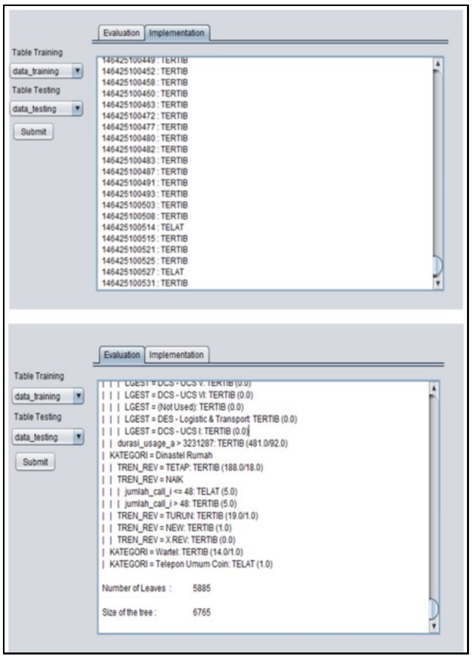

Table Training: data_training
Table Testing: data_testing

```
146425100449 : TERTIB
146425100452 : TERTIB
146425100458 : TERTIB
146425100460 : TERTIB
146425100463 : TERTIB
146425100472 : TERTIB
146425100477 : TERTIB
146425100480 : TERTIB
146425100482 : TERTIB
146425100483 : TERTIB
146425100487 : TERTIB
146425100491 : TERTIB
146425100493 : TERTIB
146425100503 : TERTIB
146425100508 : TERTIB
146425100514 : TELAT
146425100515 : TERTIB
146425100521 : TERTIB
146425100525 : TERTIB
146425100527 : TELAT
146425100531 : TERTIB
```

Table Training: data_training
Table Testing: data_testing

```
| | | LGEST = DCS - UCS V: TERTIB (0.0)
| | | LGEST = DCS - UCS VI: TERTIB (0.0)
| | | LGEST = (Not Used): TERTIB (0.0)
| | | LGEST = DES - Logistic & Transport: TERTIB (0.0)
| | | LGEST = DCS - UCS I: TERTIB (0.0)
| | durasi_usage_a > 3231287: TERTIB (481.0/92.0)
| KATEGORI = Dinastel Rumah
| | TREN_REV = TETAP: TERTIB (188.0/18.0)
| | TREN_REV = NAIK
| | | jumlah_call_i <= 48: TELAT (5.0)
| | | jumlah_call_i > 48: TERTIB (5.0)
| | TREN_REV = TURUN: TERTIB (19.0/1.0)
| | TREN_REV = NEW: TERTIB (1.0)
| | TREN_REV = X REV: TERTIB (0.0)
| KATEGORI = Wartel: TERTIB (14.0/1.0)
| KATEGORI = Telepon Umum Coin: TELAT (1.0)

Number of Leaves :        5885

Size of the tree :        6765
```

Figure 1. Decision Tree Preview.

4. Conclusions

We have applied data mining Classification using the decision tree method and C4.5 algorithm to predict delinquent customers for internet service provider company in Riau Island Province. The advantage of the decision tree is a lower error rate. The classification employs historical customer payment data with selected attribute and pre-processing to normalize and anonymize the data. The output is a decision tree model with 5,885 leaves and size of the tree 6,765. Time required to build the model is 31.5 seconds using data mining software. The decision tree than evaluated using 10-folds cross validation, the accuracy from the validation process is generated at 78.5447% with precision of 0.738 and recall 0.785.

With the accuracy of the result, the decision tree model can be implemented on the current customer data. The model can predict which customers will potentially be late in payments. The company can do more accurate dunning with various communication method, so that it is expected to increase collection ratio on internet service provider companies.

5. References

[1] Hwang H, Jung T and Suh E 2004 *Expert Systems with Applications* **26** 181-188

[2] Hung S Y, Yen D C and Wang H Y 2006 *Expert Systems with Applications* **31** 515-524

[3] Dahiya K and Bhatia S 2015 *4th International Conference on Reliability, Infocom Technologies and Optimization* 1-6

[4] Zulkifli A 2016 *Riau Journal of Computer Science* **2** 65-76

[5] Almana A M, Aksoy M S and Alzahrani R 2014 *International Journal of Engineering Research and Applications* **45** 165-171

[6] Zeng S, Melville P, Lang C A, Boier-Martin I and Murphy C 2008 *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **14** 1043-1050

[7] Han S H, Lu S X and Leung S C 2012 *Expert Systems with Applications* **39** 3964-3973

[8] Berry M J and Linoff G S 2004 *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management* (USA: John Wiley & Sons)

[9] Quinlan J R 2014 *C4. 5: Programs for Machine Learning* (USA: Morgan Kaufmann Publisher)

[10] Larose D T and Larose C D 2014 *Discovering Knowledge in Data: An Introduction to Data Mining* (USA: John Wiley & Sons)