APPLICATION OF C4.5 ALGORITHM FOR DETECTION OF COOPERATIVES FAILURE IN PROVINCE LEVEL

Anik Andriani
Manajemen Informatika, AMIK BSI Jakarta
Jl.R.S.Fatmawati No.24, Pondok Labu, Jakarta Selatan
email: anik.aai@bsi.ac.id

Abstract – Cooperative is one of the actors in Indonesia's economy is expected to become pillar of the economy in Indonesia. Based on the stastistical data on site Ministry of Cooperatives and Small and Menengah Enterprises are many cooperatives at the provincial level have failed. The purpose of this research is to create a classification of failure cooperative with C4.5 algorithm so that it can be seen that the most influential factor in the failure cooperative. The data is used to classificate are analyzed used Knowledge Discovery Databases (KDD) which consist of nine stages. Classification results was tested by using confussion matrix and ROC curves to determine the performance of the classification results. The results of the classification failure cooperatives in province level with the C4.5 algorithm shows performance with the accuracy of training data was 85,45%, and the accuracy of the testing data was 71,67%. While the results of the classification performance as measured by the ROC curve shows the results of classification rules including excellent classification.

Keywords: classification, C4.5 algorithm, confusion matrix, ROC curve

I. INTRODUCTION

Economic agents in Indonesia there are three that the State Owned Enterprises (BUMN) or Regional Owned Enterprises (BUMD), cooperatives, and Private Companies (BUMS). Seeing that the role of cooperatives is legitimate in Indonesia even expected to be a cornerstone of the economy in Indonesia [1]. For the development of cooperatives in Indonesia should be concern goverment. Even when viewed in the monetary crisis that had hit the economy in Indonesia, where many BUMN/BUMD enterprises and private enterprises were uprooted due to bear a lot of debt, cooperatives became the proponent of economy so Indonesia economy can still walk. Although cooperatives as one of the supporters of the economy, but the development of cooperatives in Indonesia has not grown up. This can be seen from the cooperative data province level in Indonesia that is retrieved from the Cooperatives Departement's website shows the number of cooperatives inactive in Indonesia still high.

The purpose of this research is to create a data classification cooperative failure to identify factors that affect of failure of cooperatives in the province level. Classification is one data mining techniques that can be used to determine the pattern of a dataset. In this study using the C4.5 algorithm which is one algorithm that can be used in data classification.

Based on the background of the above problems can be made formulation of the problems is written into the research question, among others:

1. How C4.5 algorithm can classify cooperative failure data at the provincial level in Indonesa?
2. How do performance level of the classification result?
3. What is the most influential factor in the failure of cooperatives at the provincial level in Indonesia?

II. THEORY

2.1. Cooperative

According to Mohammad Hatta, who is the founding father of Indonesian Cooperatives suggests Economy which is based on the principle of kinship is cooperative. Cooperative is a business entity that is based on mutual cooperation. The principle of the cooperative is prioritize interests together rather than the interests of the individual. So that the cooperative should have the function of educating the public in the care of shared purposes [2].

Seven task of cooperatives in Indonesia [2]:

1. Production repair, there are three main types of goods that must be corrected immediately ie food, craft items and carpentry items
2. Improving the quality of goods, so that the cooperatives role has production facilities together
3. Improve the distribution, to prevent hoarding
4. Price fixing, preventing the very high goods price
5. Getting rid of exploitation, avoid usury system
6. Strengthen the capital
7. Maintaining barn, customize a production and consumption and consumption as a buffer stock

2.2. Algoritma C4.5

Data mining is the extraction of implicit from the data to obtain information that cloud potentially have a use value that was previously unknown. Data mining is also a process for finding patterns where of data to be generated automatically or semi-automatically and provide benefits [3]. The purpose of the use of data mining there are two [4]:

1. Prediction method, which uses some existing variables to predict the future (not yet known). Examples of its use, namely classification, regression, detection bias/anomalies, and others.
2. Descriptive method, which uses patterns in the data, easily interpreted by the user. Examples of its use clustering, association rules, sequential patterns

Classification is a branch of data mining discovery. Classification itself is used to help classify the data. There are four basic components in the classification process, among others [4]:

1. Class, the dependent variable of a classification model which is a categorical which is usually represented by a label that shows the classification results, as an example: class pass and not, class buy and not, and class approved or not, etc.
2. Predictors, is the dependent variable of a classification model that is represented by the characteristics (attributes) of the data to be classified and based on the classification made, for example: predictors for a class pass or not is GPA, attendance, tuition, fees, etc.
3. Training dataset, a dataset that is used to identify a class of data pattern in the classification based on predictors that have been available.
4. Testing dataset, the new data will be classified based on the model that has been obtained from the training dataset to determine the classification accuracy of the classification results (performance model) so that the classification can be evaluated.

Classification rules can be easier to read if described with decision tree. In the decision tree into a rule makes a decision tree is not easy so before made decision trees can be made structures such as the following rule [3]:

Jika a dan b maka x

Jika c dan d maka x

Figure 1. Examples of the structure of the classification rule [3]

From the above rule can be described as a decision tree in figure 2 below:

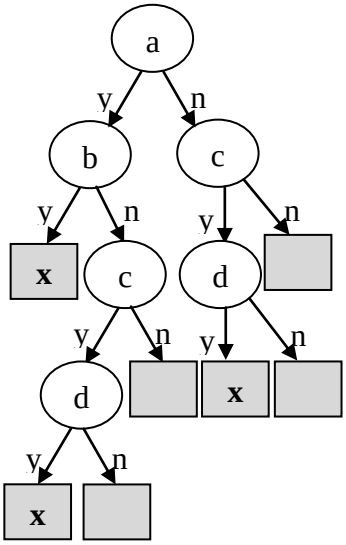

Figure 2. Decision Tree from the rule classification [3]

One algorithm that can be used for classification is the C4.5 algorithm. The working of C4.5 algorithm is learn the attributes in a data set that is subsequently mapped into attributes into a class where the class is applied to a new classification of the unknown. Learning in C4.5 algorithm is learning to map a set of data that results can be applied to other cases [5]. Stages in the making of a decision tree using the C4.5 algorithm, among others [6]:

1. Prepare the training data. Training data are usually taken from historical data that never happened before or referred to past data and are already grouped in certain classes.
2. Calculating the roots of the tree. The roots will be taken from the attribute to be selected, by calculating the value of the gain of each attribute, the highest gain value that will be the first root. Before calculating the gain of an attribute value, first calculate the entropy value. To calculate the entropy value used formula,

$$Entropy\ (S) = \sum_{i=1}^{n} - pi\ \log_2 pi \qquad (1)$$

Specification:
S = the set of cases
n = number of partitions S
pi= the proportion of Si to S

Then calculate the gain value using the formula,

$$Gain\ (S, A) = \ entropy\ (S) - \sum_{i=1}^{n} \frac{|Si|}{S} * Entropy\ (Si) \qquad (2)$$

Specification:
S = set of Case
A = feature
n = number of partitions of attributes A
|Si| = Proportion of Si to S
|S| = number of cases in S

3. Repeat step 2 and 3 until all records partitioned
4. Partitioning process will stop when the decision tree:
   a. All records in the node N gets the same grade
   b. There is no record in the attribute partitioned again
   c. There is no record in the empty branch

2.3. Evaluation

Each classification algorithms that give results to the case of unknown called classifier. The most obvious criterion used to measured classification results is the prediction accuracy. Overall prediction accuracy is a breakdown of the performance of the classifier is how often the case X is correctly classified in class X [7]. In this study to measure the performance level of the classification results can use Confusion Matrix and the ROC curve (Receiver Operating Characteristic).

1. Confusion Matrix
   Confussion matrix is a method for evaluating the results of classification using matrix table which consists of two classes as shown in Table 1 where the matrix table is considered as a first class and second class considered positive negative [7].

Table 1. Confusion Matrix Model [8]

Correct Classification	*Classified as*	
	+	-
+	*True positives*	*False negatives*
-	*False positives*	*True negatives*

In table 1 shows the confusion matrix models which are true positives correctly shows the value correctly predict the value corresponding classification results, whereas false positives appropriate one to predict the results of the classification [8].

2. ROC curve
   ROC curve shows the classification accuracy and to compare visually. ROC express confusion matrix. ROC is a two-dimensional graph with false positives as a horizontal line and a vertical line as a true positive. The area under the curve (AUC) was calculated to measure the performance difference in the methods used [9].
   ROC has a level of diagnostic value, namely [4]:
   a. Accuracy is worth 0.90-1.00 = excellent classification
   b. Accuracy is worth 0.80-0.90 = good classification
   c. Accuracy is worth 0.70-0.80 = fair classification
   d. Accuracy is worth 0.60-0.70 = poor classification
   e. Accuracy is worth 0.50-0.60 = failure

III. METHOD RESEARCH

This study uses data downloaded from the Ministry of Cooperatives there were 357 data. Whereas the stages of the research can be seen in Figure 3.

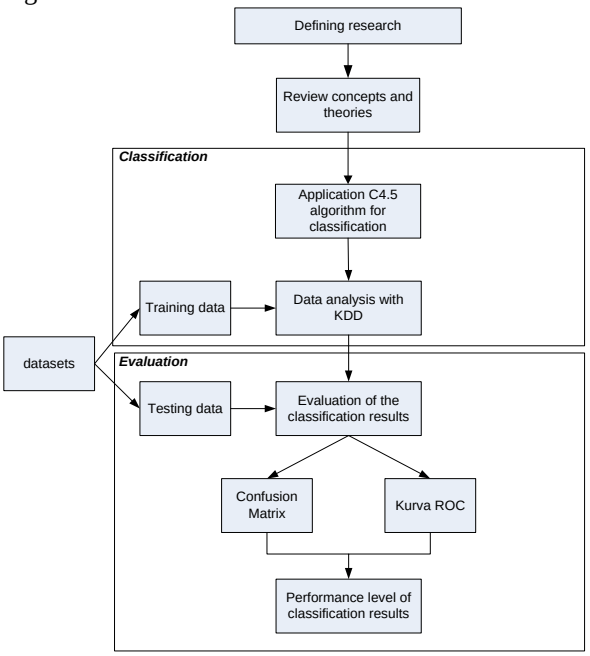

Figure 3. The stages of research

From the figure 3 we seen the research beginning with defining research and review concepts and theories. After defining the research is to collect research data is cooperatives each province dataset. Data analyzed with Knowledge Discovery in Databases (KDD) consisting of nine stages depicted in figure 4. After analysis of the data by KDD, the classification is done on the training data with C4.5 algorithm. Classification results evaluated by confusion matrix and ROC Curves. Evaluation of the training data and testing data to measure the performance level of the classification results.

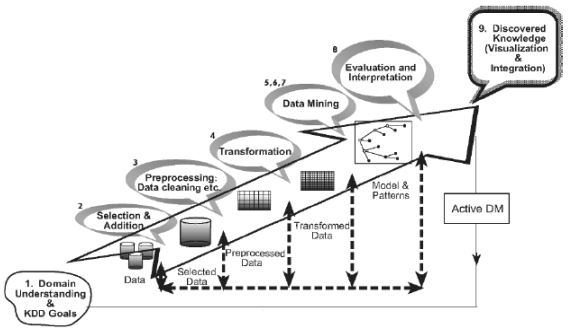

Figure 4. Knowledge Discovery in Databases [10]

Figure 4 ilustrations nine stages in KDD, namely:

1. Developing an understanding of the application domain, preparation phase to determine baseline measures
2. Selection and creating a data set on which discovery will be performed, selecting and creating research data. In this study uses data cooperatives each province from the Ministry of Cooperatives.
3. Preprocessing and cleansing, stage increases the reliability of data by deleting data that is incomplete (missing value) and incorrect data (noise). In this stages data obtained for classification amounted to 300.
4. Data transformation, stage of preparation and development for better data. Data to be used for research are transformed to the category. Examples of data that has been categorized shown on table 2.

Table 2. Data that has been categorized

Total_cooperatives	Members	RAT	Manager	Employees	Own_Capital	Outside_Capital	Business_Volume	SHU	REMARK
5001-10000	<1juta	<=5000	1001-2000	<=10000	500M-1T	500M-1T	500M-1T	50-100M	Fail
<=5000	<1juta	<=5000	<=1000	<=10000	100-500M	100-500M	500M-1T	10-50M	Fail
<=5000	<1juta	<=5000	<=1000	<=10000	100-500M	100-500M	500M-1T	10-50M	Fail
<=5000	<1juta	<=5000	<=1000	<=10000	<100M	100-500M	500M-1T	10-50M	Success
<=5000	<1juta	<=5000	<=1000	<=10000	100-500M	100-500M	500M-1T	10-50M	Success
<=5000	<1juta	<=5000	<=1000	<=10000	<100M	<100M	<100M	10-50M	Success
<=5000	<1juta	<=5000	<=1000	10001-20000	100-500M	500M-1T	500M-1T	50-100M	Fail
<=5000	<1juta	<=5000	<=1000	<=10000	<100M	<100M	<100M	<10M	Fail
5001-10000	1-2juta	<=5000	<=1000	20001-30000	100-500M	100-500M	>2T	100-150M	Fail
15001-20000	4-5juta	5001-10000	2001-3000	<=10000	1-1,5T	>2T	>2T	100-150M	Fail

Of the total data amount to 300 divided into two for the training data (80%) and data testing (20%). The division of data into training data and testing data using systematic random sampling technique. The use of systematic random sampling technique, random is performed only once when determining the first element of the sampling be taken. Determination of the elements of the next sampling reached by interval sample. Sample interval or intervals the ratio obtained by dividing the population size by the desired sample size (N/n) [11]. The results of division of data by systematic random sampling of testing data as much as 60, and as many as 240 training data.

5. Choosing the appropriate Data Mining task, the selection phase of data mining techniques. Data mining techniques selected in this study is the classification.
6. Choosing the Data Mining Algorithm, this stage of the algorithm selection. Algorithm selected for classification is C4.5 algorithm.
7. Employing the Data Mining Algorithm, application of classification with C4.5 algorithm.
8. Evaluation, classification results are evaluated by confusion matrix and ROC curves
9. Using the discovered knowledge, implementation of classification results.

IV. RESULT AND DISCUSSION

Classification result with C4.5 algorithm in the form of a decision tree like figure 5.

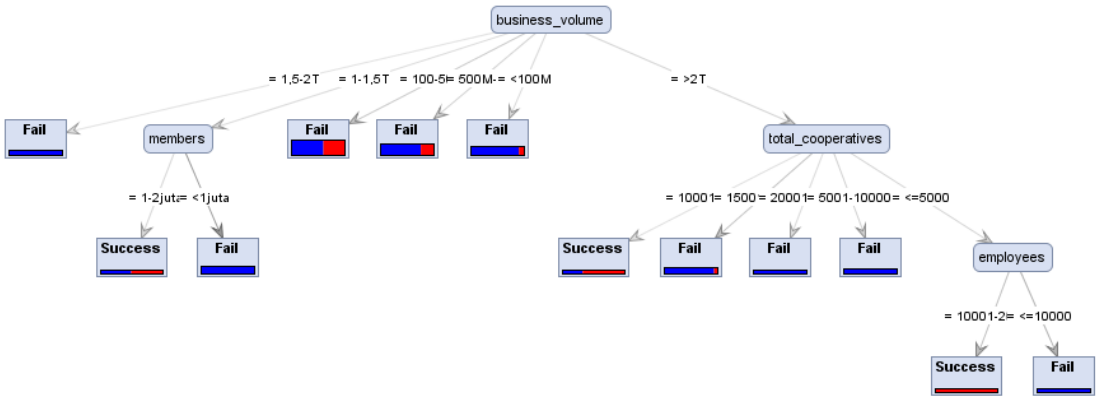

Figure 5. Decision tree of cooperative failure detection at the provincial level

In the figure 5 obtained the most influential attribute to the failure of the cooperative is a business volume. The results of classification evaluated by confusion matrix for obtain accuracy value. The first evaluation with confusion matrix is done with the training data. The results can be seen in figure 6.

accuracy: 85.42%

	true Fail	true Success	class precision
pred. Fail	173	21	89.18%
pred. Success	14	32	69.57%
class recall	92.51%	60.38%	

Figure 6. Testing results of the classification with the training data

From the figure 6 show the accuracy value from evaluation classification results with the training data of 85,42%. Evaluation classification results with ROC curve is done with the training data can be seen in figure 7.

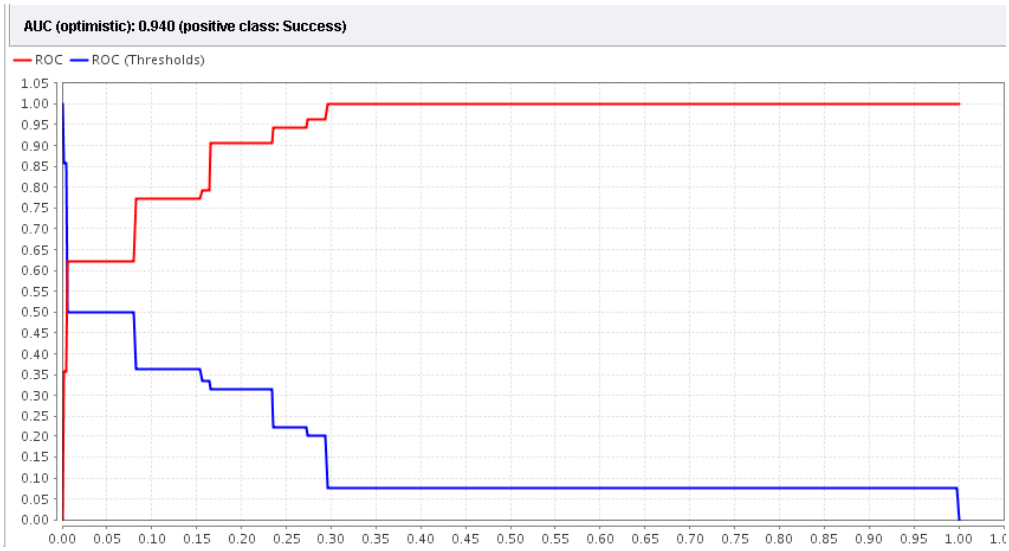

Figure 7. Testing results of the classification with the training data with ROC curves

In the figure 7 show evalution of classification results with the training data valued 0,940 so it can be classified as excellent classification. Evaluation with confusion matrix is done with the testing data can been seen in figure 8.

accuracy: 71.67%

	true Success	true Fail	class precision
pred. Success	12	4	75.00%
pred. Fail	13	31	70.45%
class recall	48.00%	88.57%	

Figure 8. Testing results of the classification with the testing data

From the figure 8 show the accuracy value from evaluation classification results with the testing data of 71,67%. Evaluation classification results with ROC curve is done with the testing data can be seen in figure 9.

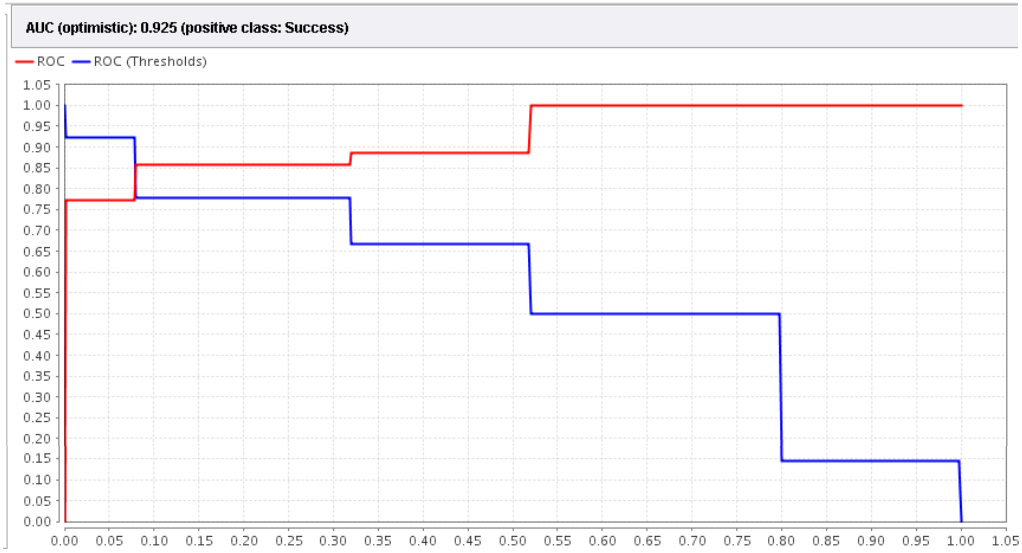

Figure 9. Testing results of the classification with the testing data with ROC curves

In the figure 9 show evalution of classification results with the testing data valued 0,925 so it can be classified as excellent classification.

IV. CONCLUSION

Based on the result achieved can be inferred using classification techniques with C4.5 algorithm can be applied in making the classification cooperative "Fail" and 'Success". The most influential factor in the failure of the cooperative is a business volume. Evaluation of the results of the classification performed by the confusion matrix and ROC curves. Classification results are evaluated by the training data shows accuracy rate of 85,42% and with the data of testing showed 71.67% accuracy rate. While the evaluation of the ROC curve shows the classification results into the excellent classification category, both evaluation on the training data and data testing.

REFERENCE

[1] P. B. Santoso, "Eksistensi Koperasi: Peluang dan Tantangan di Era Pasar Global," Dinamika Pembangunan Vol.1 No.2, pp. 111-117, 2004.

[2] Y. Harsoyo, et al., Ideologi Koperasi Menatap Masa Depan. Tangerang: PT.Agromedia Pustaka, 2006.

[3] I. H. Witten, E. Frank, and M. A. Hall, Data Mining Practical Machine Learning Tools and Techniques 3rd Edition. Burlington: Elsevier, 2011.

[4] F. Gorunescu, Data Mining Concepts, Models and Techniques. Berlin: Springer, 2011.

[5] X. Wu and V. Kumar, The Top Ten Algorithms in Data Mining. New York: CRC Press, 2009.

[6] D. T. Larose, Discovering Knowledge in Data, An Introduction to Data Mining. New

Jersey: John Wiley & Sons, 2005.

[7] M. Bramer, Principles of Data Mining. London: Springer, 2007.

[8] J. Han and M. Kamber, Data Mining Concepts and Techniques 2nd edition. USA: Elseiver, 2006.

[9] C. Vercellis, Business Intelligence, Data Mining and Optimization for Decision Making. United Kingdom: John Wiley & Sons, 2009.

[10] O. Maimon and L. Rokach, Data Mining and Knowledge Discovery Handbook. New York: Springer, 2010.

[11] D. Sugiana. (2008, Jul.) danksugiana. [Online]. http://danksugiana.wordpress.com/2008/07/08/populasi-dan-teknik-sampling/

Author, a lecture at the AMIK BSI Jakarta in Information Management courses. She's got a Master Komputer degree (M.Kom) of Computer Science courses in STMIK Nusa Mandiri Jakarta. Author concentrates on research in the field of data mining and software engineering