Classification of Data Mining with Adaboost Method in Determining Credit Providing for Customers

I Gusti Ngurah Agung Surya Mahendra[1*], Ida Bagus Leo Mahadya Suta[2], and Made Sudarma [3]

[1,2]Department of Electrical and Computer Engineering, Post Graduate Program, Udayana University

[3]Department of Electrical and Computer Engineering, Udayana University

[*]maha.gung.surya@gmail.com

Abstract Credit is the provision of funds for lending and borrowing transactions with the agreement and agreement between the bank or financial institution and its customers, and requires the borrower to pay the debt within a certain period of time and provide services. Crediting is done by identifying and assessing factors that influence credit risk. The loss of income and the threat of profitability are things that need to be wary of lending. Data mining classification can be used to help credit analysts in determining lending to customers. The classification process is carried out to obtain determinant attributes. The results of the classification process are evaluated using the adaboost method and testing using weka to obtain cross validation, confusion matrix to determine the most accurate classification in determining credit for cooperative customers.

***Index Terms*—Classification, Data Mining, Cross Validation, Confusion Matrix, Credit, Adaboost**

I. Introduction

BANK and financial institutions play a strategic role in national development. Business entities that collect funds from the public in the form of deposits and distribute them to the public in the form of loans or other forms in order to improve the lives of many people. According to Article 1 number 11 of Act Number 10 of 1998, credit is the provision of money or equivalent claims, based on an agreement or agreement to borrow or borrow from a bank or financial institution with another party that requires the borrower to repay the debt after a certain period of time with the amount of interest.

Banking regulations and regulations change with the implementation of internet banking. All customer data related to current and problem loans can be seen and verified through internet banking. So that credit risk can be reduced. On the other hand financial institutions such as cooperatives outside the banking sector do not yet have a data center, this can increase credit risk that threatens profitability. Cooperatives are a family business with the aim of prospering its members (Article 33 paragraph 1 of the 1945 Constitution). Cooperatives have different policies in granting credit. But generally credit is influenced by several factors such as trust, agreement, time period, risk and remuneration [1]. Credit analysts need to identify and assess the factors that can affect customers in returning credit (Costa et al., 2007).

Accurate measurement and good management ability in the face of credit risk is an effort to save economic operations units and benefit from a stable and healthy financial system as a whole and sustainable economic development [2]. Failure to identify credit risk leading to loss of income and expanding credit for bad credit risk is a threat to profitability [3]. Credit analysis errors can lead to credit risks, such as the disappearance of customers, uncertainty in the payment of loan funds, and even the inability of customers to repay loans. To protect credit funds, guarantees must be provided by the customer as a burden on the customer. Giving credit with collateral can be in the form of: collateral for tangible objects (land, buildings, motorized vehicles, gardens, jewelry, etc.), intangible guarantees (land certificates, share certificates, bond certificates, employment appointments, etc.) and collateral person (guarantee given by someone who states the ability to bear all risks if the credit is stuck).

Credit assessment criteria such as the nature or character of a person, ability to pay, use of funds, social, economic and political conditions and the collateral proposed are needed to provide information about good faith and the ability to pay for a customer [1]. Components that affect credit risk are the possibility that the debtor will fail to pay

(p-issn: 2579-5988, e-issn: 2579-597X)

in fulfilling the contract of payment, the claim that will be borne by the debtor if it does not meet the obligation to pay and the nominal lost due to the risk of default or default. Based on the above problems, the authors conducted a research on the feasibility of lending in the XYZ cooperative by using the boosting method, Adaboost. Adaboost method used with the aim of this method can improve accuracy in the process of classification and prediction.

II. Literature Review

A. Data Mining

Data mining is the process of analyzing data in different angles and summarizing results it into useful information. Data mining software is analytical tools which allow the users to analyze data from many different dimensions, categorize the data, and summarize the relationships among data. Technically, data mining is the process of finding correlations among many numbers of fields in huge dataset. Five major elements of data mining are: Select, Transform and store data onto the data warehouse system. Keep and handle the data in a multidimensional database system. Allow business analysts and information technology professionals to access the data. Use application software to analyze the data. Display the data in a human understandable format, such as a graph or table.

The development of Information Technology has generated large amount of databases and huge data in various areas. The research in databases and information technology has given rise to an approach to store and manipulate this precious data for further decision making. Data mining is a process of extraction of useful information and patterns from huge data. It is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data /pattern analysis.

Data mining is a logical process that is used to search through large amount of data in order to find useful data. The goal of this technique is to find patterns that were previously unknown. Once these patterns are found they can further be used to make certain decisions for development of their businesses. Three steps involved are.

- Exploration
  In the first step of data exploration data is cleaned and transformed into another form, and important variables and then nature of data based on the problem are determined.
- Pattern identification
  Once data is explored, refined and defined for the specific variables the second step is to form pattern identification. Identify and choose the patterns which make the best prediction.
- Deployment
  Patterns are deployed for desired outcome.

B. Data Mining Techniques and Algorithm

Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases.

1. *Classification*
   Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit- risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier. Types of classification models*:* Classification by decision tree induction, Bayesian Classification, Neural Networks, Support Vector Machines (SVM), Classification Based on Associations.
2. *Clustering*
   Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification. For example, to form group of customers based on purchasing patterns, to categories genes with similar functionality. Types of clustering methods: Partitioning Methods, Hierarchical Agglomerative (divisive) methods, Density based methods, Grid-based methods, Model based methods.
3. *Predication*
   Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict.

Unfortunately, many real-world problems are not simply prediction. For instance, sales volumes, stock prices, and product failure rates are all very difficult to predict because they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models. Types of regression methods: Linear Regression, Multivariate Linear Regression, Nonlinear Regression, Multivariate Nonlinear Regression.

4. *Association rule*

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value. Types of association rule: Multilevel association rule, Multidimensional association rule, Quantitative association rule.

5. *Neural networks*

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. For example handwritten character reorganization, for training a computer to pronounce English text and many real world business problems and have already been successfully applied in many industries. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs. Types of neural networks: Back Propagation

C. Adaboost

The Adaboost algorithm [4] proposed by Yoav Freund and Robert Schapire is one of the most important ensemble methods, since it has solid theoretical foundation, very accurate prediction, great simplicity (Schapire said it needs only “just 10 lines of code”), and wide and successful applications.

III. METHODOLOGY

In this research, the stages of the data collection process are carried out after the process of collecting data is continued with the initial data management process. The result of the following is the process of designing adaptive boosting systems. The entire stages of this research process can be seen in the process of research methodology.

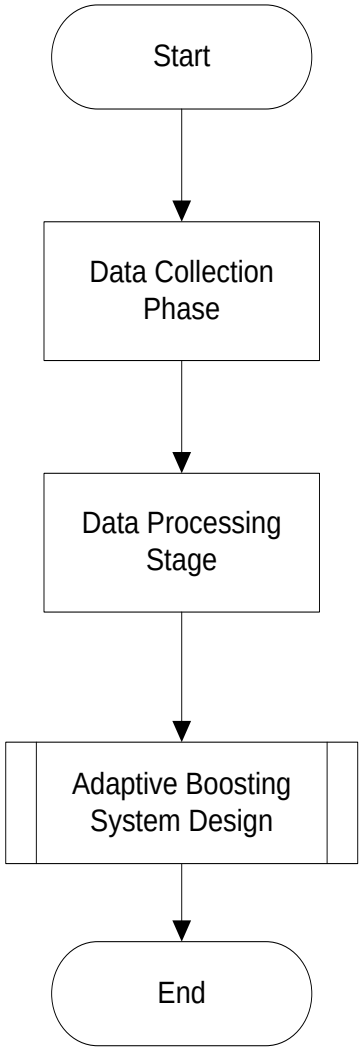

Fig.1 Research Methods

Development of the adaptive boosting system design process by reading the dataset and checking the type model if the training is in the probability number process and probing the probability table so as to produce lacar credit status or not, and vice versa if the model type is not taining so the results are whether the credit status is problematic or not.

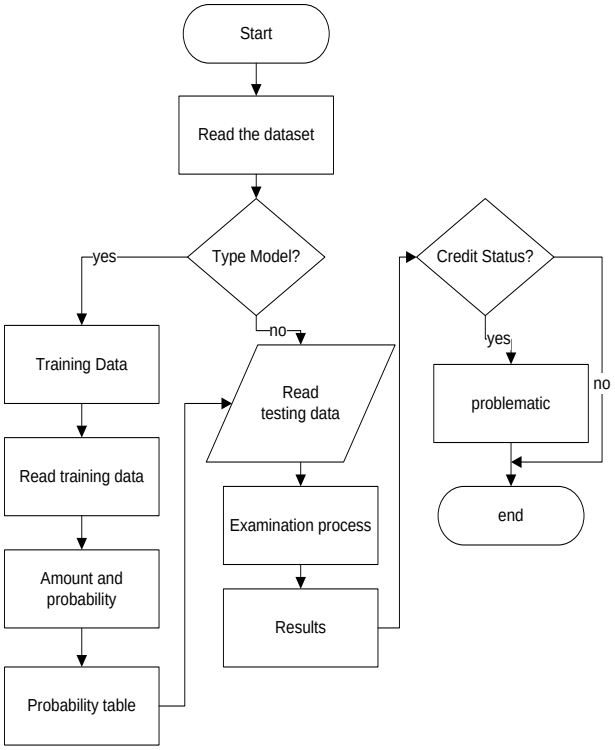

Fig.2 Adaptive Boosting system design

IV. Results and Discussion

A. Data Collection

Determining the type and source of data to obtain data that is truly accurate is very important. The data source in this study is credit data taken from the XYZ Cooperative in previous years as a reference to find certain patterns that can be used as determinant attributes. Data that can be used in this study are collateral data, loan data and current account data.

TABLE I
COLLATERAL DATA

No. Collateral	Date In	Customer Name	Insurance item
185	2/2/2018	Janur Ratnasari	BPKB Honda DK5896BN
186	8/2/2018	Putri Fridayanti	BPKB Honda DK5896BN
187	9/2/2018	Sukarji Sutejo	BPKB Honda DK5896BN
188	14/2/2018	Sudarmono	BPKB Honda DK5896BN
189	18/2/2018	Sri Wahyuni	BPKB Honda DK5896BN
190	24/2/2018	Sarimin Sukaji	BPKB Honda DK5896BN
...			

Table 1 contains the name of the customer and the item that is used as collateral. Table 2 contains the name of the customer, the start date of the loan, the amount borrowed, the arrears value and the statement has matured or is over due.

TABLE II
Collateral Data

Customer Name	date of loan	total loan	Arrears	Des.
Suyono	25/4/2015	5.000.000	1.000.000	Due date
Surya Pro	22/6/2016	5.000.000	3.390.000	Due date
Sri Purwanto	19/1/2017	45.000.000	30.000.000	Due date
Fitria Roro	31/5/2017	9.000.000	4.250.000	3month
Nurisyah	18/8/2017	4.000.000	1.042.000	5month
Sugiantoro	22/9/2017	5.000.000	1.100.000	4 month
....				

Table 3 contains the name of the customer, the person in charge, the number of loans, the basic contributions that must be paid, the amount of installments, services (interest) that must be paid and due date of payment.

TABLE III
Customer Loan Data

Customer Name	Amount of loan *	principal fee *	X	Services*	due date of payment
Suharti	2.500.000	208.000	12	55.000	Feb
Agus Jaya	4.500.000	166.000	24	88.000	Mar
Anang F	7.000.000	350.000	10	100.000	Apr
Mujiono	50.000.000	1.166.000	12	77.000	May
Surya M	2.500.000	1.240.000	3	1.100.000	Jun
Fitria R	5.000.000	208.000	24	100.000	Jul
....					

*) in thousands

B. Initial Data Processing

Data processing is needed to prepare data that is truly valid before being processed. Processing is done with clean double data, equalize data boundaries, group data, perform feature selection and data pre-processing [6].

C. Data Integration

Data that can be used in the process of determining credit are current accounts receivable data, collateral data and loan data. Data integration is a way of combining several data from different tables by looking at the similarity of data based on key attributes (primary key), guest attributes (foreign keys) to see functional dependency. Data integration is needed because feature selection is needed to get a pattern that refers to the results of lending.

D. *Feature selection (attribute)*

Feature selection is done by taking a portion of the variables on all the attributes that are there to be a decisive attribute in making a decision. The features taken are attributes that have functional dependency properties and are part of the *super key*. The following are the results of feature selection:

TABLE IV
Feature Selection

Attribute	Value	Category
Gender	1	Man
	2	Women
Collateral	1	Motorcycle
	2	Car
	3	Building
	4	Jamsostek
	5	There is no
Loan Amount	1	<= 5.000.000
	2	<= 15.000.000
	3	> 15.000.000
Time period	1	Short (<= 6 bulan)
	2	Intermediate (<= 12 bulan)
	3	Long (> 12 bulan)
Status Kredit	1	Good
	2	Problematic

E. *Cleansing data*

The cleansing process is an important step, where data is cleared from unnecessary data (such as: member number, name, address) and deletes the same data (redundancy). This matter intended to maintain the value of functional dependence.

F. *Data transformation*

In the process of transforming data, data are grouped according to the same criteria to facilitate further processing of data, which can be seen in Table V.

TABLE V
The Results of The Data Pre-Processing Process

Gender	Collateral	Loan Amount	Time period	Credit Status
Women	Motor	2.500.000	Short	Good
Man	Mobil	5.000.000	Intermediate	Problematic
Man	Bangunan	15.000.000	Intermediate	Good
Women	Motor	3.000.000	Short	Good
Women	Mobil	7.000.000	Long	Problematic

G. *The 10-fold cross validation process*

This test method follows the measurement method by measuring the level of accuracy of each algorithm based on the credit data set which is divided into decision-making variables. From the results of pre-processing data, there are 602 credit data with a total of 404 data customers who have no problems and 198 customer data having problems in paying credit.

In cross-validation, there will be a choice of how many folds to use. The default value is 10. The mechanism is as follows: Training data is divided into k subset (subset). Where k is the value of the fold. Furthermore, for each of the subset, the test data from the classification results will be made from the other k-1 subsets. So, there will be 10 tests. Where, every datum will become test data 1 time, and become training data as many as 1 times. Then, the error of the test will be averaged.

The first process is to compare the testing process using 10-fold cross validation. Based on the process of cross validation, the level of accuracy, the results of the confusion matrix in Table VI and Fig.4 .

TABLE VI
The Results of Cross Validation, The Level of Accuracy, The Results of The Confusion Matrix

	Accuracy Adaboost
Result	65.12%

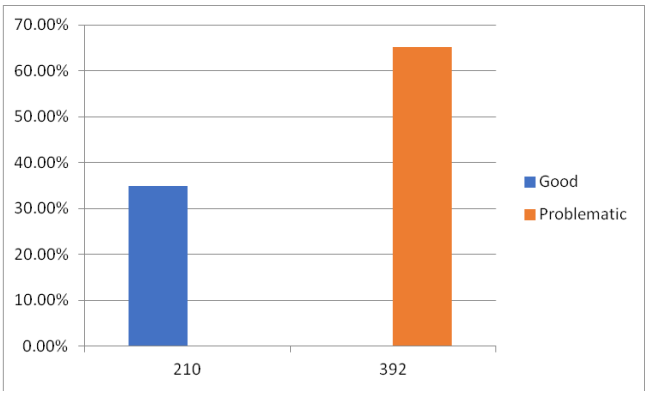

Fig.3 Cross Validation, The Level of Accuracy, The Results of The Confusion Matrix

Based on the results of the analysis, the confusion matrix obtained was 65.12%.

V. Conclusion

The results of comparative analysis using cross validation, confusion matrix, in the data mining classification adaboost algorithm it can be concluded that the algorithm used is not accurate, because it has the highest accuracy value of 65.1163%. Based on these results it can still be used as a review to combine tree algorithms, naive bayes and artificial intelligence to produce better accuracy.

References

[1] Kasmir, "Dasar-Dasar Perbankan," 1st ed., P. R. G. Persada, Ed. Jakarta.
[2] Y. Ma, H. & Guo, "Credit Risk Evaluation Based on Artificial

Intelligence Technology," 2010.
[3] J. Zurada, "Could Decission Trees Improve the Classification Accuracy and Interpretability of Loan Granting Decission," 2010.
[4] S. R. Freund Y, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 1, pp. 119–139, 1997.
[5] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.
[6] F. Gorunescu, *Data Mining: Concepts, Model and Techniques*. Berlin: Springer.