Optimization of Feature Selection Using Genetic Algorithm with Naïve Bayes Classification for Home Improvement Recipients

Luh Gede Putri Suardani[1*], I Made Adi Bhaskara[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
*Email: putrisuardani94@gmail.com

***Abstract*— The purpose of this study is to help predicting the recipients of home improvement based on the optimized criterias. The classification method is using naïve bayes method. Naïve Bayes classification is one of data mining classification technique that can predict future probabilities based on past experience. But, naïve bayes method has a disadvantage that the independence characteristics of Naïve Bayes feature can not always be applied so it will affect the accuracy. Due to the independence characteristics, Naïve Bayes classification method needs to be optimized by feature selection technique. Genetic algorithm are one of the most commonly used methods of feature selection techniques. The results achieved is Naïve Bayes Method can be applied to determine the recipients of home improvement, which seeks the greatest opportunity or alternative probability by exploiting the conditional probability of each criterion that has been optimized. This can be used as consideration, reference, and facilitate in determining the community welfare and the program rolled right on target.**

***Key words*—Naïve Bayes, Classification, Optimization, Genetic Algorithm**

I. INTRODUCTION

Based on data from BPS Bali Province, the number of poor people has increased in the last three years. The percentage of poor families in Bali Province in 2013 is 4.49%. Klungkung Regency has the highest percentage of poor people, it is 7.01% and the lowest is Denpasar which is 2.07%. The total number of poor people in Bali Province is 1.828.000 population.

Bali's Government has been gradually trying to reduce the number of poor people through strategies that are carried out with the implementation of programs. In general, poverty alleviation programs undertaken by the government have two objectives. There are charged to other parties such as the government or society and increased the income of the poor so they can get out of poverty. One of the excellent programs in Bali Mandara program is Bali Mandara Home Improvement program which is consumptive program along with Askescat, JKBM, Jamkesmas, Sembako and Raskin programs.

The problem that arises when the aid program has been initiated by the government is the determination of recipients who are sometimes not well targeted. One of the causes of this problem is the difficulty for deciding the recipient based on predetermined criteria.

Naïve Bayes classification method is one of data mining classification techniques. Naïve Bayes classification can predict future probabilities based on past experience so as to determine future consumer credit risk based on the experience of previous customers.

According to Socrates et al (2016), the advantage of this method is a simple algorithm with low computation complexity. However, there is a disadvantage that the independence characteristics of Naïve Bayes feature can not always be applied so that it will affect the accuracy of the calculation.

Due to the independence of Naïve Bayes classification, it needs to be optimized with feature selection techniques. Feature selection is an important step in the classification process. This process analyzes the features to produce features that play a role or less play a role in the classification process. Genetic algorithms are one of the most commonly used methods of feature selection techniques.

II. LITERATURE STUDY

The literature review of this journal is as follows.

A. Bali Mandara's Home Improvement Program

Home improvement program is one of the efforts to accelerate poverty reduction in Bali Province which aims to make poor families have adequate housing to fulfill their basic needs.

Home improvement program has the following benefits as follows: a habitable home is a basic need that every family must meet. Habitable house is identic with healthy dwelling that will provide a safe and comfortable atmosphere for its inhabitants, and also improve the health of its inhabitants. With good health, people will be able to work more optimally so that the welfare of the community is expected to increase. And provide better conditions for school children to learn.

B. Data Mining

Data Mining is the process of extracting information from a very large set of data through the use of algorithms and withdrawal techniques in the field of statistics, machine

learning and database management systems. Data mining is the process of analyzing data from different perspectives and summarizing it as important information that can be used to increase profits, minimize cost of spending, or even both. Another definition says Data Mining is an activity that includes the collection, use of historical data to find regularities, patterns or relationships in large data. From some of the above definition can be deduced that Data Mining is a process or activity to collect large data and then extracting the data into information that will be used.

C. Stages of Data Mining

As a series of processes, Data Mining can be divided into several stages of the process. The stages are interactive, the user is directly involved or with the intermediary knowledge base.

The first stage of data mining is data cleaning. Data cleaning is a process of eliminating noise and inconsistent data or irrelevant data. Furthermore, the integration of data that is merging data from various databases into a new database. Then do the data selection. The data that is in the database is often not all used, therefore only the appropriate data to be analyzed to be retrieved from the database. After that do data transformation. Data is altered or merged into the appropriate format for processing in Data Mining. Then doing the mining process is a major process when the method is applied to find valuable and hidden knowledge of the data. Some methods can be used based on Data Mining grouping. Then, a pattern evaluation was performed to identify interesting patterns into the found knowledge base. Finally, the presentation of knowledge is done for the visualization and presentation of knowledge of the methods used to acquire the knowledge obtained by the user.

D. Genetic Algorithm

The Genetic Algorithm is a heuristic method developed on the basis of genetic principles and the natural selection process of Darwin's Evolutionary Theory. The optimization method was developed by John Hollan around the 1960s and popularized by one of his students, David Goldberg, in the 19800s (Haupt and Haupt, 2004) in (Zukhri, 2014).

Haupt and Haupt (2004) mentioned that the basic structure of Genetic Algorithm consists of several steps, namely population initialization, population evaluation, population selection to be subjected to genetic operators, the process of crossing certain chromosome pairs, the process of a particular chromosome mutation, a new population evaluation, repeat from step 3 as long as the stop condition has not been met.

E. Naïve Bayes Method

Naive Bayes is a simple probabilistic classifier that computes a set of probabilities by summing the frequency and value combinations of the given dataset. The algorithm uses the Bayes theorem and assumes all the independent or non-dependent attributes given by the value of the class variable. Another definition says Naive Bayes is a classification with the probability and statistical methods brought by British scientist Thomas Bayes, predicting future opportunities based on past experience. Naive Bayes is based on the simplifying assumption that attribute values are conditionally independent if given output value. In other words, given the value of output, the probability of observing collectively is the product of the individual probability. The advantage of using Naive Bayes is that this method requires only a small amount of training data to determine the estimated parameters required in the classification process. Naive Bayes often work much better in most real-world situations that are complex than expected. The equation of Bayes's theorem is:

$$P(H|X) = \frac{P(X|H).P(H)}{P(X)} \quad (1)$$

Where X is the data with the unknown class. H is the data hypothesis of a specific class. P (H | X) is the probability of hypothesis H by condition X (posteriori probabilitas). P (H) is the probability of hypothesis H (prior probability). P (X | H) is the probability of X based on the condition on hypothesis H. P (X) is probability X.

To explain the Naive Bayes method, please note that the classification process requires a number of clues to determine what class is suitable for the analyzed sample. Therefore, the method of Naive Bayes above is adjusted as follows:

$$P(C|F1 \dots Fn) = \frac{P(C)P(F1 \dots Fn|C)}{P(F1 \dots Fn)} \quad (2)$$

Where Variable C represents the class, while the variable F1 ... Fn represents the characteristics of the instructions needed to perform the classification. Then the formula explains that the probability of entering a sample of certain characteristics in class C (Posterior) is the probability of the emergence of class C (before the entry of the sample, often called prior), multiplied by the probability of occurrence of sample characteristics in class C (also called likelihood) with the probability of occurrence of sample characteristics globally (also called evidence). Therefore, the above formula can also be written simply as follows:

$$Posterior = \frac{prior\ x\ likelihood}{evidence} \quad (3)$$

Evidence values are always fixed for each class on a single sample. The value of the posterior will then be compared with the posterior values of the other classes to determine to which class a sample will be classified. A further description of the Bayes formula is done by elaborating ($C \mid F1, \dots, Fn$) using the rules of multiplication as follows:

$$P(C|F1 \dots Fn) = P(C)P(F1 \dots Fn|C) \quad (4)$$
$$= P(C)P(F2, \dots, Fn|C, F1)$$

III. Analysis and Design

A. Data and Criteria for Determination of Home Improvement

The data of candidates were obtained at the Social Service of Tabanan Regency, Bali Province. The criteria used in this system are building area Criteria (K1), Floor Material Criteria (K2), Wall Material Criteria (K3), Roof Material Criteria (K4), Criteria of Lighting Type (K5), Drinking Water Source Criteria (K6), Cooking Fuel Criterion (K7), Toilet Availability Criteria (K8), Criteria of Asset Ownership / Savings / Valuables (K9), Monthly Income Criteria (K10), Criteria of Ability to Buy Apparel Per Year (K11), and Criteria for Medical Treatment (K12).

IV. Discussion

Stages of genetic algorithm starts from initializing the population by generating random numbers of binaries as much as the number of training data, so on this issue use the encoding in binary numbers. Table 1 shows an example of one of the chromosomes that has been raised.

TABLE I. Comparison Matrix Table

Kromosom i	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12
1	0	1	0	1	1	1	1	0	1	1	1	0

The next stage of the genetic algorithm is to conduct a population evaluation process. The probability of crossover is 0.65 and the probability of mutation is 0.1.

Evaluate the population formed by calculating the fitness value by using the objective function established in the case study. The value of the fitness of the chromosome i is the sum of the weight product on the gene n and the gene binary value of n on the chromosome i in the initial population. Where the weight value is determined as follows.

TABLE II. Table Weight Value

i	Gen	Weigh t	Initial Value	Sum of i (Weight i x Initial Value n)	Weight (B) Sum of i : many n initializations on i
1	Building Area	1	1	1	2,5
			2	2	
			3	3	
			4	4	
			Sum of i	10	
2	Floor Materials	2	1	2	4
			2	4	
			3	6	
			Sum of i	12	
3	Wall Materials	4	1	4	6
			2	8	
			Sum of i	12	
4	Roofing Materials	3	1	3	4,5
			2	6	
			Sum of i	9	
5	Lighting Type	5	1	5	12,5
			2	10	
			3	15	
			4	20	
			Sum of i	50	
6	Drinking Water Source	6	1	6	9
			2	12	
			Sum of i	18	
7	Cooking Fuel	7	1	7	10,5
			2	14	
			Sum of i	21	
8	Toilet Availability	9	1	9	13,5
			2	18	
			Sum of i	27	
9	Asset Ownership/Savings/Valuables	10	1	10	15
			2	20	
			Sum of i	30	
10	Monthly Income	8	1	8	16
			2	16	
			3	24	
			Sum of i	48	
11	Ability to Buy Apparel Per Year	12	1	12	18
			2	24	
			Sum of i	36	
12	Medical Treatment	11	1	11	16,5
			2	22	
			Sum of i	33	

After calculating the fitness value then conducted the process of population selection using the selection method is proportional to the value of fitness (fitness), this method is implemented with the roulette model.

The method of crossing used in this study is the one-point crossing method. First pick up a random number as much as the number of chromosomes in the population. Choose a number less than the predetermined Crossover Probability (PC) value at which, the value of PC = 0.65. Selecting a random number

again from 0 to the length of the chromosome / gen-1 is called a cross-cut crossover position where it will determine the position of the gene to be crossed between the selected parent chromosomes.

The mutation process will initiate a random switch of 1 gene with a new value. After passing through the mutation process, a new population has been generated called first generation, if it has been determined that the maximum generation total = 10 generations, the genetic algorithm process will stop The process of genetic algorithm is complete and stopped when it reaches 10 generations by producing the highest fitness value on chromosome 5 which are selected for the testing process are: Building Criteria, Criteria of Floor Material, Criteria of Wall Materials, Roof Criteria, Criteria Type of Lighting.

TABLE III. TRAINING TABLES OF OPTIMIZED FEATURES

No	K1	K2	K3	K4	K5	Status
1	3,8	Soil	Bambu	Roof Tile	No Electricity	Accepted
2	4	Cement	Bambu	Tin Roof	900 VA	Not Accepted
3	3,8	Soil	Batako	Roof Tile	No Electricity	Accepted
4	3,8	Soil	Batako	Tin Roof	900 VA	Not Accepted
5	3,5	Soil	Batako	Roof Tile	900 VA	Not Accepted
6	3,6	Semen	Batako	Tin Roof	900 VA	Not Accepted
7	3,7	Semen	Bambu	Roof Tile	No Electricity	Accepted
8	3,8	Soil	Bambu	Tin Roof	No Electricity	Accepted
9	4,2	Soil	Batako	Roof Tile	No Electricity	Accepted
10	4	Soil	Bambu	Tin Roof	900 VA	Accepted
11	2,5	Semen	Batako	Roof Tile	900 VA	Not Accepted
12	2	Semen	Bambu	Tin Roof	900 VA	Not Accepted
13	2,7	Soil	Bambu	Roof Tile	No Electricity	Accepted
14	2,8	Soil	Batako	Tin Roof	900 VA	Not Accepted
15	2,9	Semen	Bambu	Roof Tile	No Electricity	Accepted

Table III explains about 15 training data which feature has been optimized. Table IV is a test table that will predict its acceptance status using Naive Bayes.

TABLE IV. TESTING TABLE

3,5	Cement	Bamboo	Roof Tile	900 VA	???

The first step is to count the number of classes or labels. P (Y = Accepted) is the Number of Data "Received" divided by the Number of Data. Obtained P (Y = Accepted) is 8/15. P (Y = Not Received) is the Number of Data "Not Received" divided by the Number of Data. Obtained P (Y = Not Received) is 7/15. The second stage is to count the same number of cases with the same class.

- P (K1 = 3.5 | Y = Accepted) = total data K1 = 3.5 and Y = Accept divided by amount of data Accepted = 0/8
- P (K1 = 3.5 | Y = Not Received) = total data K1 = 3.5 and Y = Not Accepted divided by the amount of data Not Received = 1/7
- P (K2 = Cement | Y = Accepted) = total data K1 = 3.5 and Y = Accept divided by amount of data Accepted = 2/8
- P (K2 = Cement | Y = Not Received) = total data K1 = 3.5 and Y = Accept divided by the amount of data Accepted = 4/7
- P (K3 = Bamboo | Y = Accepted) = total data K1 = 3.5 and Y = Accept divided by amount of data Accepted = 6/8
- P (K3 = Bamboo | Y = Not Received) = total data K1 = 3.5 and Y = Accept divided by amount of data Accepted = 2/7
- P (K4 = Tile | Y = Accepted) = total data K1 = 3.5 and Y = Accept divided by amount of data Accepted = 6/8
- P (K4 = Tile | Y = Not Received) = total data K1 = 3.5 and Y = Accept divided by amount of data Accepted = 2/7
- P (K5 = Power Electric 900 VA | Y = Accepted) = total data K1 = 3.5 and Y = Accept divided by amount of data Accepted = P
- P (K5 = Power Electric 900 VA | Y = Not Received) = total data K1 = 3.5 and Y = Accept divided by amount of data Accepted = 7/7

The next stage is multiplying All Results Variable Accepted Not Accepted

- P (K1 = 3,5), (K2 = Cement), (K3 = Bamboo), (K4 = Roof Tile), (K5 = Power Electric 900 VA) | accepted) = {K1 = 3.5 | Y = Accepted). (K2 = Cement | Y = Accepted). (K3 = Bamboo | Y = Accepted). (K4 = Roof Tile | Y = Accepted). (K5 = Electricity Power 900 VA | Y = Accepted)}= 0/8 . 2/8 . 6/8 . 6/8 . 1/8 = 0
- P (K1 = 3,5), (K2 = Cement), (K3 = Bamboo), (K4 = Roof Tile), (K5 = Power Electric 900 VA) | Not accepted)
  = (K3 = Bamboo | Y = Not Accepted). (K4 = Tile | Y = Not Accepted). (K4 = C = Y = Not Received). . (K5 = Power Electric 900 VA | Y = Not Received)}
  = 1/7. 4/7. 2/7. 2/7. 7/7 = 0,006529

The last stage is to compare the class results Accepted Not Accepted. Since the result (P | No Accepted) is greater than (P | Accepted) then the decision is Not Accepted

V. KESIMPULAN

Naïve Bayes method can be applied in the determination of a home improvement recipient, which seeks the greatest opportunity or probability of an alternative by exploiting the conditional probability of each criterion optimized with the Genetic Algorithm that can be used as consideration, reference and simplify in determining the welfare of the community and programs that are scored right on target.

REFERENCES

[1] Wulan, S. T., UMRAH, F., Bettiza, M., Si, S., & Nurul Hayaty, S. T. 2017. Selection of Naïve Bayes Classification Features Using Genetic Algorithm for Consumer Credit Risk Prediction.

[2] Iskandar, D., & Suprapto, Y. K. 2013. The average accuracy level between C4 algorithm. 5 and Naïve Bayes Classifier. JAVA Journal of Electrical and Electronics Engineering, 11 (1).

[3] Witary, V., Rachmat, N., & Inayatullah, I. 2013. Optimizing Lecturing Scheduling Using Genetic Algorithms (Case Study: AMIK MDP, STMIK GI MDP and STIE MDP).

[4] Maricar, M. A., Wahyudin, W. dan Sudarma, M. 2016. Decision Support System of the Employees Acceptance using Analytical Hierarchy Process (AHP) and Multi Factor Evaluation Process (MFEP) dalam International Journal of Engineering and Emerging Technology Vol.1 No. 1, pp.48-54, July 2016

[5] Wati, M., & Hadi, A. 2017. Implementation of the Bayesian Naive Algorithm in Determining the Beneficiaries of the Government Aid Program. STMIK KHARISMA Makassar, 3 (1), 22-26.

[6] Simatupang, F. J., Wuryandari, T., & Suparti, S. 2016. Classification of Houses Worth Living in Brebes District By Using Learning Vector Quantization And Naive Bayes Methods. Gaussian Journal, 5 (1), 99-111.

[7] Setiawati, W. 2015. Course Scheduling System Based on Student Activity Using Association Methods Improved with Genetic Algorithm. Journal of Informatics Dian Nuswantoro University Semarang.

[8] Optimization of Naïve Bayes Algorithms by Using Genetic Algorithms for Fertility Predictions. Evolution-Journal of AMIK Science and Management BSI Purwokerto, 4 (1).

[9] Thomas, S. L. dan Luis, Vargas G. “Models, Methods, Concepts & Applications of The Analytic Hierarchy Process,”. International Series in Operations Research & Management Science. Second Edition. 2012.

[10] Turban, Efraim dan Aronson, J.E. 2004. Decision Support Systems and Intelligent Systems (11 ed). Yogyakarta: Penerbit Andi

[11] Chen, J. et al., 2009. Expert Systems with Applications Feature selection for text classification with Naïve Bayes. Expert Systems with Applications,36(3), pp.5432–5435.