ISSN 2714-9714

Optimization of Naïve Bayes Method using Genetic Algorithm to Diagnose Cattle Disease

Safit Firmansyah [1*], Endang Sugiharti [1], Riza Arifudin [1]

[1]Department Of Computer Science, Faculty of Mathematics and Natural Sciences, Universitas Negeri Semarang, Semarang. Indonesia
*Corresponding author: firmansyahsafit@gmail.com

ARTICLE INFO

Article history
Received 19 February 2020
Revised 5 March 2020
Accepted 6 April 2020

Keywords
Expert System
Naïve Bayes
Genetic Algorithm
Cow Disease

ABSTRACT

Technological advances that so fast encourage people to create a breakthrough, one of which is the expert system. An expert system can be used to solve problems in diagnosing disease problems, one of which is cattle disease. Lack of knowledge of breeders regarding this can result in considerable losses to the breeder. An expert system is needed to diagnose cow disease. The method that can be used to create an expert system is the Naïve Bayes method. The naïve Bayes method is a classification with probability methods and statistics to predict future opportunities based on previous experience. But there are weaknesses; namely, the independence feature is often wrong, and the probability estimation results cannot be optimal. To overcome this problem, one way is to use Genetic Algorithms. Genetic algorithms are random search forms that mimic the principle of natural biological evolution processes to find optimal solutions. The number of attributes used is 24 attributes consisting of the name of the disease and 23 symptoms. The accuracy of using the Naïve Bayes method is 90%, while the accuracy of using the Naïve Bayes and Genetic Algorithm is 95%. It can be concluded that there is an accuracy increase of 5%.

1 Introduction

The use of a new system that is more practical and fast in its services and can provide assistance to users begins to develop quite rapidly. One branch of computer science that can help human performance is an expert system. According to Pramesti, Arifudin, and Sugiharti (2016), expert systems, namely artificial intelligence programs that combine knowledge bases and inference machines. In the application of expert systems, experts' problem is not just an algorithm problem but sometimes a problem that is difficult to understand (Setiabudi, Sugiharti, & Arini, 2017). Expert systems that are sub-fields of artificial intelligence are also designed to mimic an expert's abilities such as in the fields of health, plantations, agriculture, and so on. So that with existing capabilities, the expert system can be used to solve problems in diagnosing disease problems, one of which is a disease in cow disease. The lack of knowledge of breeders regarding diseases that attack their livestock and handling solutions to these diseases can considerably damage the breeder.

The Naïve Bayes method is a classification with probability and statistical methods to predict future opportunities based on previous experience (Sugiharti, Firmansyah, & Devi, 2017). In the Naïve Bayes calculation, there are 3 calculation steps, namely (1) calculating the prior value, (2) calculating the likelihood value, and (3) calculating the posterior value. The results of class classification using the Naive Bayes method are done by comparing the existing classes' posterior values. The highest posterior value was selected as a result of classification. But in its application, the Naïve Bayes method has weaknesses; namely, the attributes or independence features are often wrong, and the probability estimation results cannot run optimally. To overcome this, one of the ways is by attribute weighting method to improve the accuracy of the Naïve Bayes method, and later the weight of the attribute will be used to select influential and non-influential features or attributes. One method of weighting attributes or feature selection used is the Genetic Algorithm (GA).

GA is a random search form that mimics the principle of natural biological evolution processes to find optimal solutions. For a complex problem, this algorithm starts with a set of parameters called chromosomes or strings, then each of them is evaluated for its level of resilience by a predetermined objective function. In general, the GA consists of three stages, namely determining the initial population randomly, calculating each chromosome's fitness value, and applying genetic surgery to get a new alternative solution. Operators used in GA include (1) Selection, (2) Crossover, and (3) Mutations.

Based on this background, this study aims to optimize the naïve Bayes method's accuracy using GA and find out the workings of the combination of the naïve Bayes method with GA to diagnose diseases in cattle based on symptoms in cows disease.

2 Methods

2.1 Naïve Bayes

Naïve Bayes is a classification with probability and statistical methods proposed by British scientist Thomas Bayes, namely predicting opportunities in the future based on previous experience, so that it is known as the Bayes Theorem. Naïve Bayes for each decision class calculates the probability provided that the decision class is correct, given the vector information object. This algorithm assumes that the object attribute is independent. The probability involved in producing the final estimate is calculated as the number of frequencies from the "master" decision Table. The Naïve Bayesian method will calculate the probability in each case value of the target attribute in each sample data (Trihartati, Adi, 2016).

Bayes calculations can be done using the following steps.

1. Look for the prior value for each class by calculating the average of each class using Equation 1.

$$P = \frac{X}{A} \tag{1}$$

2. Look for the likelihood value for each class using Equation 2.

$$L = \frac{F}{B} \tag{2}$$

3. Look for the posterior value of each class that exists using Equation 3.

$$(H|E) = (H) \times (E|H) \tag{3}$$

The results of class classification using the Naive Bayes method are done by comparing the existing classes' posterior values. The highest posterior value was selected as a result of classification.

2.2 Genetic Algorithm (GA)

GA is a random search form that mimics the principle of natural biological evolution processes to find optimal solutions. For a complex problem, this algorithm starts with a set of parameters called chromosomes or strings, then each of them is evaluated for its level of resilience by a predetermined objective function.

In general, the GA consists of 3 stages, namely determining the initial population randomly, calculating each chromosome's fitness value, and applying genetic surgery to get a new alternative solution. In GA, fitness can usually be an objective function of the problem to be optimized. The chromosomes are selected according to their respective fitness values. Strong chromosomes have a high chance of surviving in the next generation, but weak chromosomes can survive. New chromosomes then determine the selection process through a crossover process and mutations of the selected chromosomes. From the two processes mentioned above, a new generation is formed, which will be repeated continuously until it reaches a convergence, which is as much as the desired generation (Aziz, 2014).

In this research, the method used for the selection process is the roulette wheel method. The roulette wheel is the simplest method and is often also known as stochastic sampling with replacement. This method selects one individual with possible proportions directly for its fitness value, namely choosing the best chromosome by calculating each chromosome value and comparing it with other chromosome values. Roulette wheel selection selects parents based on fitness value. Better chromosomes have larger chosen presentations. In this method, individuals are mapped in a line segment sequentially so that each individual segment has the same size as its fitness size. While the crossover process uses the one-point crossover method, the method used in the mutation process is the random mutation method.

The steps of GA (Ashari, Muslim, & Alamsyah, 2016) as follows:

1. Generating the initial population

The initial population is randomly generated so that the initial solution is obtained, consisting of a number of chromosomes that represent the desired solution.

2. Form a new generation

Reproductions/selection, crossover, and mutation operators are used. This process is carried out repeatedly so that the number of chromosomes is sufficient to form a new generation, where the new generation is a representation of new solutions. The new generation is known as the offspring.

3. Evaluate the solution

In each generation, chromosomes will go through an evaluation process using a measuring device called fitness. The fitness value of a chromosome describes the quality of the population. This process will evaluate each population by calculating each chromosome's fitness value and evaluating until the criteria are stopped. Before AG is done, there are two important things that must be done, namely the definition of chromosomes, is a symbolic solution, and the function of fitness or objective functions.

The operators contained in the genetic algorithm are as follows.

1. Selection

Selection aims to provide greater reproductive opportunities for members of the fittest population. The first step in the selection is the search for fitness values. Each individual in a selection container will accept the reproductive probability that depends on the object's value. The fitness value will be used in the next stage.

2. Crossover

Crossover aims to increase string diversity in one population by crossing strings obtained from previous reproductions. There are several types of crossover:

1. One-point Crossover
2. Two-point Crossover
3. Uniform crossover

The cross-breeding process is one of the most important components in AG. The presence of crosslinking causes this, and the resulting solution will converge at a certain point randomly, different from the conventional iteration method in the form of hill climbing. More details can be seen in Figure 1.

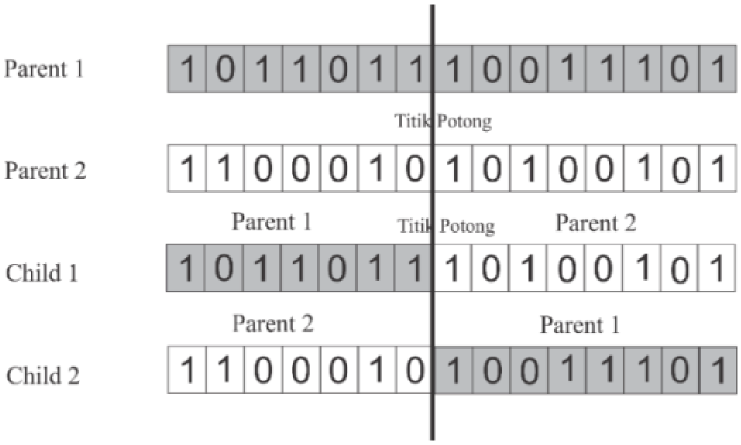

Figure 1. Crossover

3. Mutation

Mutations are the process of changing the value of one or several genes on a chromosome. Mutation operations aim to obtain new chromosomes as solution candidates for future generations with better fitness.

a. Mutations in binary coding
b. Mutations in permutation coding
c. Mutations in value coding
d. Mutations in tree coding

The process of mutation in GA here is also like genetic processes in general. In GA, the mutation process is expressed by replacing the gene's value affected by the mutation with the opposite value.

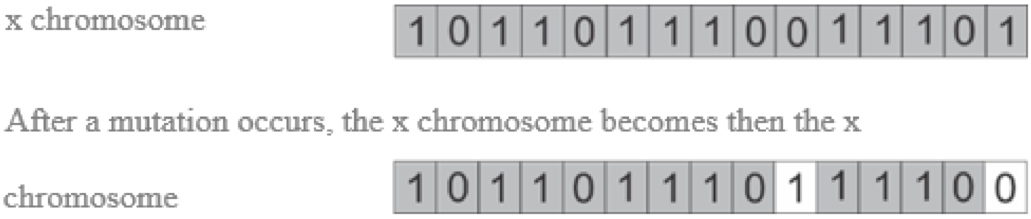

Figure 2. Mutation

On the X chromosome on the 10th gene and the 15th gene, there is a mutation, so that at first, the binary index of the X chromosome in the 10th gene that has a value of 0 changes to 1, and so to the 15th gene the binary index which is 1 changes to 0.

2.3 Waterfall Model

The development of the Naïve Bayes method and Genetic Algorithm as an expert system for diagnosing bovine disease with a case study in the Semarang District Agriculture, Fisheries and Food Service uses the Waterfall Model approach. The waterfall is an approach method based on the assumption that the main decision must be made before the coding begins (Vedayoko, Sugiharti, & Muslim, 2017). The reason the author uses the waterfall model is that this method has steps that are clear, real, and practical. The waterfall model is carried out when a new process phase is complete, it can proceed to the next stage, so if a process is not carried out, then the next process cannot be carried out. According to (Pressman, 2001), software development using the waterfall method can be seen in Figure 3.

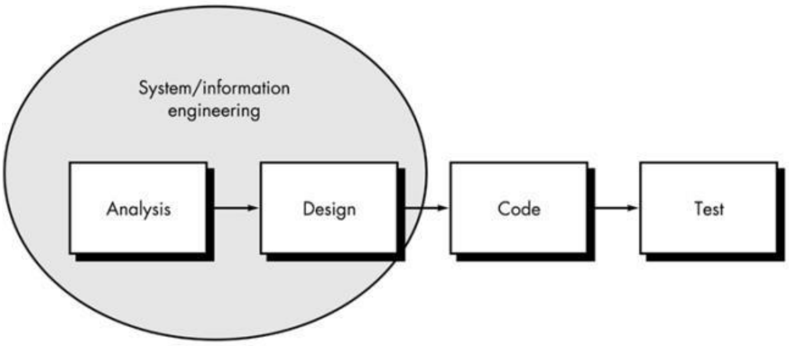

Figure 3. Waterfall Model

3 Results and Discussion

3.1 Results

The data processed is data on cow disease and each cow disease's symptoms that occur most in Semarang Regency. The data collection obtained seven cow diseases in Semarang Regency, which will later be used as training data, and 20 disease data from expert diagnosis to be used as test data. The list of diseases obtained can be seen in Table 1.

Table 1. List of diseases

Disease ID	Name of the disease
P1	BEF
P2	Helminthiasis
P3	Scabies
P4	Milk Fever
P5	Pink Eye
P6	Indigestion
P7	Bloat

Then to list the symptoms of each disease can be seen in Table 2

Table 2. Symptoms of the disease

Symptoms ID	List of symptoms	List of Diseases						
		P1	P2	P3	P4	P5	P6	P7
G1	Fever	v						
G2	Abstaining from Food	v					v	v
G3	Limp	v						
G4	Teary-Eyed	v				v		
G5	Excessive Saliva Out	v						
G6	Dull Fur		v					
G7	Alopecia		v	v				
G8	diarrhea		v					
:	:	:	:	:	:	:	:	:
G23	Abdominal Swell							v

In addition to obtaining disease and symptom data, the results of the interviews also obtained weights from each disease and symptom by giving weights according to the weight categories listed in Table 3.

Table 3. Weight category

Category	Weight
Clueless	0 - 0.1
Possibly	0.2 - 0.3
Potentially	0.4 - 0.5
Very likely	0.6 - 0.8
Definitely	0.9 – 1

In the calculation using the Naïve Bayes method, the results of the diagnosis of cattle in the system can be seen in Table 4.

Table 4. Comparison of diagnoses using naïve Bayes

Testing Data Id	Diagnosis of Naïve Bayes Calculations in the System	Expert diagnosis	Result
1	BEF	BEF	Relevant
2	Helminthiasis	Helminthiasis	Relevant
3	Scabies	Scabies	Relevant
4	Milk Fever	Milk Fever	Relevant
5	Pink Eye	Pink Eye	Relevant
6	Indigestion	Indigestion	Relevant
7	Bloat	Bloat	Relevant
8	BEF	BEF	Relevant
:	:	:	:
20	BEF	Indigestion	Irrelevant

The results of the comparison of the diagnosis above obtained the correct amount of data is 18, then the calculation of the accuracy value is.

Accuracy $= \frac{}{}$ x 100 %
$= 90\%$

So that the accuracy value obtained from the calculation of Naïve Bayes is 90%. While the calculation between the combination of Genetic Algorithm and Naïve Bayes in the system, the highest accuracy value is taken. The results of the diagnosis on the system can be seen in Table 5.

Table 5. Comparison of diagnoses using GA and Naïve Bayes

Testing Data Id	Diagnosis of Calculation of Genetic Algorithms and Naïve Bayes in the System	Expert diagnosis	Result
1	BEF	BEF	Relevant
2	Helminthiasis	Helminthiasis	Relevant
3	Scabies	Scabies	Relevant
4	Milk Fever	Milk Fever	Relevant
5	Pink Eye	Pink Eye	Relevant

6	Indigestion	Indigestion	Relevant
7	Bloat	Bloat	Relevant
8	BEF	BEF	Relevant
:	:	:	:
20	BEF	Indigestion	Irrelevant

The diagnosis comparison result above obtained the correct amount of data is 19, then the calculation of the accuracy value is.

Accuracy = x 100 %

= 95%

So that the accuracy value obtained from the calculation of Naïve Bayes is 95%.

3.2 Discussion

Trials of genetic and Naïve Bayes Algorithms were carried out with existing data, with training data of 7 data and test data as many as 20 data. In Naïve Bayes' calculation, to look for prior and likelihood values based on the value of the weight of the symptoms and diseases produced during interviews with experts. Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersted. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.

In determining input parameters for crossover probabilities and mutation probabilities from GA, experiments were conducted from a combination of crossover probability values and mutation probability with crossover probability values including 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9. At the same time, the value of the probability mutation is 0.01, 0.06, 0.11 … 0.96. After calculating the combination of each mutation probability and crossover probability, the results of the accuracy of the combination between the two are shown in Table 6.

Table 6. Combination of crossover probabilities and mutation probabilities

Mutation	*Crossover Probability*								
Probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.01	70%	65%	70%	65%	65%	65%	65%	65%	65%
0.06	95%	95%	95%	95%	95%	95%	95%	95%	95%
0.11	95%	95%	95%	95%	95%	95%	95%	95%	95%
0.16	90%	95%	90%	95%	95%	90%	95%	95%	95%
0.21	95%	95%	95%	95%	95%	85%	90%	95%	95%
0.26	90%	95%	85%	95%	95%	90%	90%	90%	90%
0.31	95%	95%	90%	90%	95%	90%	95%	95%	95%
0.36	95%	95%	95%	95%	90%	95%	95%	95%	95%
0.41	95%	95%	95%	95%	95%	95%	95%	95%	95%
0.46	95%	95%	95%	95%	95%	95%	95%	95%	95%
0.51	95%	95%	95%	95%	95%	95%	95%	95%	95%
:	:	:	:	:	:	:	:	:	
0.96	85%	85%	85%	85%	85%	85%	75%	75%	85%

From the results of the experiment, it was found that the probability of mutation resulted in the highest and stable accuracy value for each crossover probability, namely 0.06, 0.11, 0.41, 0.46, 0.51. Whereas the probability of crossover which has an average high accuracy value, is found in the crossover probability value of 0.5 and 0.9. From this experiment, it was found that the higher the value of the probability mutation, the lower the average value of the accuracy obtained, and the higher the value of the probability of crossover, the average value for the accuracy obtained increased or equaled. After the results of the crossover probability and optimal mutation probability were obtained, the experiments conducted for the calculation of the GA used a crossover probability of 0.9 and a mutation probability of 0.06.

From the results of testing 20 test data on seven training data using the system that has been created, using the Naïve Bayes method as many as 18 accurate test data with a diagnosis from experts, 2 test data is not in accordance with the diagnosis of experts. While with the Naïve Bayes method and GA as many as 19 accurate test data with expert diagnosis, 1 test data is not in accordance with the diagnosis of experts because the system can select the most influential features in the diagnosis process using GA. Accuracy comparison between naïve Bayes algorithm and combination of naïve Bayes Algorithm and GA can be seen in Table 7.

Table 7. Comparison of accuracy

Naïve Bayes Algorithm	**Combination of Naïve Bayes Algorithm with Genetic Algorithms**
90%	**95%**

4 Conclusion

Based on the research that has been done related to the optimization of the Naïve Bayes method using GA to diagnose disease in cattle. It can be concluded that in implementing a combination of the Naïve Bayes Method and GA, the initial process is to process GA to select the most influential and not influential in diagnosing disease in cattle. After the features are successfully selected, the next process is to do the classification process using the Naïve Bayes Method to diagnose cow disease from the symptoms there. Accuracy results using the Naïve Bayes method that is equal to 90%. While the accuracy of the combination of methods of GA and Naïve Bayes is 95%. So, the increase in accuracy obtained is 5%.

References

Ashari, I. A., Muslim, M. A., & Alamsyah. (2016). Comparison Performance of Genetic Algorithm and Ant Colony Optimization in Course Scheduling Optimizing. *Scientific Journal of Informatics*, 3(2), 149-158. doi:10.15294/sji.v3i2.7911

Aziz, M. (2014). Pemodelan Algoritma Genetika Pada Sistem Penjadwalan Perkuliahan Prodi Ilmu Komputer Universitas Lambugmangkurat [Genetic Algorithm Modeling in the Lecture Scheduling System of the Computer Science Study Program, Lambugmangkurat University]. *Jurnal Ilmu Komputer*, 1(1), 67-78. doi:10.20527/klik.v1i1.8

Pramesti, A. A., Arifudin, R., & Sugiharti, E. (2016). Expert System for Determination of Type Lenses Glasses using Forward Chaining Method. *Scientific Journal of Informatics*, 3(2), 177-188. doi:10.15294/sji.v3i2.7914

Pressman, Roger S. (2001). *Software Engineering: A Practitioner's Approach (6th Ed.)*. Singapore: McGraw-Hill, Inc.

Setiabudi, W. U., Sugharti, E., & Arini, F. Y. (2017). Expert System Diagnosis Dental Disease using Certainty Factor Method. *Scientific Journal of Informatics*, 4(1), 43-50. doi:10.15294/sji.v4i1.8463

Sugiharti, E., Firmansyah, S., & Devi, F. R. (2017). Predictive Evaluation of Performance of Computer Science Students of UNNES using Data Mining Based on Naïve Bayes Classifier (NBC). *Journal of Theoretical and Applied Information Technology*, 95(4), 902–911. Retrieved from http://lib.unnes.ac.id/id/eprint/33089

Trihartati, S. A., Adi, C. K. (2016). An Identification of Tuberculosis (TB) Disease in Humans using Naïve Bayesian Method. *Scientific Journal of Informatics*, 3(2), 99-108. doi:10.15294/sji.v3i2.7918

Vedayoko, L. G., Sugiharti, E., & Muslim, M. A. (2017). Expert System Diagnosis of Bowel Disease Using Case Based Reasoning with Nearest Neighbor Algorithm. *Scientific Journal of Informatics*, 4(2), 134-142. doi:10.15294/sji.v4i2.11770