Optimization of Naïve Bayes uses the genetic algorithm for classification data

by Abdurrahman Salim

Submission date: 12-Jun-2022 02:16PM (UTC+0700)

Submission ID: 1855156257

File name: f_Optimization_of_Na_ve_Bayes_uses_the_genetic_algorithm_for.pdf (625.78K)

Word count: 2666

Character count: 12985

Optimization of Naïve Bayes uses the genetic algorithm for classification data

Optimization of Naïve Bayes uses the genetic algorithm for classification data

A Salim[1,*], M R Alfian[2], H Andriani[3] and N Afifah[4]

[1]Department of Plant Production, Polytechnic of Jember State, Mastrip Street 164 Jember, Indonesia.
[2]Diploma in computer Engineering, University of Mataram Technology, Mataram City, Indonesia.
[3]Diploma in Medical Record and Health Information Management, Polytechnic of Medica Farma Husada Mataram, Mataram City, Indonesia.
[4]Diploma in Information Technology, Polytechnic of Mitra Global Jember, Jember City, Indonesia.

[*]Corresponding author: abdurrahman.salim@polije.ac.id

Abstract. Classification is one of the statistical methods to classify data systematically. However, if there is a large amount of data and various features, it often results in low accuracy. For this reason, methods are needed that can handle the data with various types. One method that can handle this problem is Naïve Bayes. Naïve Bayes is one of the methods used for classification data. This method requires a stage of selection of independent variables in increasing the accuracy of the model from Naïve Bayes. So we need an excellent method to fix these deficiencies uses a Genetic Algorithm (GA). Genetic algorithm is one of the metaheuristic methods used in optimization techniques. The data used are septic tank data in East Surabaya with eleven independent for classification data. The result of classification accuracy using Naïve Bayes is 72.7%. When Naive Bayes was used with a genetic algorithm, the classification accuracy was increased is 90.9%

1. Introduction

Classification is a statistical method for group or classifies the data systematically arranged. The classification problems arise when several rules that consist of one or more categories can not be identified immediately but must use a size. In statistics, several classification methods are used to perform data classification, such as discriminant analysis, logistic regression, Naïve Bayes, etc. [1].

Naïve Bayes is a simple probabilistic classification that calculates probabilities by summing the frequency and the combined value of a given dataset. In its application, Naïve Bayes is a method that requires a small number of training data to determine the estimated parameters required in the process classification. Naïve Bayes often works better in most real-world situations complex than expected [2]. The Naive Bayes method required stages of selecting variables to improve the model's accuracy in explaining the data. So that needed another method in variable selection one of them is Genetic Algorithm (GA). Genetic Algorithm (GA) is one of the metaheuristic methods used in optimization techniques. According to Haupt and Haupt [3], a Genetic Algorithm is an optimization technique based on genetic principles and natural selection. In the Genetic Algorithm, the population comprises many individuals who develop according to specific selection rules by maximizing fitness. According to

Sivanandam and Deepa [4], the advantages possessed by Genetic Algorithm compared to other methods are very suitable for solving global problems optimum, easy to change, or flexible to implement on various problems more expansive solution space. According to Xu and Zhang [5], in their research stated that the selection of variables using the genetic algorithm approach is better than the method of selection of backward elimination, forward selection, and stepwise method for multiple regression. According to research [6-10], naïve Bayes with genetic algorithms is a good combination of implementing applications. According to this study [11,12], the genetic algorithm is used to optimize logistic regression, and the results are better than without the genetic algorithm.

This study aims to know the optimization of naïve bayes using a genetic algorithm for classification data based on the above description. The data used to see the effect is the data on domestic waste results in the Surabaya area, where this data is classified as data [13].

2. Methods

2.1. Naïve Bayes

Naïve Bayes is a simple odds classification that calculates multiple odds by adding up the frequency and value combinations from a specific data set. Naive Bayes uses the Bayes theorem and assumes all independent attributes are assigned to class variables[14]. Naïve Bayes is a classification with probability and statistical methods [6].

The advantage of using Bayes is that this method only requires a small amount of training data to determine the classification process's estimated parameters. Naïve Bayes often works far better in most complex real-world situations than expected [2].

The equation from the Bayes theorem is:

$$P(H \mid X) = \frac{P(X \mid H).P(H)}{P(X)} \tag{1}$$

To explain the Naïve Bayes method, it is essential to know that the classification process requires several clues to determine what class is suitable for the analyzed sample. Therefore, the Naïve Bayes method above is adjusted as follows:

$$P(C \mid F1\ldots Fn) = \frac{P(C)P(F1\ldots Fn \mid C)}{P(F1\ldots Fn)} \tag{2}$$

The variable C represents the class, while the variable F1 ⋯ Fn presents the characteristic instructions needed to do the classification. Further elaboration of the Bayes formula is carried out by describing (C | F1, ⋯, Fn) using the multiplication rule. A very high independence assumption (naïve) is used, that each of the instructions (F1, F2, ⋯, Fn) is independent of each other. With these assumptions, the following similarities apply:

$$P(F_i \mid F_j) = \frac{P(F_i \cap F_j)}{P(F_j)} = \frac{P(F_i)P(F_j)}{P(F_j)} = P(F_i),\ \text{for } i \neq j, \tag{3}$$

So that,

$$P(F_i \mid C, F_j) = P(F_i \mid C) \tag{4}$$

The equation above is a model of the Naïve Bayes theorem, which will then be used in the classification process. For classification with continuous data, the Gauss Density formula is used:

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{-\frac{(x_i-\mu_{ij})^2}{2\sigma^2_{ij}}} \tag{5}$$

2.2. Genetic Algorithm (GA)

A genetic algorithm is an optimization technique based on genetic principles and natural selection. Genetic algorithms are formed from many individuals who develop according to specific selection rules by maximizing fitness [15]. This algorithm is also used to get the optimum global value by repeating or iterating the Darwinian concept of evolution.

According to Trevino and Falcian [16], there are 7 stages for carrying out genetic algorithms, namely:

1. It was forming an initial population consisting of several chromosomes which contain chromosome genes in a genetic algorithm that is used to indicate candidates for a group of genes that can be used as solutions to problems. Genes in genetic algorithms contain variables that want to be optimized. In this study, genes contain independent variables.
2. Each chromosome in the population is evaluated for its ability by using the fitness function. In this study, the fitness function is in the form of a misclassification.
3. When a chromosome has a more optimum fitness value than its initial value, then the chromosome is stopped, but if not, then the analysis stage is continued to stage 4. The smallest fitness value chosen as a solution to this study's problem because the fitness function used is the level of misclassification.
4. It was choosing chromosomes with optimum fitness values that parents use.
5. They were combining genetic information in parent's replication through crossing over. Two parents are randomly selected and used to form two new chromosomes.
6. We are doing mutations to introduce new gene elements to chromosomes randomly.
7. Stages are repeated from step 2 to the chromosomes that provide the most optimal fitness value or have reached convergence

2.3. Evaluation of Classification Method Performance

Actual data and predictive data from the classification model are presented using a cross-tabulation (Confusion matrix), which contains information about the actual data class represented on the matrix row and the predictive data class in column [17]. Classification accuracy can be seen in Table 1.

Table 1. Classification Table

Actual	Prediction	
	Positif	Negatif
Positif	*TP*	*FN*
Negatif	*FP*	*TN*

where,

$$accuration = \frac{TN + TP}{TN + TP + FN + FP} \quad (6)$$

Besides accuracy, classification performance can also be assessed based on sensitivity and specificity values. Sensitivity is the accuracy of the positive class, while specificity is the accuracy of the negative class. The formula for sensitivity and specificity is as follows.

$$Sensitivity = \frac{TP}{(TP + FN)} \times 100\% \quad (7)$$

$$Specificity = \frac{TN}{(TN + FP)} \times 100\% \quad (8)$$

Also, performance evaluation of the classification model can be done using the G-mean. G-mean is the average geometric sensitivity and specificity. If all positive classes cannot be predicted, the G-mean will be zero, so a classification algorithm is expected to achieve a high G-mean value [11]

$$G - Mean = \sqrt{Sensitivity \times Specificity} \quad (9)$$

3. Results and Discussion

3.1. Analysis Naïve Bayes

Naïve Bayes methods of use in classifying a ratio of 90%for the training data with 10 % for the data testing. Then we count the prior probability of Y in Table 2.

Tabel 2. Prior Probability of *Naïve Bayes*

y	
0	1
0,308	0,692

And then, calculate the odds of each of the marginal opportunities for each variable X against the variable Y. The following is given. The results of the opportunities for each variable are summarized in Table 3.

Table 3. Marginal probability

y	x_1	
	0	1
0	0,750	0,250
1	0,619	0,381

y	x_2			
	0	1	2	3
0	0,179	0,286	0,357	0,179
1	0,286	0,238	0,429	0,048

y	x_3			
	0	1	2	3
0	0,214	0,321	0,250	0,214
1	0,016	0,254	0,333	0,397

y	x_4	
	0	1
0	0,214	0,786
1	0,508	0,492

y	x_5	
	0	1
0	0,607	0,393
1	0,873	0,127

y	x_6			
	0	1	2	3
0	0,500	0,214	0,214	0,071
1	0,492	0,318	0,191	0,000

y	x_7	
	0	1
0	0,000	0,286
1	0.079	0,143

y	x_8	
	0	1
0	0,107	0,893
1	0,064	0,937

y	x_9			
	0	1	2	3
0	0,357	0,036	0,036	0,571
1	0,302	0,206	0,079	0,413

y	x_{10}	
	0	1
0	0,321	0,679
1	0,191	0,809

y	x_{11}	
	0	1
0	0,607	0,393
1	0,318	0,683

After calculating the prior and marginal opportunities, then the prediction results obtained from the Naïve Bayes probability model are used to compare the probability of testing data 10% of 102 observations, 11 observations, and are written in Table 4, as follows:

Table 4. Prediction y

y	$\hat{y}$
1	1
1	0
0	1
1	1
0	0
0	1
0	0
1	1
1	1
1	1
1	1

from the results of the predictions in Table 4, then the actual data is compared, and the Naïve Bayes contingency table is formed in Table 5, as follows:

Table 5. Table of contingency*Naïve Bayes*

Real	**Prediction**	
	0	**1**
0	2	1
1	2	6

From Table 5, the accuracy of the Naïve Bayes classification will be calculated summarized in Table 6, as follows:

Table 6. Classification Accuracy of*Naïve Bayes*

Akurasi	APER	Specificity	Sensitivity	G-Mean
72.7%	27.3%	0.9	0.5	0.7

3.2. *Selection analysis of GA - Naïve Bayes*

The genetic algorithm stage will stop if the minimum fitness value converges from the previous and next generation. The results of the genetic algorithm with the Naïve Bayes model can be seen in Table 7 as follows:

Table 7. There are 100 Chromosome Results and Fitness Values

Chromosome	x_1	x_2	x_3	⋯	x_9	x_{10}	x_{11}	Fitness Values
1	0	1	0		1	1	1	0,134
2	0	0	1		1	1	1	0,134
.				.				.
.				.				.
.				.				.
99	0	0	1	⋯	1	1	1	0,134

Chromosome								Fitness Values
	x_1	x_2	x_3	⋯	x_9	x_{10}	x_{11}	
100	0	0	1	⋯	1	1	1	0,134

From Table 7, the results of the Genetic Algorithm selection above obtained 5 selected variables, namely x_6, x_7, $x_{9,}$ and x_{11} contained on the first chromosome with a fitness value of 0.134. The following Table 8 will summarize the results of the classification accuracy.

Table 8.The Classification Accuracy of Naive Bayes with Genetic Algorithm

Accuracy	APER	Specificity	Sensitivity	G-Mean
90,9%	9,1%	1	0,8	0,9

3.3. Comparison of Classification accuracy

From the results obtained, the comparison of the methods used between Naïve Bayes with Genetic Algorithms - Naïve Bayes in Table 9, as follows:

Table 9. Comparison of Classification accuracy

Method	Selection Variable	The measure of classification accuracy				
		1	2	3	4	5
Naïve Bayes	x_1, x_2, …, x_{11}	72,7 %	27,3 %	0,9	0,5	0,7
GA - *Naïve Bayes*	x_6, x_7, x_9, x_{11}	90,9 %	9,1 %	1	0,8	0,9

Information :
1: Accuracy
2: APER
3: *Specificity*
4: *Sensitivity*
5: *G-Mean*

Based on Table 9, it can be seen that naïve Bayes with a genetic algorithm has a better classification accuracy than the results without a genetic algorithm. Then the variables selected smaller than the total generated by the naïve Bayes only.

4. Conclusion

The results of data analysis result from classification accuracy Genetic Algorithm - Naïve Bayes has an accuracy of 90,91% dan APER is 9,09%. But the result of variable selection Naïve Bayes is less than is 4 variable.

References

[1] Salim A 2017 *Pengoptimalan Naive Bayes Dan Regresi Logistik Menggunakan Algoritma Genetika Untuk Data Klasifikasi*. Unpublished.

[2] Patterkari S A and Parveen A 2012 *Internastional J. Adv. Comput. Math. Sci*. **3** 290 294.

[3] Haupt R L and Haupt S E 2003 *Practical Genetic Algorithms* (Hoboken, NJ, USA: John Wiley & Sons, Inc).

[4] Sivanandam S N and Deepa S N 2008 *Genetic Algorithms in Introduction to Genetic Algorithms* (Berlin. Heidelberg: Springer Berlin Heidelberg. 15 37)

[5] Xu L and Zhang W J 2001 *Anal. Chim. Acta*. **446** 475.

[6] Bustami 2013 *J. Penelitian Teknik Inform.* **3** 127 146
[7] Buani D C P 2016 *EVOLUSI J. Sains dan Manaj* **4** 1
[8] Choubey D K, Paul S and Kumar S 2017 *Classification of Pima indian diabetes dataset using naive bayes with genetic algorithm as an attribute selection*.
[9] Govindarajan M 2013 *Sentiment Analysis of Movie Reviews using Hybrid Method of Naive Bayes and Genetic Algorithm*.
[10] Busono S 2020 *J. Ilm. Teknol. Inf. Asia.* **14** 31
[11] Wardhani R P S, Sudarno S and Maruddani D A I 2019 *J. Gaussian* **8** 506 51
[12] Salim A and Alfian M R 2019 *J. Teknol. Inf. dan Terap.* **6** 2580 2291
[13] Kusumawati Y 2013 *Pemodelan Faktor-Faktor Yang Mempengaruhi Rumah Tangga Membuang Limbah Domestik Menggunakan Regresi Logistik Dan Algoritma Genetika.* Unpublished.
[14] Patil T R and Shereker M S 2013 *Int. J. Comput. Sci. Appl.* **6** 256 261
[15] Han J, Kamber M and Pei J 2006 *Data mining: concepts and techniques, Second ed.* (San Francisco)
[16] Trevino V and Falciani F 2006 *Bioinformatics* **22** 1154 1156
[17] Miroslav K and Matwin S 1965 *Br. J. Psychiatry.* **111** 1009 10

Optimization of Naïve Bayes uses the genetic algorithm for classification data

ORIGINALITY REPORT

20%	24%	20%	10%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	S A Komarudin, D Anggraeni, A Riski, A F Hadi. "Classification of genetic expression in prostate cancer using support vector machine method", Journal of Physics: Conference Series, 2020 Publication	5%
2	link.springer.com Internet Source	3%
3	trepo.tuni.fi Internet Source	3%
4	ejurnal.dipanegara.ac.id Internet Source	3%
5	D H Sulaksono, A C P Siregar. "Power Loss Classification on Shifts Based on SMS (Singlemode-Multimode-Singlemode) Structured Fiber Optic Using Gaussian Naïve Bayes Method", IOP Conference Series: Materials Science and Engineering, 2019 Publication	2%

6	gfzpublic.gfz-potsdam.de Internet Source	2%
7	iptek.its.ac.id Internet Source	2%

Exclude quotes	On	Exclude matches	< 2%
Exclude bibliography	On		