IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 6, Ser. III (Nov. –Dec. 2021), PP 33-37
www.iosrjournals.org

Naïve Bayes Gaussian Attribute Weighting with Gain Ratio Optimization

Agung Kurniawan Faisol[1], M. Arief Soeleman[2]
[1]*(Magister Teknik Informatika, Universitas Dian Nuswantoro, Indonesia)*
[2]*(PascaSarjana Teknik Informatika, Universitas Dian Nuswantoro, Indoneisa)*

Abstract: *Naïve Bayes is algorithm of a classification model, which works by predicting opportunities that will occur based on previously obtained data. the final result or class that is determined independently by the combination of the various attribute values in the data being classified. Naive Bayes is included in the 5 important roles of the algorithm. Basically, naive Bayes is very easy to use in everyday implementation, but it is very difficult to ensure accurate predictions, due to the existence of attribute freedom. With this, to eliminate this freedom, this study proposes giving attribute weights in each naive Bayes experiment by adding a gain ratio to each attribute. in this study using a public dataset from the UCI Repository. This research can help to reduce the freedom of the naive Bayes attribute*

Key Word*: naïve bayes; Information Gain.*

Date of Submission: 10-12-2021 Date of Acceptance: 24-12-2021

I. Introduction

Naïve bayes [1][2][3] including the top 10 trending data mining that is often used for general predictions. To build from a naive Bayes algorithm, a parameter is needed, attributes/features/dimensions are used for secondary data and must be present in the use of naive Bayes, a class or label is a primary primary data used to help calculate secondary data. The basic concepts used by naive bayes [4] is a theorem in statistics to calculate the probability of one class from each of the existing attribute groups, and determine which class is the most optimal. Naive bayes [5] is a network method that is often used in classification problems. One of the most effective classification methods in terms of predictive performance [6] .Naive Bayes is very effective dalam permasalahan yang sangat complex and very effective to teach in real world applications proves that the condition of the independent naive Bayes assumption is rarely true [5][7][8][1]. Naïve bayes [5] also called naive because each input attribute is assumed to have independent properties from one another, so this is a strong assumption that is real and unrealistic.

Independent assumptions on attributes [8] resulting in the effect of classification performance so that to deal with these problems many method approaches have been proposed. The researchers tried to solve the independent assumption of this naive Bayes attribute by using various enhancement techniques. These techniques can be grouped into six main categories. The first category of extension of the structure [4], Second Selection attribute [9], Third Attribute Weighting [7], Fourth Instace Selection [9], Fifth Selection Attribute[10] dan Sixth Atribute Selection [11] optimization of attribute settings. Researchers have proposed many improvements to the naive Bayes method [8] by taking an approach in the form of assigning a frequency value to naive Bayes attribute testing. Attribute weighting method [5] needs to be redeveloped by using attribute weighting based on the naive Bayes ratio, with the presence of a ratio each attribute can be given a more appropriate weight, so that the independent assumptions of each attribute can be minimized. Attribute weighting [5] [12] which performs an attribute search to optimize the weight of the attribute by maximizing the classification accuracy of the final model. By taking an instantaneous approach to naive Bayes classification [11], then the use of the naive Bayes model has a higher accuracy value than the previous naive Bayes model. Approach to attribute weighting [5] by knowing the attribute mean square error using gradient search to maximize terms from naive Bayes logs [7] on attributes. Attribute weighting emphasizes more on highly predictive attributes than on less predictive attributes so that attribute weighting correlations [8] determines that the weight for an attribute is proportional to the difference between the class attribute correlation and the attribute intercorrelation.

II. Related Work

To reduce weakness in naive bayes, some researchers propose improvements to naive bayes

FEATURE SELECTION

Naive Bayes optimization using attribute selection [13] by selecting attributes that are not used so that the attributes used are not excessive and the use of naive Bayes is more relevant, for example by (1)

$$c(x) = \arg\max P\ (c) \prod_{j=1}^{s} p\ (aj\ |c)$$

From the description of the algorithm above, c is a probability condition, argmax is the maximum argument from the attribute probabilities that have been provided to find the attribute selection by previous researchers using a subset of attributes. [4] by identifying attributes that are not used during naive Bayes data processing, then by evaluating the selection of these

STRUCTURE EXTENSION

The limitations of the assumption of conditional independence, the assumption in NB, extending the NB structure is a direct method to expect better performance from the final model. In this method can be added to explicitly represent the dependencies between attributes. In contrast to NB, this method uses Eq. To classify for simplicity, we assume that each attribute node can have at most one other attribute parent node.

In addition to the addition of attribute selection also find algorithms [14] by adding the feature extension used, optimization of the naive bayes model with the name hidden naive bayes (HNB), to reduce the assumption of independent naive bayes [5] by proposing Toward naive Bayes.

III. Methods

The method proposed in this paper is divided into two, namely the data collection method and the settlement method masalah. This study uses a data set or public dataset. The *UCI repository* is used as a reference for public data sources. The design method that will be proposed is by using the naive Bayes algorithm to reduce the independent assumptions on the naive Bayes algorithm by adding attribute weighting. It starts by dividing the dataset that has been provided from the *UCI Repository* and is divided into two parts, namely the training dataset and the testing dataset using split data. Then test the data from several tests using the 10-fold-cross-validation method, then apply the gain ratio method to each attribute that has been processed in data processing, both testing data and training data. To find out the data can be seen in table 1

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

Image 1 : Formula Of Entthropy

Dataset	*attribute*	*Type*	*instance*	*Jclass*
Breast-Caner	10	Nominal	286	2
Credit-Approval	15	Categorical	690	1
Hepatitis	19	Categorical	155	1
Iris	4	Real	150	1

Table no 1: Dataset use to implementation.

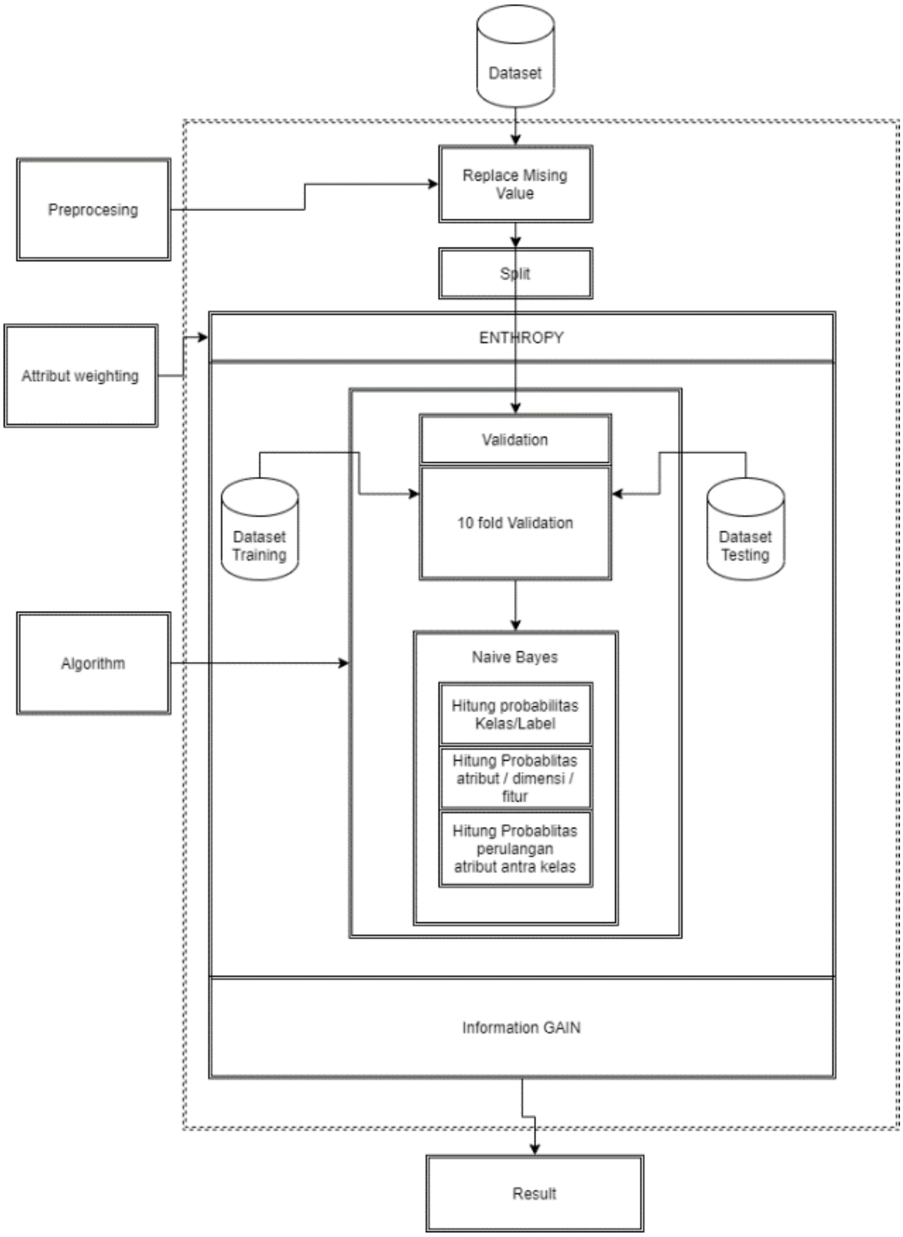

Image 2 : Proposed Algorithm

By knowing the dataset above, the data is ready to be used for the proposed research To find out which chat will be used, it is as follows Image 2 Research Suggest.

In this study, a computer was used to perform the calculation process against the proposed method with the computer specifications shown in Table 2

$$Gain\ (\ A\ ) = Entropy\ (S) - \ \sum_{i=1}^{k}\frac{Si}{s}$$

Image 3 : Formula of Gain Ratio

Table 2 : Spesification Computer

Processor	Intel(R) Core (TM) i5
Memory	12 GB
Harddisk	1 TB
SSD	256 GB
Operating System	Windows 10
Application	RapidMinner 9.9.0

Measurement of model performance using a confusion matrix table. In this study, a computer was used to perform the calculation process against the proposed method with the computer specifications shown in table 2.

IV. Experiment and Result

The experiment in this study used a dell series latitude 3490 laptop using an Intel(R) Core (TM) i5-8520 CPU @ 1.60 GHz (8 CPUs) ~ 1.8GHz. And use DDR4 memory with 12 GB. And use a 1000 GB sata hard drive, and use additional data for operating system using Windows 10 enterprise ori and a 256 GB SSD. For applications using Rapid minner with version 9.9.0.

The steps for this test are to enter the existing training dataset into the rapid minner, to find missing values or data that is empty, the researcher uses the features in the rapid minner menu Replace Mising Values, to eliminate missing data. After the errors have been found and corrected, the next step is to convert the data into two parts, the first part is for training data and the second is testing data by dividing 10 percent for testing and 90 percent for training data. The data being tested is data from the UCI Repository University California Irvinne machine learning dataset which can be obtained from the https://archive.ics.uci.edu/ site. The dataset is widely used by researchers to test the research that will be made. The dataset used consists of

1. Dataset Breast Cancer
2. Credit Approval
3. Dataset Hepatitis
4. Dataset Iris

The Breast Cancer dataset is a collection of data related to the classification of breast cancer, the attributes owned are of nominal type, consisting of 286 instances, 10 attributes, and 2 classes. Credit Approval dataset is a collection of data related to credit approval classification, the attributes owned are of real type, consisting of 690 instances, 15 attributes, and 1 class. Hepatitis Dataset is a collection of data related to the classification of hepatitis disease, the attributes owned are of real type, consisting of 155 instances, 19 attributes, and 1 class. In the experiment that will be carried out this research is to enter training data into the application used, the application used is rapid miner version 9.9.0. The dataset is divided into two using split data, dividing the data into two parts. After the training data is entered, then enter the data into 10 fold cross validation, then enter the existing naive Bayes algorithm. after entering into naive bayes, the additional weighting of attributes is entered into the algorithm by adding the entropy and gain that exist in the naive bayes.

Table 3 : Perfomance Naïve Bayes

Dataset	*Accuracy*	*Kappa*
Breast-Cancer	90.33%	0.93
Credit-Approval	88.25%	0.8
Hepatitis	93.66%	0.89
Iris	90.98%	0.92

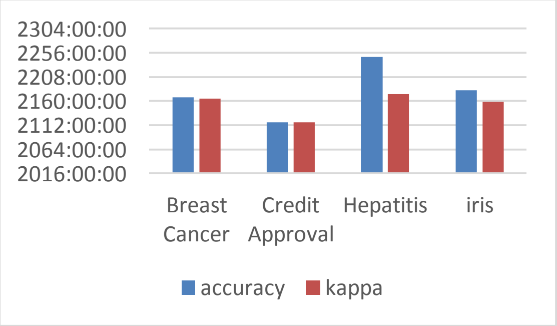

Image 4 : Chart Algorithm Naïve Bayes

Next at this stage is to combine nave Bayes by adding weighting to each attribute in the training data. And it will be proven by testing the average accuracy value

Table 4 : Perfomance Naïve Bayes + GAIN RATIO

Dataset	*Accuracy*	*NB + GAIN*
Breast-Cancer	90.33%	95.29 %
Credit-Approval	88.25%	93.67 %
Hepatitis	93.66%	96.21 %
Iris	90.98%	94.25%

Seen from table 3, the performance of nave Bayes has different accuracy in each dataset. The Brease cancer dataset has an accuracy of 90.33%. Credit approval dataset has an accuracy value of 88.25%, Hepatiti has an accuracy value of 93.66%, and Iris has a value of 90.98%. And it can be seen from the graph that Figure 3 shows the performance of the nave Bayes value.

Seen from table 4, the comparison of nave Bayes with the addition of nave Bayes added attribute weighting by using a gain ratio, so it gets a very significant comparison value. The breast cancer dataset has an accuracy value of 95.29%, the Credit Apporval dataset has a value of 93.67%, the Hepatitis dataset has a value of 96.21%, and the Irisi dataset has a value of 90.98%.

V. Conclusion

In naive Bayes research, the classification method contains a lot of data noise, to improve this, data cleaning is needed in data preparation. This test includes the stages of searching for training data and stages for using the attribute weighting algorithm. Based on the results of experiments and evaluations in this study in general, it can be concluded that the application of the Naïve Bayes algorithm given the weighting of the gain ratio can optimize the performance of the independent Nave Bayes assumption. From these results, there is an increase in each dataset Breast-cancer 90.33% (Increase 4.96%), Credit Approval 88.25% (Increase 5.42%), Hepatitis 93.66% (Increase 2.55 %), Iris 90.98% (An increase of 3.27%. Thus it can be concluded that the accuracy value of nave Bayes is given the weight gain ratio slightly increased seen from the optimization

References

[1] X. Wu *et al.*, *Top 10 algorithms in data mining*, vol. 14, no. 1. 2008.

[2] L. Jiang, L. Zhang, C. Li, and J. Wu, “A Correlation-Based Feature Weighting Filter for Naive Bayes,” *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 2, pp. 201–213, 2019, doi: 10.1109/TKDE.2018.2836440.

[3] Y. Wu, S. Huang, H. Ji, C. Zheng, and C. Bai, “A novel Bayes defect predictor based on information diffusion function,” *Knowledge-Based Syst.*, vol. 144, pp. 1–8, 2018, doi: 10.1016/j.knosys.2017.12.015.

[4] C. S. Division, M. Park, P. Langley, and P. Smyth, “Bayesian Network Classifiers *,” vol. 163, pp. 131–163, 1997.

[5] L. Yu, L. Jiang, D. Wang, and L. Zhang, “Toward naive Bayes with attribute value weighting,” *Neural Comput. Appl.*, vol. 31, no. 10, pp. 5699–5713, 2019, doi: 10.1007/s00521-018-3393-5.

[6] W. Xu, L. Jiang, and L. Yu, “An attribute value frequency-based instance weighting filter for naive Bayes,” *J. Exp. Theor. Artif. Intell.*, vol. 31, no. 2, pp. 225–236, 2019, doi: 10.1080/0952813X.2018.1544284.

[7] N. A. Zaidi, J. Cerquides, M. J. Carman, and G. I. Webb, “Alleviating Naive Bayes attribute independence assumption by attribute weighting,” *J. Mach. Learn. Res.*, vol. 14, pp. 1947–1988, 2013, doi: 10.13039/501100000923.

[8] H. Zhang, L. Jiang, and L. Yu, “Attribute and instance weighted naive Bayes,” *Pattern Recognit.*, vol. 111, 2021, doi: 10.1016/j.patcog.2020.107674.

[9] M. E. C. Santos, A. Chen, T. Taketomi, G. Yamamoto, J. Miyazaki, and H. Kato, “Augmented reality learning experiences: Survey of prototype design and evaluation,” *IEEE Trans. Learn. Technol.*, vol. 7, no. 1, pp. 38–56, 2014, doi: 10.1109/TLT.2013.37.

[10] C. Elkan, “Boosting and naive Bayesian learning,” *Proc. Int. Conf. Knowl. Discov. Data Min.*, pp. 1–11, 1997, [Online]. Available: http://www-cse.ucsd.edu/users/elkan/papers/bnb.ps.

[11] K. El Hindi, “Fine tuning the Naïve Bayesian learning algorithm,” *AI Commun.*, 2014, doi: 10.3233/AIC-130588.

[12] J. Wu and Z. Cai, “Attribute weighting via differential evolution algorithm for attribute Weighted Naive Bayes (WNB),” *J. Comput. Inf. Syst.*, vol. 7, no. 5, pp. 1672–1679, 2011.

[13] L. Jiang, C. Li, S. Wang, and L. Zhang, “Deep feature weighting for naive Bayes and its application to text classification,” *Eng. Appl. Artif. Intell.*, vol. 52, pp. 26–39, 2016, doi: 10.1016/j.engappai.2016.02.002.

[14] H. Zhang, L. Jiang, and J. Su, “Hidden naive Bayes,” *Proc. Natl. Conf. Artif. Intell.*, vol. 2, pp. 919–924, 2005.

Agung Kurniawan Faisol, et. al. “Naïve Bayes Gaussian Attribute Weighting with Gain Ratio Optimization.” *IOSR Journal of Computer Engineering (IOSR-JCE)*, 23(6), 2021, pp. 33-37.