Analysis of Attribute Reduction Effectiveness on The Naive Bayes Classifier Method

D Syafira1*, S Suwilo1* and P Sihombing1*

1 Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia

*Email: defysyafira561@gmail.com, saib@usu.ac.id, poltak@usu.ac.id

Abstract. Naïve Bayes is a prediction method that contains a simple probabilistic based on the application of the Bayes theorem (Bayes rule) with the assumption that the dependence is substantial. Classification is a technique in data mining to form a model from a predetermined data set. Data mining techniques are the choices that can overcome in solving this problem. Therefore, researchers make a comparison of Naïve Bayes by modifying using PCA. Improving the comparative performance of the Naive Bayes method by weighting using Principal Component Analysis (PCA) and Naive Bayes was carried out in this study. Research conducted on the classification model of Naïve Bayes (PCA + Naïve Bayes) using the Iris dataset that simplified into two attributes, four classes and 147 instances with an accuracy rate of 97.78% with a classification error rate of 2.22%. Meanwhile, the Conventional Naïve Bayes classification model uses four attributes with four classes from the Iris dataset, which has an accuracy rate of 95.45% with a classification error rate of 4.55%. Based on the test results of the classification model, it can be concluded that the success of the PCA can be used as a reference to improve the accuracy performance of the Naïve Bayes classification model.

1. Introduction

Classification algorithms mostly have problems with data by having many attributes, for example, in reducing the accuracy of classification [5]. One method that allows classification algorithms to work more quickly and effectively and increases the accuracy and performance of a classification algorithm is by reducing its attributes [3]. Attribute reduction can eliminate irrelevant features, reduce noise, and reduce the curse of dimensionality. Attribute reduction can also reduce the amount of time and memory needed by the classification algorithm.

Naïve Bayes is a classification method based on Bayesian probability and theorem with the assumption of attribute independence [6]. The attribute independence assumption will eliminate the need for a large amount of training data from all the attributes needed to classify data. The training data for the Bayes theorem requires at least the Cartesian multiplication of all possible attribute groups so that the fewer attributes used will reduce the training data needed. However, in reality, the assumption of an independent attribute on the Naïve Bayes Classifier is often ignored. It causes the assumption of attribute independence in the real world seldom happens.

Taheri used the Discretization Method to optimize the performance of the Naïve Bayes algorithm [4]. The research focuses on classifying binary type datasets. The results obtained using the model obtain a better level of accuracy than conventional Naive Bayes.
Handayani discussed the implementation of the Naïve Bayes Classifier algorithm for the classification of automatic text complaints and public reporting through call center services [2]. Based on the results of the study, obtained an accuracy rate for recall of 93%, precision of 93%, f-measure of 92%, and an average accuracy of 92.67%.

Babajide used Naïve Bayes Classifier to predict the level of hypertension [1]. The accuracy level obtained in the study was 83.67% for Naïve Bayes while using Decision Trees obtained an accuracy rate of 77.55%.

Yofi, in his research, used the Naïve Bayes Classifier and K-Nearest Neighbor methods to classify the Healthy Indonesia Card (KIS) [5]. The results obtained in this study are an accuracy rate of 96% by combining the Naïve Bayes method on the K-Nearest Neighbor algorithm.

In this study, information gain is used to reduce the attributes that are not relevant so that it is expected to improve the effectiveness of the performance of the Naïve Bayes Classifier method.

2. Research methods

Bayes is a simple probabilistic based prediction technique that is based on the application of the Bayes theorem (or Bayes rule) with strong (naïve) independent assumptions. In other words, Naïve Bayes, the model used, is an "independent feature model." Naïve Bayes is a classification with the probability and statistical methods presented by British scientist Thomas Bayes, namely predicting opportunities in the future based on previous experience so that it is known as the Bayes Theorem.

In Bayes (especially Naïve Bayes), the purpose of healthy independence in features is that a feature in data is not related to the presence or absence of other features in the same data. Bayes predictions are based on the Bayes theorem with the general formula as equation 1:

\[ P(H|E) = \frac{D(E|H) \times P(H)}{P(E)} \]  

The explanation of the formula is as follows: parameter description

- \( P(H|E) \): The final probability of conditional probability a hypothesis \( H \) occurs if the evidence is given \( E \) occurs.
- \( P(E|H) \): The probability of an \( E \) proof occurring will affect the \( H \) hypothesis.
- \( P(H) \): The initial probability (prior) of hypothesis \( H \) occurs regardless of any evidence.
- \( P(E) \): The initial probability (prior) of proof \( E \) occurs regardless of the other hypothesis/evidence.

The basic idea of Bayes’ rule is that the results of a hypothesis or event (\( H \)) can be estimated based on some evidence (\( E \)) observed. There are several essential things from the Bayes rules, namely:

- An initial probability / prior \( H \) or \( P(H) \) is the probability of a hypothesis before the evidence is observed.
- A final probability \( H \) or \( P(H|E) \) is the probability of a hypothesis after the evidence is observed.

This research uses Naïve Bayes because, in the classification process in probabilistic calculations, naïve Bayes has more advantages. One of the advantages is the classification of statistics that can be used to predict the probability of membership in a class. Naïve Bayes is based on the Bayes theorem, which has classification capabilities similar to decision trees and neural networks. Naïve Bayes proved to have high accuracy and speed when applied to databases with extensive data. Also, the following advantages are found in the naïve Bayes as a whole, namely:
• Quantitative handling and discrete data.
• Sturdy for the noise point isolated, for example, a point averaged when estimating the opportunity for conditional data.
• It only requires a small amount of training data to estimate parameters (mean and variance of variables) needed for classification.
• Deal with missing values regardless of the agency during the calculation of opportunity estimates.
• Fast and space efficiency.
• Sturdy against irrelevant attributes.

The link between Naive Bayes and classification, correlation of hypotheses, and evidence of classification is that the hypothesis in the Bayes theorem is a class label that becomes a mapping target in classification, whereas evidence is a feature that is included in the classification model. If X is an input vector that contains features, and Y is a class label, Naive Bayes is written with $P(X | Y)$. This notation means that the class Y label probability is obtained after the X features are observed. This notation is also called the final probability (posterior probability) for Y, while $P(Y)$ is called the initial probability Y.

During the training process $P(Y | X)$, final probability learning must be carried out on the model for each X and Y combination based on information obtained from the training data. By building the model, an X test data can be classified by looking for Y by maximizing the $P(X' | Y)$ obtained.

Classification with Naïve Bayes works based on probability theory, which views all features of the data as evidence in probability. It gives the characteristics of Naïve Bayes as follows:

• The method of working hard against private data is usually outliner data. Naïve Bayes can also handle incorrect attribute values by ignoring training data during the model building and prediction processes.
• Tangguh faces irrelevant attributes.
• Attributes that correlate can degrade Naïve Bayes classification performance because the assumption of independence of these attributes is gone.

3. Identification of problems
One of the problems with the Naïve Bayes classification algorithm is that there are many attribute data. A large number of attributes greatly influence the classification process. Therefore, we need a method to reduce attributes that are not relevant to the dataset to improve the performance of the Naïve Bayes Classifier algorithm so that it can provide scientific input for researchers in the future towards classification using the Naïve Bayes Classifier algorithm.

4. Result and discussion
The process of mining data in this study was carried out using an existing system. Therefore, to ensure that the system has implemented the algorithm correctly, it is necessary to compare the results between the systems built. The selected system used as a comparison is RapidMiner. The authors get the results as follows:

4.1. PCA analysis results
Calculate the correlation value between attributes using the covariance equation. Covariance is used to measure the magnitude of the relationship between two attributes.

Insert the covariance value calculation for each attribute pair into a covariance matrix (Covariance Matrix) measuring 8x8 where the Cov (X1, X2) is the same as Cov (X2, X1), Cov (X1, X3) is the same as Cov (X3, X1) and so on in the same way also applies to each attribute pair. The following table shows the results of covariance values:
Table 1. Covariance values.

| Attribute  | Sepal length | Sepal width | Petal length | Petal width |
|------------|--------------|-------------|--------------|-------------|
| sepal length | 1            | -0.109320823 | 0.871304556 | 0.817058302 |
| sepal width  | -0.109320823 | 1           | -0.421057394 | -0.35637616 |
| petal length | 0.871304556  | -0.421057394 | 1            | 0.961882752 |
| petal width  | 0.817058302  | -0.35637616  | 0.961882752  | 1           |

The results of calculating the correlation between attribute pairs use the covariance equation. In the Iris dataset, the Sepal Length attribute has a variance-covariance correlation of 1, while the Sepal Width correlation with the Petal Length correlates (variance-covariance) of -0.011 and so on in the same way also applies to each attribute pair. The table above is the result of calculating the correlation between attribute pairs using the covariance equation. In the Iris dataset, the Petal Length attribute has a variance-covariance correlation of 1, while the correlation of the Petal Width attribute with the Petal Length has a correlation(variance-covariance) of 0.817 and so on, in the same way, applies to each attribute pair.

The Principal Component (PC) formed is a linear combination of the mth attribute that has an eigenvalue and the maximum variance proportion of the total overall variance data (total variability). The following is the result of Iris's decomposition:

Table 2. Distribution table.

| Component | Standard deviation | Propotion of variance | Cumulative variance |
|-----------|--------------------|-----------------------|---------------------|
| PC 1      | 1.706              | 0.727                 | 0.727               |
| PC 2      | 0.96               | 0.23                  | 0.958               |
| PC 3      | 0.385              | 0.037                 | 0.995               |
| PC 4      | 0.145              | 0.005                 | 1                    |

4.2. Naïve Bayes analysis results

In this study, only two preprocessing processes were carried out. The first is handling missing value. Missing values for numeric attributes are replaced by mean values of the attributes in the same column. Whereas the missing value of the nominal value attribute is replaced by the highest possible value of the attribute in the same column. Next is the cleaning process carried out by removing data duplication.

The next process is to provide a form of categories for each subset/attribute to facilitate the mining process and the classification accuracy. The following are the results of preprocessing data from the Iris dataset. The accuracy of the Naïve Bayes algorithm is 97.78%.

Table 3. Confusion Matrix Naïve Bayesian Method uses Iris.

|                  | true Iris-setosa | true Iris-versicolor | true Iris-virginica | class precision |
|------------------|------------------|----------------------|---------------------|-----------------|
| pred. Iris-setosa| 15               | 0                    | 0                   | 100.00%         |
| pred. Iris-versicolor | 0        | 15                   | 1                   | 93.75%          |
4.3. Testing the accuracy of the PCA + Naïve Bayes Classification model

The results obtained from the PCA analysis in the form of Iris dataset dominant factors were selected based on the highest eigenvector values. Comparison of Naïve Bayes Accuracy Chart (factor loading) generated from PC 1 and PC 2 that meet the proportion of variance-covariance of 95.45%.

Table 4. The detailed data results from the PCA analysis.

| no  | class               | PC1   | PC2   |
|-----|---------------------|-------|-------|
| 1   | Iris-setosa         | -2.28385 | -0.48492 |
| 2   | Iris-setosa         | -2.10822 | 0.663745 |
| 3   | Iris-setosa         | -2.38791 | 0.33136 |
| 4   | Iris-setosa         | -2.32449 | 0.585706 |
| 5   | Iris-setosa         | -2.40717 | -0.6517 |
| 6   | Iris-setosa         | -2.08778 | -1.48768 |
| 7   | Iris-setosa         | -2.46389 | -0.05702 |
| 8   | Iris-setosa         | -2.25347 | -0.22939 |
| 9   | Iris-setosa         | -2.36288 | 1.10031 |
| 10  | Iris-setosa         | -2.21012 | 0.459482 |
| …   | …                   | …     | …     |
| 147 | Iris-virginica      | 0.938287 | 0.031708 |

So we can get the accuracy of the classification of PCA+Naïve Bayes:

Table 5. Confusion Matrix PCA + Naïve Bayesian Method uses Iris.

|            | true Iris-setosa | true Iris-versicolor | true Iris-virginica | class precision |
|------------|------------------|----------------------|---------------------|-----------------|
| pred. Iris-setosa | 14               | 0                    | 0                   | 100.00%         |
| pred. Iris-versicolor | 0               | 14                   | 1                   | 93.33%          |
| pred. Iris-virginica | 0               | 1                    | 14                  | 93.33%          |
| class recall | 100.00%          | 93.33%               | 93.33%              |                 |
5. Result

After this research has been done, the authors produce several conclusions in the form of research conducted on the Naïve Bayes classification model (PCA + Naïve Bayes), using the Iris Dataset which has been simplified into two attributes, four classes and 147 instances with an accuracy level of 97.78% with a level of accuracy classification error of 2.22%. Meanwhile, the Conventional Naïve Bayes classification model using four attributes with four classes from the Iris dataset has an accuracy rate of 95.45% with a classification error rate of 4.55%. Based on the test results of the classification model, it can be concluded that the success of the PCA can be used as a reference to improve the accuracy performance of the Naïve Bayes classification model.

References

[1] Babajide, O. A., Thomas, I.A., & Iyanuoluwa, A.O. (2017). Hypertension Prediction System Using Naïve Bayes Classifier. JAMCS. Vol 24, No 2. 1-11.
[2] Handayani, F. & Pribadi, S.F. 2015. Implementation of the Naïve Bayes Classifier Algorithm in Classifying Automatic Text Complaints and Reporting Communities through Call Center Services 110. Journal of Electrical Engineering. Vol 7 No. 1. 19-24.
[3] Prasetyo, E. 2014. Dimension Reduction of Data Sets with DRC on the SVM Classification Method with Efforts to Add Third Components. SNATIVE Proceedings, 293-300.
[4] Taheri, S. & Mammadov, M. (2013). Learning The Naïve Bayes Classifier With Optimization Models. IJAMCS. Vol 23, no 4. 787-795.
[5] Yofi, F. S., Riza, A., & Much, A.M. (2018). K-Nearest Neighbor and Naïve Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor. SJI. Vol 5 No 1. 9-18. Imbalanced Data Classification. Advanced-Data Mining and applications – 7th International Conference (pp. 83-96). Beijing, Springer.
[6] Khotimah, B.K. and Suprajitno, H., 2019, March. Modeling Naïve Bayes Imputation Classification for Missing Data. IOP Conference Series: Earth and Environmental Science Vol. 243, No. 1, p. 012111.