Analysis of classification and Naïve Bayes algorithm k-nearest neighbor in data mining

Lotar Mateus Sinaga*, Sawaluddin, Saib Suwilo

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

*lotarmateus88@gmail.com

Abstract. Naïve Bayes is a prediction method that contains a simple probabilistic that is based on the application of the Bayes theorem (Bayes rule) with the assumption that the dependence is strong. K-Nearest Neighbor (K-NN) is a group of instance-based learning, K-NN is also a lazy learning technique by searching groups of k objects in training data that are closest (similar) to objects on new data or testing data. Classification is a technique in Data mining to form a model from a predetermined data set. Data mining techniques are the choices that can be overcome in solving this problem. The results of the two different classification algorithms result in the discovery of better and more efficient algorithms for future use. It is recommended to use different datasets to analyze comparisons of naïve bayes and K-NN algorithms. the writer formulates the problem so that the research becomes more directed. The formulation of the problem in this study is to find the value of accuracy in the Naïve Bayes and KNN algorithms in classifying data.

1. Introduction

According to (Little, 1970) said that a set of model-based procedures for data processing and assessment to help names take a decision. A system will be successful if a system must be simple, fast, and easy to control, adaptive, complete with important contents, and easy to communicate.

Data Mining is part of the KDD (Knowledge Discovery in Database) process which consists of various stages such as in data selection, pre-processing, transformation, data mining and evaluation results. (Maimon. 2000).

Classification is a directional learning step. Classification serves to predict the class of objects whose classes are unknown (Raviya, 2013). The classification methods commonly used are Decision Tree, K-Nearest Neighbor, Naïve Bayes, Neural Network and Support Vector Machines (Sahu, 2015).

According to (Vladimir Nikulin, 2008) classification can only be applied to stronger training set data which are explained in the "positive" class already representing the minority without losing attributes. (Dunja Mladenic, 1999) conducts research related to selecting features that contribute to classification using certain specifications and learning abilities from classifiers on text data whose distribution is uneven. It was found that when the domain and characteristics of the classification algorithm take into account the performance of the classifier increases.

Naïve Bayes is a prediction method that contains a simple probabilistic that is based on the application of the Bayes theorem (Bayes rule) with the assumption that the dependence is

strong. K-Nearest Neighbor (K-NN) is a group of instance-based learning, K-NN is also a lazy learning technique by searching groups of k objects in training data that are closest (similar) to objects on new data or testing data.

2. Research Methods

a. Naïve Bayes

Bayes is a simple probabilistic based prediction technique that is based on the application of the Bayes theorem (or Bayes rule) with strong (naive) 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 strong independence in features is that a feature in a 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 follows:

$$P(H|E) = \frac{\mathrm{D(E|H)xP(H)}}{P(E)}$$

The explanation of the formula is as follows: Parameter Description

P (H | E)	The final probability of conditional probability a hypothesis H occurs if 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 (priori) of hypothesis H occurs regardless of any evidence.
P (E)	The initial probability (priori) 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 important 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 have more advantages. One of the advantages is the classification of statistics that can be used to predict the probability of membership of a class. Naive Bayes is based on the Bayes theorem which has classification capabilities similar to decision trees and neural networks. Naive Bayes proved to have high accuracy and speed when applied to databases with large data. In addition, the following advantages are found in the naïve bayes as a whole, namely:

1. Quantitative handling and discrete data
2. Sturdy for the noise point isolated, for example a point averaged when estimating the opportunity for conditional data.
3. Only requires a small amount of training data to estimate parameters (mean and variance of variables) needed for classification.

4. Deal with missing values regardless of agency during calculation of opportunity estimates
5. Fast and space efficiency
6. 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 value (X '| Y') obtained.

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

1. The method of working hard against isolated data which is usually an outliner data. N can also handle incorrect attribute values by ignoring training data during the model building and prediction processes.
2. Stromg faces irrelevant attributes
3. Attributes that have correlation can degrade Naïve Bayes classification performance because the assumption of independence of these attributes is gone

b. K-Nearest Neighbor

K-Nearest Neighbor (KNN) is an instances-based learning group. K-Nearest Neighbor is also one of the lazy learning methods. KNN is done by searching for groups of objects in the training data closest to the object on new data or testing data. K-Nearest Neighbor Algorithm is a method that functions to classify objects based on learning data which is the closest distance to the object being tested. Nearest Neighbor is an approach to finding cases by calculating the proximity between new problems and old problems in matching weights to the number of features available. To define the distance between two points, namely the point on the training data (x) and the point on the testing data (y), the Euclidean formula is used, as shown in the following equation:

$$\boldsymbol{D(x,y) = \sqrt{\sum_{k-1}^{n}(X_k - Y_k)^2}}$$

With D is the distance between points on x training data and data testing points y which will be classified, where x = x1, x2, x3,....., xi and y = yy1, y2, y3, ... yi and I represent attribute values and n are attribute dimensions.

In the training phase, this algorithm only stores feature vectors and classifies sample training data. In this phase, the same features are calculated for testing the data that is classified as not yet known.The distance from the new vector to which all the training sample vectors are calculated and the closest number of k pieces is taken.

Steps to calculate the K-Nearest Neighbor Algorithm method:

a. Determine Parameter K (number of closest neighbors)
b. Calculates the square of the Euclid distance (query instance) of each object against the sample data provided
c. Sort these objects into groups that have the smallest euclid distance.
d. Collect Y categories (Nearest Neighbor Classification)
e. Using the main Nearest Neighbor category, you can predict the calculated query instance value

3. Identification of Problems

Based on the background that has been explained, the writer formulates the problem so that the research becomes more directed. The formulation of the problem in this study is to find the value of accuracy in the Naïve Bayes and KNN algorithms in classifying data.

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. In analyzing the comparison between these two algorithms, the authors get the results as follows:

Table 1.Analysis Results Table Naïve Bayes

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	50	0	0	100.00%
pred. Iris-versicolor	0	47	4	92.16%
pred. Iris-virginica	0	3	46	93.88%
class recall	100.00%	94.00%	92.00%	

Table 2.K-NN Analysis Results

	true Iris-setosa	true Iris-versicolor	true Iris-virginica	class precision
pred. Iris-setosa	50	0	0	100.00%
pred. Iris-versicolor	0	47	3	94.00%
pred. Iris-virginica	0	3	47	94.00%
class recall	100.00%	94.00%	94.00%	

Table 3.Comparison of Naïve Bayes and K-NN Accuracy

Comparison of Accuracy Values	
Naïve Bayes	95.33% +/- 4.27% (micro: 95.33%)
K-NN	96.00% +/- 3.27% (micro: 96.00%)

5. Conclusion

Classification is a technique in Data mining to form a model from a predetermined data set. Data mining techniques are the choices that can be overcome in solving this problem. The results of the two different classification algorithms result in the discovery of better and more efficient algorithms for future use. It is recommended to use different datasets to analyze comparisons of Naïve Bayes and K-NN algorithms.

References

[1]. Little, J.D.C. 1970. *Models and Managers: The Concept Of Decision Calculus*. Management Science vol 16.

[2]. Maimon, O. & Last, M. 2000. *Knowledge Discovery and Data Mining, The Info-Fuzzy Network (IFN) Methodology*. Dordrecht: Kluwer Academic.

[3]. Mladenic D, Grobelnik M. 1999. *Feature Selection for Unbalanced Class Distribution and Naive Bayes.* Proceedings of the 16th International Conference on Machine Learning (ICML)

[4]. Mladenic, Dunja. 1999. Text - *Learning and Related Intelligent Agents: A Survey*. IEEE.

[5]. Nikulin, & Mc Lachlan. 2008. *Classification of Imbalanced Marketing Data with Balanced Random Sets*. Department of Mathematics, University of Queensland, Australia.

[6]. S. Sahu, et al. 2015. *Twitter Sentiment Analysis, A More Enhanced Way of Classification and Scoring*. IEEE International Symposium on Nanoelectronic and Information Systems

[7]. Syeda Farha Shazmeen, Mirza Mustafa Ali Baig, M.Reena Pawar. 2013 *Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis.* Balaji Institute of Technology and Science, Warangal, A.P, India.

[8]. Turban Efraim., E Jay., Aronson., Liang Ting-Peng. 2005. *Decision Support System and Intelligent System.* Andi Offset

[9]. Vapnik, V. 1998. *Statistical Learning Theory*. Wiley New York