

Research on Weighted Naive Bias Classification Algorithm Based on Big Data

Bo He^{1, a}, Chao Xue^{1, b}

¹School of Computer Science and Engineering, Chongqing University of Technology, 400054, Chongqing, China

^a29659807@qq.com, ^b52980551@qq.com

Abstract

Naive Bayes algorithm is one of the most common classification algorithms in many classification models. Aiming at the problem that the conditional independence assumption is flawed or not, The paper uses the big data in association rules to improve the traditional naive Bayesian algorithm, and a naive Bayesian classification algorithm based on Apriori weighting is proposed. The improved algorithm eliminates the high degree of conditional attributes, and combines the confidence of the rules, and gives the weight of each attribute separately. The Pima Indians's hyperglycemia classification test shows that the improved algorithm is feasible and effective.

Keywords

Naive Bayes; Classification algorithm; Apriori; Weighted.

1. Introduction

The Bayesian algorithm is actually a classification method that similar to the way in which statistics are used to classify data by calculating probabilities. The naive Bayes algorithm is based on the Bayesian theorem, which reduces the assumptions of an independent algorithm between the attribute values to be classified. In all classification algorithms of machine learning, Naive Bayes algorithm is largely different from other classification algorithms. Common machine learning classification algorithms are KNN, Logistic regression, SVM support vector machine, decision tree, etc. For these algorithms, they tend to learn the direct relationship between the feature output y and the feature item X by learning, or by the decision function and the distribution of conditions to classify, the process is often very complicated. Relative to these classification algorithms, Naive Bayesian classification algorithm has the advantages of simplicity, ease of use and high efficiency, and in some areas of classification problems, it not weaker than the neural network, classification ability of decision tree. In short, the idea of naive Bayes algorithm is to assume that the attributes of the items to be classified are independent of each other, and calculates the posterior probability that the variable X belongs to a certain category according to the prior probability.

But Naive Bayes is convenient because it assumes that the n dimensions of the items to be categorized are independent of each other, and we know that there is a big loophole in such assumptions in real life, in many cases it is impossible to exist, and there are some connections between more or less attributes, which leads to a deviation from the results of the simple Bias classification. In order to improve the computing precision of naive Bayes, many scholars have made great efforts in this field. Zhao Wentao et al proposed a algorithm for Naive Bayesian in underflow and accuracy problems, consider the core vocabulary of categories into the question of text categorization, the basic formulas in TFIDF algorithm are optimized and improved, and the CIT-NB algorithm is proposed to improve the performance of the algorithm [1]. Guo Feifei et al. Improved the naive Bayes algorithm by studying the weighting of distance measurement,

and got the distance function, improved the accuracy of the algorithm [2]. Jiang and others get a higher performance of the simple Bias classifier by weighting the feature of the training set [3]. Wang Chaoxue et al. proposed an improved SMOTE algorithm to improve the number of synthetic samples [4]. SonaTaheri et al. based on the attribute structure of Bias classifier and combining conditional probability weighted simple Bias classifier, it is also a very effective method [5]. The algorithm improves the performance of the naive Bias classifier to some extent, but it also improves the advantage of the simple Bias algorithm. For this problem, in this paper, the association rules and big data are used to find the relationship between each attribute. The strong association rules are divided into association rules with taxonomy and no taxonomy. By eliminating the associated attributes that do not wait for the taxonomy, the degree of association between attributes can be reduced. After that, weighted core attributes are added to increase the accuracy of naive Bayes classification, which improves the classification performance of naive Bayes algorithm based on the simple and easy use of naive Bayes algorithm.

2. Correlation Theory

2.1. Bayesian Classification Algorithm

Naive Bayesian model is different from the decision tree model, It derives from the classical mathematical theory and has a stable classification efficiency [6]. The naive Bayes formula is:

$P(C/X) = \frac{P(CX)}{P(X)} = \frac{P(X|C)P(C)}{P(X)}$, where $P(X)$ is the probability of the occurrence of the X event, it is a priori probability of X, $P(C)$ is the probability that X belongs to a certain category, it is a priori probability of C, $P(C/X)$ is the probability of the occurrence of C under the precondition of X, which belongs to the posterior probability. $P(X/C)$ is the posterior probability of the occurrence of X under the precondition of C.

(1) The simple workflow of the naive Bayes algorithm is defined as follows:

(2) The known category set $C = \{Y_1, Y_2, Y_3, \dots, Y_n\}$, where Y in the C set is an already existing category item;

The existing to-be-classified item $X = \{A_1, A_2, A_3, \dots, A_n\}$, where X is a to-be-classified item and each A represents a characteristic attribute of X and assumes that the attributes are directly independent from each other;

(3) $P(Y_1 | X)$, $P(Y_2 | X)$, $P(Y_3 | X)$, ..., $P(Y_n | X)$ are calculated from known conditions;

(4) It is concluded that if $P(Y_m | X) = \max \{P(Y_1 | X), P(Y_2 | X), P(Y_3 | X), \dots, P(Y_n | X)\}$, then $X \in Y_m$.

The key step in the algorithm is how to calculate the conditional probability of each item to be classified. According to the naive Bayes formula: $P(C/X) = \frac{P(CX)}{P(X)} = \frac{P(X|C)P(C)}{P(X)}$, We can find a set of

taxonomy that is known as a training sample set. Then we calculate the conditional probability of the characteristic attributes of each classification item in the training sample set to belong to each category, is $P(A_1|Y_1), P(A_2|Y_1), P(A_3|Y_1), \dots, P(A_n|Y_1); P(A_1|Y_2), P(A_2|Y_2), P(A_3|Y_2), \dots, P(A_n|Y_2); P(A_1|Y_3), P(A_2|Y_3), P(A_3|Y_3), \dots, P(A_n|Y_3)$. Because the premise of simple Bias is to assume that each characteristic attribute is conditional, then the following inference can be

obtained according to the Bias theorem: $P(Y_i/X) = \frac{P(Y_iX)}{P(X)} = \frac{P(X|Y_i)P(Y_i)}{P(X)}$, because the denominator $P(X)$ is a tiring multiplication of the probability of all category attributes: $\prod_{j=1}^n P(A_j|Y)$, Y is the item to be classified, and for the item to be classified, the result is a constant. Therefore, we only need to maximize the numerator for accurate classification. So there is: $P(X|Y_i)P(Y_i) = P(A_1|Y_i)P(A_2|Y_i)P(A_3|Y_i)\dots P(A_n|Y_i)P(Y_i) = P(Y_i) \prod_{j=1}^n P(A_j|Y_i)$. It can be seen that the classification

process of naive Bayes is divided into two stages. The first phase is the preparation stage, at which the feature attributes are determined and the training samples are obtained. Then calculate for each category $P(Y_i)$, Then the conditional probability of each partition is calculated for each feature attribute. The second phase is the application phase, at which the $P(X|Y_i)P(Y_i)$ is calculated for each category, and finally the $P(X|Y_i)P(Y_i)$ maximum is used as the category of X [7].

2.2. Big Data

An association rule can be simply expressed as a relationship between two disjoint itemsets A and B . There are two important concepts in correlation analysis - support and confidence, Support s refers to the transaction database D in $s\%$ of the transaction support project set A and project set B ; Confidence is expressed in C , and C indicates that transactions with $c\%$ in D support the project set A while supporting the project set B [8]. Support degree refers to the frequency of a transaction appearing in the transaction database, and is the ruler of the measurement range in the association rules. Confidence refers to the accuracy of an association rule and a ruler of the strength of a rule in association rules [9]. Their expressions are: Support $(A \rightarrow B) = P(A \cup B) = \frac{D(X)}{|D|}$, (Where $D(X)$ is the number of transactions of the X in the database D , D is a set of things). Confidence $(A \rightarrow B) = P(B|A) = \frac{\text{support}(A \cup B)}{\text{support}(A)}$. The big data is based on the association analysis of frequent itemsets. It finds frequent itemsets (That is, the support degree of this set is greater than the minimum support threshold value set) through a layer by layer iterative search to find strong association rules. The algorithm idea of Apriori can be divided into two steps: The first step is to get the candidate set from the self connection, that is, to find the set of "1- item sets" from the data set D , which is recorded as $L1$, and link the "2- item set" in order until you can't find the "K- item set". The second step is the pruning step, and it is a frequent itemset to make any nonempty subset of any frequent item set. On the other hand, if the non - empty subset of a candidate set is not frequent, then the candidate set is not a frequent itemset [10]. The big data reduces the amount of computation by these two steps and improves the efficiency of the algorithm.

3. Improved Weighted Naive Bayesian Classification Algorithm based on Big Data

3.1. Algorithm Ideas

One of the most important prerequisites for naive Bayesian algorithms is the independence of conditional attributes, but in real life, basically, there is nothing without any connection. In view of this problem, by reducing the degree of association between attributes, eliminating the associated attributes of higher degree of Association, and the core attributes are weighted to increase the classification accuracy of naive Bayes. The APBNC algorithm proposed in this paper is a simple Bias classification algorithm, which is improved by the way of weighting and eliminating. The algorithm is divided into two steps. First of all, the first step is to find the degree of dependence between the conditional attributes, through the big data, and all attributes are weighted by the calculated results, and each attribute is given a weight value. The algorithm is classified by the naive Bayes algorithm, which is called the APBNC algorithm in the paper. An important part of the APBNC algorithm is the calculation of the weight of each attribute, and the emphasis is to find the weight of the core attribute with the classification item. The taxonomy means that the results of the strong association rules found through the big data include the categories, the results are made up of a number of structures $\rightarrow$ categories. So without a taxonomy, it means that the resulting association results are not included in the category, the results are made up of an attribute $\rightarrow$ an attribute. First of all, the APBNC algorithm

uses the big data in the association rules, sets support and high confidence threshold, and according to all subsets of a frequent item set must be frequent itemsets and non frequent itemsets is a superset of the non frequent itemsets to find all frequent itemsets, and find strong association rules according to confidence. According to APBNC algorithm idea, the strong association rules are divided into two types, that is, the association rules with classified items and the association rules without the taxonomy. The strong association rules are obtained by combining the conditional attributes of the classified terms with the big data, and the strong association rules of each condition attribute are found. And the weight of each condition attribute is calculated to reduce its influence on the final classification results. The calculation method of attribute weight with classification term is that the number of condition attributes obtained in strong association rules is divided by the total number of strong association rules, $w_a = \frac{X(ai)}{X}$. The result w_a is the value of the attribute. The strong association rules obtained as: $x_1, x_3 \rightarrow c_3$, $x_2, x_4 \rightarrow c_1$, $x_1, x_3 \rightarrow c_2$, by calculating the weights of the associated conditional attributes, we can obtain $w_{x_1} = \frac{2}{3}$, $w_{x_2} = \frac{1}{3}$, $w_{x_3} = \frac{2}{3}$. The idea of this method is to improve the degree of independence of the categorization condition attributes to the naive Bias algorithm, so as to improve the accuracy of naive Bias classification. This kind of condition attribute is considered as the core attribute because the strong association rules are obtained, so it has a strong influence on the classification. Because each attribute in the taxonomy has a different weight, the key attributes will have a greater impact on the results of the classification. The second step is a strong association rule that is found without a taxonomy, which has a greater impact on the classification results with the core attributes. The formula of naive Bayes $P(C_i/X) = \frac{P(X|C_i)P(C_i)}{P(X)}$, we can find that when the molecular $P(X|C_i)$ value is large, the $P(C_i/X)$ results in greater, the more likely the taxonomy is to belong to X. In view of this point, the paper first analyzes the structure without the classification, in this way, we can calculate the correlation between attributes by the confidence of the association rules. For example, the confidence of X_2 to X_5 is 60%, and the confidence threshold is 50%, so we can make sure that when x_2 appears, X_5 has a great chance of appearing together. If at this time it is determined that X_5 is not the core attribute in the strong association rules with the taxonomy, we can think of X_5 as the associated attribute of x_2 and delete it in the computing sample set. After some adjoint attributes are deleted by calculation, the total number of attributes is reduced, the value of the numerator $P(X|C_i)$ is increased, and the accuracy of the classification is increased. And this approach can improve the accuracy of the algorithm by increasing confidence. For example, the following case: Training samples for health and disease users T , $\frac{1}{3}T$ is a healthy user, and $\frac{1}{3}T$ is a disease user, the attributes contained by a single user are $P\{a_1, a_2, a_3 \dots a_n\}$, these attributes include age, sex, blood pressure, blood sugar, family history, smoking or not, stay up or not. The standard range of these attributes can be found in the medical category, the data is preprocessed by this range, and the attributes are subdivided. Preconditioning results, such as high blood pressure, low blood pressure, high blood sugar and low blood sugar, etc, and delete abnormal attributes in health and disease users. For example, when the association rules are calculated, strong association rules are found to stay up overnight $\rightarrow$ high blood pressure, stay up and smoke. Because the high blood pressure is one of the categories, so staying up late is a core attribute that can't be eliminated. But staying up night $\rightarrow$ smoking is a association rule without classification, and the confidence is greater than the threshold, therefore, smoking can be regarded as the associated attribute of staying up late, eliminate it and improve the classification accuracy of the follow-up naive Bayes algorithm.

3.2. Algorithmic Description

The detailed steps of the improvement can be divided into six steps:

(1) Preprocessing the collected health data, according to the standard value of medical properties, the blood pressure, blood glucose were divided into high blood pressure value, low blood pressure value and high blood sugar, blood sugar value is low, other attributes are also pre-treated to similar standards;

(2) Set the support and confidence thresholds, scan the pre-processed data, find the association rules (such as Equations 1 and 2) and the big data to find the pre-treatment health attributes and strong association rules;

$$\text{Support } (A \rightarrow B) = P(A \cup B) = \frac{D(X)}{|D|} \quad (1)$$

$$\text{Confidence } (A \rightarrow B) = P(B|A) = \frac{\text{support}(A \cup B)}{\text{support}(A)} \quad (2)$$

The pseudo code for the principle of improvement is shown below:

```

Input: < Data set to be classified as D, Training set sample as T, A set of
       conditional probabilities for all attributes in known
       classification as K>.
Output: < All categories of D >
Begin:
1:   According the medical standards to preprocess the health data
    // To preprocess the training set of known classification
2   : if (T>0)
        Pretreatment T
        s←Confidence of rule
        Endif
    // To find frequent itemsets between attributes of training set
3:   for (i=0; F is exist; i++)
        Si = Smax
    // To find out the rule with the highest confidence of frequent items
4:   for (j=1; F is exist; j++)
        If (Sj>Si)
            Smax = Sj
        Endif
    Endfor
    // Remove the condition attributes in the rule
5:   delete Smax from F and D
    // Find the weight of the training set with posterior attributes is
    obtained
6:   wa←L    s←(1-si)
    // The attribute weight multiplied by the confidence of the weight
    wk←s*wa
    p= argmax P(Ci) ∏k=1n P(xk|ci)wk
    return p
end

```

(3) The classification results of the strong association rules of health attributes and classification items are divided into association rules with classification items and association rules without classification items;

(4) The number of core conditional attributes in strong association rules is divided by the total number of strong association rules $w_a = \frac{X(ai)}{X}$, and the weights of core attributes are obtained w_a .

(5) It is known from the naive Bayes formula : $P(C_i/X) = \frac{P(X|C_i)P(C_i)}{P(X)}$, $P(C_i/X)$ increases with the value of $P(X|C_i)$, so we take fourth steps to eliminate the attributes associated with the classification to improve $P(X|C_i)$ value, at the same time, this method is also advisable in practical application;

(6) Based on the results obtained above, we get a new naive Bayes formula with a weighted core attribute in a classified association rule : $P(C_i/X) = \arg\max P(C_i) \prod_{k=1}^n P(x_k|C_i)^{w_k}$, $C_i \in C$, $w_k = w_a$, from the formula, the bigger the w_k , $P(X|C_i)$ smaller, and $w_a, s \in [0, 1]$, so we can take $s = 1 - s_i$ for the confidence s .

4. Experiment and Result Analysis

The physical environment of the experiment was carried out on a laptop computer with a win10 system. The main parameters of the system are Intel(R) Core (TM) i5-2430M CPU @2.40GHz, Memory 8GB, Hard disk 500GB. The experimental data source from Pima Indians diabetes data set, is to predict the incidence of diabetes than Indians in 5 years. The results of the classification were two categories of diabetes and non diabetes. 768 data sets are selected as the training set, including 500 diabetic data sets and 268 non diabetic data sets, and each training set includes 8 attributes.

Experimental steps: First, the data set is preprocessed, and the missing values are encoded with 0. Normalization of data sets based on the standard values specified in medicine, as shown in Table 1.

Table 1. Standard reference value for health attributes

Attribute	Number of pregnancies	2 hour plasma glucose concentration	Diastolic pressure	Skin fold thickness of triceps	Serum insulin content for 2 hours	BMI	Family history of diabetes	
Standard value	3	7.8-11	60-90	16.5	9.9-124.9	18.5-24	0 or 1	

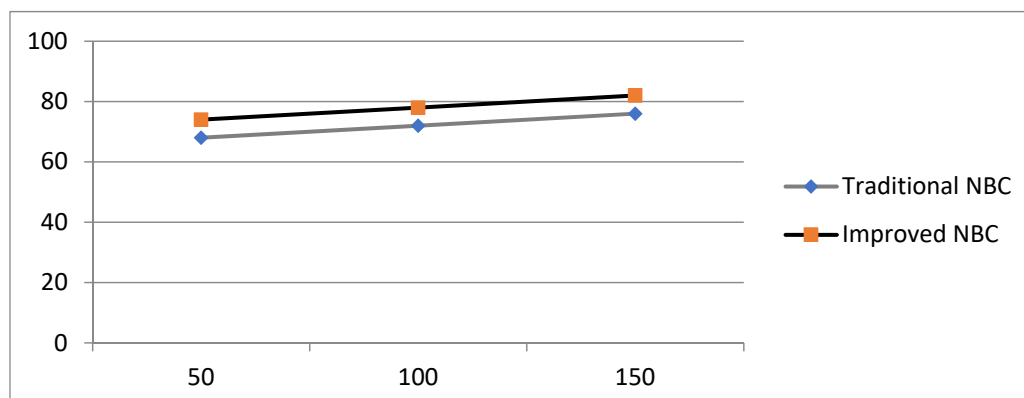


Figure 1. Traditional NBC and improved NBC classification accuracy comparison diagram

This paper selected respectively 50, 100, 150 Pima Indians health data, were classified by using Naive Bayesian algorithm and improved algorithm of traditional Naive Bayesian, the results of the experiment are shown in Figure 1. It can be seen that the accuracy of the improved algorithm was higher than the traditional naive Bayes algorithm, and the accuracy of the classification was improved by 5%-6%.

5. Conclusion

Aiming at the shortcomings of the traditional naive Bias classification algorithm's attribute independence, this paper introduced association rules, and used the big data in association rules to find the association between conditional attributes. A weighted naive Bayes algorithm for joint association rules is proposed by eliminating the condition attributes in the rules with the highest confidence level. Finally, experiments show that the classification accuracy of the improved naive Bayes algorithm is increased by 5%-6%, which proves that the improved weighted naive Bayes algorithm is effective and feasible.

Acknowledgments

This research is supported by the research fund for humanities and social sciences of the ministry of education under grant No.19XJA910001.

References

- [1] Zhao WT, Meng LJ, Zhao HH. Improvement and application of naive Bayes algorithm [J]. Measurement and control technology, 2016, 35(2): 143~147.
- [2] Guo FF, G WG. Improvement of local weighted naive Bayes algorithm based on distance measurement [J]. Computer development and Application, 2009, 22(10): 46-48.
- [3] Jiang, Qiaowei. Deep feature weighting in Native Bayes Chinese text classification[J]. Cloud Computing and intelligence System, 2016, 4: 160-164.
- [4] Wang CX, Pan ZM, Dong LL. Research on non balanced dataset classification based on improved SMOTE [J]. Computer engineering and Application, 2013, 49(2): 184-187.
- [5] Taheri S, Yearwood J, Mammadov M. Attribute weighted naïve Bayes classifier using a local optimization [J]. Neural Computing and Applications, 2014, 24(5): 995-1002.
- [6] Guo XK, Fan BB. Distributed parallel implementation of a naive Bayes text classification algorithm[J]. Computer application and software, 2016, 33(11): 240-243.
- [7] Zen SF, Zhang XY, Du XF. Research on improved naive Bayes incremental algorithm[J]. Journal of communication, 2016, 37(10): 81-91.
- [8] Tian WD, Chen H. A simplified association rule representation model[J]. Computer application research, 2015, 32(3): 658-661.
- [9] Olson D L. Association Rules[M]. Descriptive Data Mining. Springer Singapore, 2017.
- [10] Xu KY, Gong XR, Cheng MC. Audit log Association Rules Mining Based on improved big data [J]. Computer application, 2016, 36(7): 1847-1851.