A new Bayesian classification algorithm based on attribute reduction

Hongmei Nie[a], Jiaqing Zhou[b]

Math,Physics and Information Engineering College, Zhejiang Normal University, Jinhua,China

[a]nhm@zjnu.cn, [b]jhzjq@zjnu.cn

Keywords: Naive Bayesian classifier, attribute reduction,PCA, Attribute correlation coefficient

Abstract. Naive Bayesian classifier is a simple and efficient classification method. However, the assumption of the independence of its attributes is difficult to be satisfied, which influences the classification performance. In this paper, a new classification algorithm is proposed, which is based on the attribute correlation coefficient and principal component analysis. By the algorithm, we can remove the attributes that are not related to the class, and make sure that the retained attributes are independent of each other. By removing redundant attributes, the obtained attribute subset meets the assumption of Naive Bayesian classifier, and ultimately improves the classification performance of Naive Bayesian classifier.

Introduction

Classiffication is an elementary and important task in pattem recognition , machine learning and data mining. As one of the best classification methods，Bayesian classifiers have meaningful model and high classification precision. Especially Naive Bayesian classifier(NB) [1], as the foremost and simplest one, has very high classification accuracy matching or even exceeding that of other mature classifiers，such as decision tree, k-nearest neighbor, neural network. Furthermore, NB has strong ability to counteract noise data [2,3,4,7].

NB has been applied to many areas since it was proposed and its effectiveness has been verified in practice. With the increase of its application, however its disadvantage becomes more and more clear. A strong conditional independence assumption for NB is made like this: attribute variables are independent of each other [11,15]. Datasets in reality howeve usually do not satisfy this assumption，which often reduce the classifieation effect of NB obviously. One of methods to resolve this problem is to delete redundant attributes by selecting attributes and construct NB on remaining attributes. Some effective algorithms about selective Bayesian classifiers have been proposed. Langley and Sage used the Wrappers algorithm to propose a selective Bayesian classifier FSS [5]. Pazzani also gave two selective classification algorithms FSSJ and BSEJ [7] based on the Wrappers. In the two algorithms, the attributes joint and attributes reduction were used to reduce the redundant attributes, so that the classification effect was improved. Singh and Provan proposed a selective Bayesian network classifier K2-AS [6], which improved the classification effect, but it also increased the computational complexity. Later, in order to improve the efficiency of K2-AS, Singh and Provan proposed Info-AS algorithm [8]. By using the conditional information gain, conditional gain rate and conditional distance to select attributes, this algorithm made the classification efficiency of Info-AS be improved obviously. Ratanamahatana used the decision tree to carry on the attribute selection, and on this basis constructed the selective Naive Bayesian classifier SBC [10].

The majority of the algorithms mentioned above need to confirm the accuracy rate of the classification of each step, so as to increase the computational complexity. In order to reduce the computational cost, bying selecting main eigenvalues, this paper proposes a new algorithm based on Naive Bayesian classifier. It is represented as ARNB. The experiments show that the algorithm can reduce the dimension of the attributes, and improve the classification precision.

The outline of this paper is as follows: Section 2 presents the Naive Bayes classifier. Section 3 is a introduction to the prerequisite knowledge. Section 4 presents the new algorithm based on attribute reduction. Section 5 illustrates the experimental results obtained with the UCI datasets. To conclude, section 6 gives the conclusions and discusses further future work.

Naive Bayesian Classifier

Bayes Formula

$$p(C \mid X) = \frac{p(X \mid C)p(C)}{p(X)} \tag{1}$$

Formula (1) is called Bayes formula.

Naive Bayesian Classifier. The work process of Naive Bayesian classification is as follows:

1) Let $A=\{A_1,\ldots, A_n\}$ be a set of attributes, each A_r having sample values x_{rj}, $j=1,\ldots, NX_r$. Let $C=\{ C_1,\ldots, C_m\}$ be a set of classes.

2) Using $C=\{ C_1,\ldots, C_m\}$ and the formula (1), get the formula (2):

$$p(C_i \mid X) = \frac{p(X \mid C_i)p(C_i)}{p(X)},\ i = 1,\ldots,m \tag{2}$$

For each sample X, $X \in C_i \leftrightarrow p(C_i \mid X) > p(C_j \mid X)$, $1 \le i, j \le m$, $i \ne j$.

3) Since $p(X)$ for all the classes can be considered as a constant, so we only need to determine whether $p(X \mid C_i)p(C_i)$ is maximum.

4) The process of calculating $p(C_i|X)$ usually is very complex. In order to reduce the computational complexity, generally we assume that the attributes are independent of each other. So get

$$p(X \mid C_i) = \prod_{r=1}^{n} p(x_{rj} \mid C_i),\ j = 1,\ldots, NX_r \tag{3}$$

In (3), the probability $p(x_{1j} \mid C_i),\ldots, p(x_{nj} \mid C_i)$ can be estimated by the training sample.

a) If A_r, $r=1,\ldots,n$ are discrete attributes, then [16]

$$p(x_{rj} \mid C_i) = \frac{S_{rj}}{S_r} \tag{4}$$

Under the premise of the class C_i, S_r represents the number of training sample for attribute A_r, and S_{rj} represents the number of value x_{rj} for the attribute A_r.

b) If A_r, $r=1,\ldots,n$ are continuous attributes, then [16]

$$p(x_{rj} \mid C_i) = p(x_{rj}, \mu_{C_i}, \sigma_{C_i}) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_{rj}-\mu_{C_i})^2}{2\sigma_{C_i}^2}} \tag{5}$$

In (5), under the premise of the class C_i, μ_{Ci} and σ_{Ci} respectively represent the mean value and variance of x_{rj}. Or $p(x_{rj} \mid C_i)$ is divided by the sample interval.

5) In order to determine the class label for the sample X, for each class C_i ($i=1,\ldots, m$), calculate $p(C_i \mid X) \approx p(X \mid C_i)p(C_i)$ If $p(C_i \mid X) > p(C_j \mid X)$, $1 \le i, j \le m$, $i \ne j$, then $X \in C_i$.

Prerequisite Knowledge

In order to obtain a more accurate classiffication result, an attribute reduction algorithm must meet two conditions at the same time:① There is a correlation between the selected attribute and the class; ② The selected attributes are independent of each other. Through the above analysis, this paper proposes a classification algorithm based on attributes reduction. It is represented as ARNB. First of all, the prerequisite knowledge is given.

In 3.1 section, the concept of attribute correlation coefficient is introduced. In 3.2 section, principal component analysis is discribed.

Attribute Correlation Coefficient Based on χ^2 Statistics. For two attributes A, B, their values respectively are a_i ($i=1,\ldots,n$) and b_j ($j=1,\ldots,m$). Their frequencies are shown in table 1.

TABLE1. FREQUENCY TABLE

	b_1	b_2	...	b_m	$n_{i.}=\sum_{j=1}^{m} n_{ij}$
a_1	n_{11}	n_{12}	...	n_{1m}	$n_{1.}$
a_2	n_{21}	n_{22}	...	n_{2m}	$n_{2.}$
...	...	...	...	...	...
a_n	n_{n1}	n_{n2}	...	n_{nm}	$n_{n.}$
$n_{.j}=\sum_{i=1}^{n} n_{ij}$	$n_{.1}$	$n_{.2}$	...	$n_{.m}$	$n=\sum_{i=1}^{n}\sum_{j=1}^{m} n_{ij}$

χ^2 statistic [9,12,14] is defined as follows:

$$\chi^2=\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(n_{ij}-\frac{n_{i.}\cdot n_{.j}}{n}\right)^2}{n_{i.}\cdot n_{.j}/n} \tag{6}$$

In (6), n is the total number of training sample, n_{ij} is the frequency of a_i and b_j, $n_{i.}$ is the frequency of a_i , and $n_{.j}$ is the frequency of b_j.

Therefore, In accordance with the table 1, the correlation coefficient between every row and column is defined as [9,12,14]:

$$\varphi(A,B)=\begin{cases}\dfrac{n_{11}n_{22}-n_{12}n_{21}}{\sqrt{n_{1.}n_{2.}n_{.1}n_{.2}}}, & n=m=2,\\ \sqrt{\chi^2/n}, & other.\end{cases} \tag{7}$$

Principal Component Analysis. The main idea of the principal component analysis(PCA) is [13]:

Too many attributes will increase the complexity of the issue. People naturally want to have more information about the less number of attributes. In many cases, there is a certain correlation between any two attributes. When there is a certain correlation between the two attributes, then the information that the two attributes reflect has certain overlap. Principal component analysis(PCA) is to delete the redundant attributes, and establish the new attributes that are as few as possible, so that these new attributes are not relevant, and these new attributes are as far as possible to maintain the original information.

The dimension reduction steps of PCA are as follows:

1) Calculate the attribute correlation matrix R;

2) Calculate the eigenvalues λ_i, i=1,…,n of the matrix R and the eigenvectors e_i, i=1,…,n corresponding λ_i, i=1,…,n;

3) Sort the eigenvalues: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, and select the eigenvectors to construct the eigenvector matrix;

4) Select q main components.The selection of the q value mainly is based on the size of the cumulative contribution rate. That is, the general requirement of the cumulative contribution rate is more than 85%, so as to ensure that the new attributes can provide the most information of the original attributes.

Attribute Reduction Algorithm

For attributes A_i, i=1,…,n, we assume that $\varphi(A_i,A_j)$ (i, j=1,…,n) is the correlation coefficient between the attribute A_i and A_j. It is represented as φ_{ij}.

The algorithm of attribute reduction is described as follows:

1) According to the table 1 and the training data set, the frequency between attribute values of any two attributes is counted.

2) According to the formula (6), the formula (7) and the above frequencies, calculate the correlation coefficient $\varphi_{ij}=\varphi(A_i,A_j)$ $(i, j=1,\ldots,n)$ between the attribute A_i and A_j. And then get the correlation matrix $R=(\varphi_{ij})$.

3) Calculate all eigenvalues λ_i $(i=1,\ldots,n)$ in $|R-\lambda I|=0$.

4) The eigenvalues are sorted: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$. Calculate the accumulative contribution rate Q of the main components, making

$$Q=\frac{\sum_{i=1}^{q}\lambda_i}{\sum_{i=1}^{n}\lambda_i} \geq 0.85.$$

5) Eventually we can select all the desired attributes $A_{k1},\ldots, A_{kq}$ corresponding to $\lambda_1, \ldots, \lambda_q$.

6) For any one testing sample X, according to Naive Bayesian classifier, calculate

$$p(X \mid C_i)p(C_i) = p(C_i)\prod_{r=1}^{q} p(x_{rj} \mid C_i).$$

7) calculate

$p(C_i \mid X) \approx p(X \mid C_i)p(C_i)$, If $p(C_i \mid X) > p(C_j \mid X)$, $1 \leq i, j \leq m$, $i \neq j$, then $X \in C_i$.

Simulation Experiment

In order to verify the validity of the new algorithm (ARNB), the author makes a comparison test between the NB algorithm and the new algorithm(ARNB). The experimental data sets are selected from the 4 data sets(Car, Dermatology, Chess and Connect-4) in UCI database (ftp：//ftp.ice.uci.edu/pub/machine-learning-datebases). And the precision is used as the evaluation criteria. For all the sample data sets , each dataset is randomly split into 70% of instances for learning and 30% of instances for testing. For continuous datas in the dermatology data set, this paper implements the segmentation process, and then transformes the continuous datas into the discrete datas. For incomplete datas in the dermatology data set, this paper respectively sets up the specific values(the maximum frequency value), so that the data set is complete.The basic testing information of the data is shown in Table 2.

TABLE 2. EXPERIMENTAL RESULTS FOR NB AND ARNB

Dataset	Number of attributes	Number of the classes	Number of the selected attributes(ARNB)	NB%(The precision)	ARNB %(The precision)
Car	6	4	6	82.12	81.71
Dermatology	33	6	20	62.36	70.21
Chess	36	2	22	61.13	73.01
Connect-4	42	3	28	56.16	70.32
Average				65.44	73.81

Table 2 shows that the performance of the classifier is improved by ARNB. By this new algorithm, the redundant attributes can be removed, and the remaining attributes of the data subsets are ensured to be independent of each other. So the premise of Naive Bayesian classifier is well met. In addition, the analysis shows that for the data set with less attributes, the performance of the two algorithms are similar, on the contrary, for the data set with multidimensional attributes, the classification precision of ARNB is higher than that of NB. Therefore, for the data set with multidimensional attributes, the proposed algorithm in this paper can obtain more significant classification precision. In summary, table 2 shows that the precision of the ARNB is 8.37% higher than that of the NB algorithm, so the proposed new algorithm is effective.

Conclusion

The assumption of the independence of the attribute has a large effect on the classification ability of Naive Bayesian classifier. Therefore, by selecting the main eigenvalues, this paper proposes a new Bayesian classification algorithm, which can remove redundant attributes, and ensure that the selected attributes are independent of each other, so as to well meet the assumption of attribute independence. The experimental results also show that the classification precision of the proposed algorithm is higher than that of NB.

Our next work is: 1) to propose more good algorithms to deal with single label classification; 2) to study more accurate methods for dealing with incomplete data sets; 3) to propose more efficient methods for dealing with high-dimensional text classification; 4) to propose more efficient algorithms for multi labels classification.

References

[1] R.0.Duda and PE.Hart. Pattern classification and scene analysis.Newyork:Wiley.(1973)

[2] P.Clark and T.Niblett.The CN2 Induction algorithnl. Machine Learning,(1989), p. 261-283.

[3] B.Cestnik.Estimating probabilitics:A crucial task in machine learning. In: Proeeedings of the Ninth European Conference on Artificial Intelligence.(1990), p. 147-149.

[4] P.Langle, WIba and K.Thompson.An analysis of bayesian classifiers. In: Proceedings of Tenth National Conference on Artificial Intelligence,MenloPark, CA:AAAIPress,(1992), p. 223-228.

[5] P.Langley and S.Sage.Induction of selective Bayesian classifiers. in: Proeeedings of the 10th Conference on Uncertainty in Altificial Intelligence: Morgan Kaufmann, (1994), p. 399-406.

[6] M.Singh and G.M.Provan.A Comparison of induction algorithms for selective and non-selective Bayesian classifiers. in: Proceedings of the 12th International Conference on Machine Learning, Morgan Kaufmann,(1995), p. 497-505.

[7] M.J.Pazzani. Searching for dependencies in Bayesian Classifiers. Learning from Data，Artificial Intelligence and Statistics V, NewYork, Springer-Verlag, (1996), p. 239-248.

[8] M.Singh and G.M.Provan. Efficient learning of seleetive Bayesian networks classifiers. In: Proceedings of the 13th International Conference on Machine Learning, MorganKaufinan.(1996)

[9] Daling Wang,Ge Yu,Yubin Bao and Guoren Wang. A decision tree classification method based on correlation metric. Journal of Northeastern University(Natural Science), Vol. 22(2001), p. 481-484.

[10] C.Ratanamahatana and D.Gunopulos. Feature selection for the Naïve Bayesian classifier using decision trees, APPlied Artificial Intelligence, 17(5-6)(2003), p. 475-487.

[11] Xun Liang. Algorithm and application of data mining, Peking University Press, (2006), p. 164-174.

[12] Yan Wang and Silian Sui. Mathematical statistics and MATLAB data analysis. Tsinghua University Press, (2006), p. 104-110.

[13]Jing Zhang,Guang Li and Wu Chao.Automatic text classification model based on principal component analysis. Journal of Beijing University of Posts and Telecommunications, Vol. 29 (22)(2006),p. 136-138.

[14] Jun Wang. A restricted Bayesian classification model based on strong attributes. Computer Technology and Development, Vol. 17(2007), p. 205-211.

[15] Ming Zhu. Data Mining, University of science and technology of China press,(2008), p.63-156.

[16] Lili Rao, Xionghui and Liu. Dongzhan Zhang. Weighted Naive Bayesian classification algorithm based on feature correlation. Journal of Xiamen University,Vol.51(2012), p. 682-685.