Improving the Accuracy of Naïve Bayesian Classifier Using Fisher Score

Youping huang [1,2], Zhongzhi Shi[1]

(1 Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China , 100080;
2 Graduate School of the Chinese Academy of Sciences, Beijing, China, 100080)

Abstract

Classification is widely used in e-services and e-commerce. This paper proposed a new method(FS-NBC) to improve the performance of NBC. The original attributes are mapped to new attribute set according to the Fisher score, and new NBC are constructed on the new attributes. We further prove that these new attributes are condition independent on each other under certain conditions. Then we analyzed the condition independence of new attributes for two special distributions: discrete distribution without any prior information and the distribution in which the attributes are condition independent already. And illustrate that our approach will perform better than the basic NBC in theory for these two common distributions. The experiment results show that this method has excellent performance.

Key words: Naïve Bayesian Classifier, Condition Independence, Fisher Score

1. Introduction

Classification, as one of the most important branch of decision support technologies, plays a very significant role in e-services and e-commerce. The naïve Bayesian classifier (NBC) is a simple but widely applied classification technology based on Bayes' theorem. It shows excellent performance in many application domains[1][2]. The so-called "naïve" refers to its condition independence assumption, but which rarely holds in real world problems.

The naïve Bayesian classifier is based on Bayes' formula:

$$P(C = c_k \mid X = x) = P(C = c_k)\frac{P(X = x \mid C = c_k)}{P(X)}$$

(X is attributes vector and C is class label)

It implies a condition independence assumption: the non-class attributes are condition independent while given the class label.

The naïve Bayesian classifier predicts the class of a new example X=(x1, x2, … , xm) by computing the posterior probability for each class, and picks up the one with maximum probability.

$$P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)} \propto \prod_{j=1}^{m} P(x_j \mid C_i)P(C_i)$$

The class of the instance should be: $C_k = \arg\max_{C_i} P(C_i \mid X)$.

Although the NBC may have satisfactory prediction accuracy even when the condition independence assumption is strongly violated[3], many researches show that its performance can be improved by extending the NBC algorithm[4][5][6].

This paper presents a new algorithm, FS-NBC (Fisher Score-NBC), to construct new attributes from the original attributes based on information geometry theorem and Fisher score. The naïve Bayesian classifier is built on these new attributes.. Also, we investigate the condition independence of the new attributes, prove that these new attributes are condition independent of each other on certain conditions, and discuss some examples of special distribution. Then the experiment result of this algorithm is given.

2. Related Work

2.1. Scaling up the Accuracy of Naïve Bayesian Classifier

There are two main approaches to scale up the accuracy of NBC. One is to relax the condition independence assumption of NBC to support certain correlation arcs between attributes. The other is to improve the condition independence of the training data through deriving new attributes from the original attributes.

Most present researches concentrate on the first method of improving NBC. The Semi-NBC divides the attributes into groups, assuming that the attributes between different groups are condition independent[4]. Friedman investigates the tree augmented Naïve Bayesian Classifier (TAN)[2]. It modifies the NBC to a tree-structured Bayesian network. Several other methods extent TAN, such as Bayesian network

augmented naïve Bayesian classifier (BAN), Bayesian multi-net (BMN), etc[6]. All these methods improve the performance of NBC by taking into account additional conditional dependencies among non-class attributes. But searching conditional dependencies among attributes is a NP-hard problem, so these algorithms usually search only a restrict area to make a tradeoff between prediction accuracy and search complexity.

The main idea of the second way is changing the representation of the instances by creating new attributes from the original attributes, then constructing NBC on the new attributes. The simplicity of the classification model is acquired by keeping the condition independence assumption in this approach. BSEJ[5] algorithm constructs new attributes based on the Cartesian product of existing attributes. It uses the "wrapper model" for deciding which attributes to join.

2.2. Fisher Score

Assume S is a probability distribution family with parameters of a random variable X: $S = \{P(X \mid q) \mid q \in \Theta\}$, in which X is a random variable in m-dimension sample space and Θ is an open set of n-dimension Euclidean space. While the probability distribution P satisfied several regular conditions, S forms a differential manifold, called statistic manifold, and q forms the natural coordinate of S[7]. One of the most important Riemannian metric on S is Fisher information matrix. In fact, Fisher information matrix is the only appropriate Riemannian metric on S in view of keeping invariance with the transforming of sufficient statistics[8].

Let $P(X \mid q)$, $q = (q_1, q_2, ..., q_n)$ be the probability distribution of the random variable X, its log-likelihood function is denoted as:

$$l(X \mid q) = \log p(X \mid q)$$

By denoting $\partial_i = \frac{\partial}{\partial q_i}$, $\partial_i l = \frac{\partial l(X \mid q)}{\partial q_i}$, the Fisher information matrix $I = [g_{ij}(q)]$ of this probability distribution family can be defined as:

$$g_{ij}(q) = E_q[\partial_i l \partial_j l]$$

$$= \int_x \partial_i \log p(X \mid q) \partial_j \log p(X \mid q) p(X \mid q) dx$$

The Fisher information matrix defined a Riemannian metric of statistic manifold $S = \{P(X \mid q) \mid q \in \Theta\}$. Different from the Euclidean metric, this metric is invariant with the transform of coordinate. So the Fisher information matrix embeds the intrinsic features of the instances.

By denoting $U_x = \nabla_q l(X \mid q) = (\partial_1 l, \partial_2 l, ..., \partial_n l)$, the Fisher information matrix I can be written as: $I = E_q[U_x U_x^T]$. U_x is called Fisher score. As a method of feature exaction, Fisher score is widely used in many domains of machine learning[9][10][11]. Due to the definition of Fisher score, it realizes mapping an instance of m-dimension sample space to a point of n-dimension Euclidean space:

$$f(X) = (\frac{\partial \log p(X \mid q)}{\partial q_1}, \frac{\partial \log p(X \mid q)}{\partial q_2}, ..., \frac{\partial \log p(X \mid q)}{\partial q_n})$$

Tsuda argued that the Fisher score could preserve all information of the class distribution, and can separate the important dimensions from the nuisance dimensions[9][10].

3. A Naïve Bayesian Classifier Based on Fisher Score

An m-dimension supervised training instance is denoted as: (x1, x2, ..., xm , C), where xi denotes the ith attribute, C denotes the class label. Let X=(x1, x2, ..., xm). $S = \{P(X \mid q) \mid q \in \Theta\}$ is the probability distribution of X, where Θ is an open set in n-dimension Euclidean space, S forms a statistic manifold.

We denote the Fisher score that corresponding to attribute set X as Y= (y1, y2, ..., yn), where yi= $\frac{\partial \log p(x \mid q)}{\partial q_i}$. The new instance produced from Fisher score mapping is denoted as (Y, C). We are to show that for the exponential family - a probability family that is widely applied in statistics, the attributes of the instance's Fisher score are independent to each other under certain conditions.

Theorem 1: A probability distribution of exponential family- $P(X \mid q)$, the attributes of its Fisher score are independent to each other while the sufficient statistics are independent to each other.

(The proof is omitted because of the limit of the length of this paper.)

In theorem 1, what we have shown is with no given the condition of class. In fact, this is also tenable while given the condition of class. From this theorem.

In practice, many common distributions belong to exponential family, such as normal distribution, discrete distribution, and so on. The sufficient statistics in a statistic experiment often correspond to parameters of the distribution family, and have good independency in many cases.

Based on Fisher score mapping, we can adopt a common way to construct NBC. The algorithm is described as follows:

Step1: Determining the prior probability distribution model of the sample. The model can come from expert's experience, or be induced by common way in statistics, or be decided by combining both methods.

Step2: Complementing the missing values in data sets. From the computation of Fisher score, it can be seen that the lack of a value can produce the failure of the whole Fisher score. So, it is necessary to adopt some common ways to complement missing values, such as computing expectation and conditional expectation etc.

Step3: Changing the representation of the instances by the Fisher score mapping. After the prior probability distribution family of samples are decided, the new attributes can be calculated by the formula $f(X)=(\frac{\partial \log p(x|q)}{\partial q_1}, \frac{\partial \log p(x|q)}{\partial q_2},\ldots, \frac{\partial \log p(x|q)}{\partial q_n})$.

Step4: Feature selection. As the Fisher score mapping can produce many dimensions, it is necessary to adopt some feature selection algorithm to keep some important dimensions and get rid of some dimensions that are not related to classification.

Step5: Discretizing the Fisher score. As the common NBC can only tackle discretized attributes, the Fisher score, which is continuous, need be discretized. But this is not necessary, which is dependent on the algorithm adopted in the next step. If the algorithm can handle continuous attributes, this step can be omitted. There are several ways to discretize continuous values.

Step6: Constructing NBC. After finishing the above steps, we have obtained new discretized attributes, and then we can construct NBC by adopting a general way. In fact, we can also construct other extended NBC, such as TAN, BSEJ etc.(In the experiment, We have used TAN based on Fisher score, that is TA-FSNBS algorithm)

We consider that an instance of one class corresponds to a group or groups of special distribution parameters. Since Fisher score is dependent on the sample's distribution parameters, the new attributes can reflect sample's class well. The constructed NBC based on Fisher score can not only keep the information contained in sample, but also combine the information of the prior probability distribution by choosing the prior distribution family of the instances.

4. Analysis of Two common Distributions

1. Discrete Distribution Without any Prior Information

Let an instance be denoted as (X; C), where X=(x1, x2, ..., xm) denotes the attributes of the instance, and C is the class label. X can take n+1 values v1, v2, ..., vn. We represent the discrete distribution with the form of exponential family[7], and calculate it's Fisher score: (the steps of calculating Fisher score is simple and omitted)

$$f(x)=(d_1(x),d_2(x),\ldots,d_n(x))$$

$$d_i(x)=\begin{cases}1 & X=v_i\\ 0 & otherwise\end{cases}$$

Namely, to an instance X=vi, the Fisher score of it is:

f(X)=(0,0,...0,1,0,...,0)

(Only the ith attribute is 1, and the others are 0)

So the Fisher score holds all the information contained in the original sample. Therefore, it is obvious that the classification accuracy of this classifier will be higher than that of NBC. In fact, the obtained Fisher score can be viewed as extension of the BSEJ algorithm proposed by Pazzani which takes the Cartesian product of all the attributes as the new attributes.

As the BSEJ algorithm, attributes can be classified into different groups according to actual situations. The dependence possibility between groups is low, while that among a group is high. Then calculate the Fisher score of each group, and all groups are integrated into the whole Fisher score of the instance. This will greatly reduce the dimensions of Fisher score.

2. Attributes of the Distribution are Condition Independent of Each Other

Without loss of the generality, we assume that the sample has only two non-class attributes: (X; C), X=(x1, x2), where x1, x2 are mutually condition-independent. We omit the condition that given the class label in following discussion, this will not affect the processing and result of our discussion.

Suppose that the probability distribution of x1 is $P(x1|q_1)$, that of x2 is $P(x2|q_2)$, q_1 and q_2 are real values, The distribution of X can be denoted as:

$$p(x|q)=p(x|q_1)p(x|q_2)$$

Then Fisher score of $p(x|q)$ is:

$$f(x)=(\frac{\partial \log p(x_1|q_1)}{\partial q_1}, \frac{\partial \log p(x_2|q_2)}{\partial q_1})$$

It is obvious that under the conditions that x1 and x2 are mutually (condition) independent, the attributes of Fisher score still keep theirs (condition) independency. Namely, from the point of (condition) independency, the performance of NBC constructed on Fisher score will not be worse than that of basic NBC.

5. Experiment Results

The results of FS-NBC and TA-FSNBC on ten data sets of UCI database are given in this section. As

mentioned above, determining prior distribution family is an important factor in FS-NBC. In the experiments, we simply adopt the method described in section 4.1 to tackle discrete attributes. First, discrete attributes are grouped by computing the mutual information, and then, the Fisher score is calculated. We adopt normal distribution to tackle continuous attributes. To deal with the missing value, we fill in a single special value.

Besides constructing NBC based on Fisher score, we also construct the TAN based on Fisher score. The result is as follows:

Table 1. Experiment results (the value is the classification precision)

Domain	NBC	FS-NBC	TAN	TA-FSNBC
Anneal	96.24	96.32	96.24	96.34
Glass	68.65	82.75	67.78	83.96
Iris	91.35	92.00	93.62	92.05
Letter	74.50	85.60	85.83	85.77
Mushroom	94.0	93.41	96.45	93.81
Pima	75.51	75.52	75.52	75.52
Post-op	69.88	70.00	70.01	70.00
Segment	91.16	93.55	95.56	94.05
Vote	90.21	87.78	91.95	88.45
Waveform	76.04	86.33	78.10	87.52

(The TA-FSNBC Algorithm is Tree Augment NBC[2] based on Fisher Score.)

From the experiment results, it is quite obvious that the classification accuracy has been improved greatly in FS-NBC compared with NBC and TAN, while the TA-FSNBC extended by tree structure doesn't really improve the FS-NBC. The reason should be that there has already existed excellent independency between new attributes of the instances after the mapping of Fisher score. The experiments also show that the accuracy of FS-NBC decreases a bit when FS-NBC is applied to data sets with more missing values, this is because the error is increased when computing Fisher score in such cases.

6. Conclusion

Naïve Bayesian classifier (NBC) has been widely used in e-services and e-commerce because of its simplicity and good classification accuracy, such as consumer classification and commodity classification in recommender systems, text classification in Web mining, and so on. But its condition independence assumption, which is impractical usually, restricts its further application. Many researches have been done to improve the classification accuracy of NBC. Two primary ways are as followed: one is improving NBC by relaxing restrictions of condition independence assumption; the other is improving the condition independence of the training set through creating new attributes from original ones. This paper presents a new method to create new attributes of the instances based on Fisher score. The Fisher score can improve the condition independency of attributes while keep the information of the instance's distribution, so the FS-NBC improved the performance of NBC.

7. References

[1] Langley, P., Iba, W., & Thompson, K.: An Analysis of Bayesian Classifiers. Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA , AAAI Press (1992) 223-228

[2] Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning, 29 (1997) 103-163

[3] Domingos, P., Pazzani, M.: Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Proceedings of the Thirteenth International Conference on Machine Learning, Morgan Kaufmann Publishers, Inc. (1996) 105–112

[4] Kononenko, I.: Semi- Naïve Bayesian Classifier. Y. Kodratoff (Ed.), Proceedings of Sixth European Working Session on Learning, Springer-Verlag (1991) 206-219

[5] Pazzani, M.: Constructive Induction of Cartesian Product Attributes. Information, Statistics and Induction in Science, Melbourne, Australia (1996)

[6] Cheng, J., Greiner, R.: Comparing Bayesian Network Classifiers. Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence (UAI'99), (1999) 101-107

[7] Amari, S.: Differential-Geometrical Methods in Statistics, Lecture Notes in Statistics, Springer-Verlag, Berlin, Vol.28 (1985)

[8] Wei, B., Some invariance of statistic manifold, Application of probability statistics, Vol.3 (1987) 106-112 (in Chinese)

[9] Tsuda, K., Kawanabe, M., R¨atsch, G., Sonnenburg, S., Muller, K.-R.: A New Discriminative Kernel from Probabilistic Models. Neural Computation (2002)

[10] Tsuda, K., Kawanabe, M., Muller, K.-R. Clustering with the Fisher Score. In S. Becker, S. Thrun, and K. Obermayer, (eds.), Advances in Neural Information Processing Systems 15. MIT Press (2003)

[11] Muller Sonnenburg, S., R¨atsch, Jagota, A., M¨uller, K.-R.: New Methods for Splice Site Recognition. In ICANN'02 (2002)