http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (17): 4336-4339, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4336.4339

An Improved TANC Method Based on Bayesian Equivalence Theorem

Zhao Xiaoq-iang, Yang Jia-min and Zhou Jin-hu
College of Electrical Engineering an Information Engineering, Lanzhou University of Technology, Lanzhou, 730050, Gansu, China

Abstract: TANC (Tree Augmented Naive Bayes Classifier) is efficient extension of NBC (Naïve Bayes Classifier). This method not only inherits the simple and high efficiency performances of NBC, but also enhances the generalization ability. However it ignores the correlation between weaken attributes. So, an improved TANC method is proposed in this study according the dependence degree and the correlation between attributes. This method can set up a correspond dependencies to effectively improve the classification accuracy by selecting the appropriate attributes. Compared with NBC and TANC, experimental results showed this method is better than TANC and NBC in performance.

Key words: Machine learning, TANC, NBC, attribute dependency relationship

INTRODUCTION

With the rapid development of machine learning and computer technology, the application of classification is now becoming more and more widely in recent years (Fan, 2009). The technology of classification can be viewed as a mapping function which can maps instance space to one or some label of category space. Lots of methods of classification techniques is used to build the classification model, for instance, The Neural Network, The Decision Tree, Support Vector Machine, Bayesian (Chen, 2012; Deng *et al*., 2012) and so on. Among them, Bayesian method is gradually becoming one of research hot spots in machine learning area, because Bayesian method is based on solid mathematics theory and it has the ability of integrating data information.

Among numerous classification methods, NBC method of Bayesian classification is one of the most simple, effective and successful classifier in actual applications (Perez *et al*., 2009; Hsu *et al*., 2008). NBC method is based on the property to maintain relative independence assumption which is often inconsistent with actual situations and lead to bad classification performance. Aiming at the problems of NBC, researchers have proposed many improved NBC methods. One of the most representatives is proposed by Friedman called TANC method based on distribution (Friedman and Goldszmidt, 1996). It is a natural extension of NBC method, TANC method is combined attribute ability of dependent relationships with Bayesian simplicity. Essentially speaking, it can improve the performance of classification by reducing the qualification of attribute and expanding the optimal scope and at the same time it can weaken the independence assumptions in the Naïve Bayesian according to rational selecting augmenting arcs and attribute subset to expand the structure of Bayesian. A literature proposes a new model based on Bayesian theorem which it is the extension of the Bayesian model that enhances correlation degree between the attributes to improve the classification performance (Shi *et al*., 2004). A literature proposes a TAN model based on rough set which it can reduce misclassification rate by filling the defect value of data and dealing with some defective attribute to improve the classification performance (Wang and Zhang, 2004). However, these methods don't indicate relationship between the attributes completely and ignore the weak part of the relative relationship among the dependence properties. So, an improved TANC method is proposed in this study. According to the mutual dependence between attributes, this method is to find some strong attributes which can make strong effect on other attributes by searching the attribute space, so it can use these attributes to express dependencies among other attributes.

TANC METHOD

TANC is introduced by Friedman and Koller (2003). Experiments indicate that TANC is significantly better than NBC in terms of classification accuracy. By calculating condition mutual information between attributes as a power, this method structures the complete undirected graph and maximum spanning tree. Then by choosing a root node, the undirected tree is changed to a

Corresponding Author: Zhao Xiaoq-iang, College of Electrical Engineering an Information Engineering, Lanzhou University of Technology, Lanzhou, 730050, Gansu, China

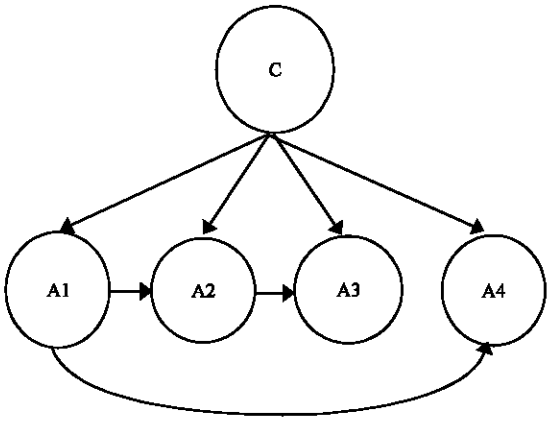

Fig. 1: TANC model

directed tree. Tree topology is formed between the attribute nodes and the largest number of the parent nodes is limited to 2, in other words, each attribute has only one as its parent in addition to the class node as its parent node. The classification model is shown in Fig. 1. Each attribute variable at most has two points to point its associated edge. This model has a good comprehensive performance which embodies an appropriate compromise between the learning efficiency and the ability that can accurately describe the correlation of attributes. TANC classification principle is as follows.

Given a set of random variables $U = \{X_1, X_2, \ldots, X_n, C\}$, C is a class variable and its value range is $\{c_1, c_2, \ldots, c_m\}$ and m is the total number of $X_1, X_2, \ldots, X_n$, which show the characters of classification, n is the attribute number of classification. $P(c_j / x_{i1}, x_{i2}, .., x_{in})$ is posteriori probability. A posteriori probability stands an object belonging to a kind of probability and a class with maximum posteriori probability is the object's class.

TANC is supposed that the structure of Bayesian network composed of attribute nodes $X_1, X_2, \ldots, X_n$ is a tree, that is to say each attribute variable has no more than one attribute parent except the class parent. So, the classification formula of TANC is:

$$c^* = \arg\max_{c_j \in c} \frac{P(x_{i1}, x_{i2}, \ldots, x_{in} / c_j) \times P(c_j)}{P(x_{i1}, x_{i2}, \ldots, x_{in})} = \arg\max_{c_j \in c} P(x_{i1}, x_{i2}, \ldots, x_{in} / c_j) \times \prod_{k=1}^{n} P(x_{ik} / \prod x_{ik}) \quad (1)$$

However, this kind of construction can not fully reflect the dependencies between attributes.

AN IMPROVED TANC METHOD BASED ON BAYESIAN EQUIVALENCE THEOREM

Improved TANC method: Let A_p and A_q be the division of property set $\{A_1, A_2, .., A_n\}$.p and q represent the value of A_p and A_q. Instance $(a_1, a_2, \ldots, a_n)$ represents the probability of the class C. This can be made of Bayesian theorem as follows:

$$P(c_j / a_1, a_2, \ldots, a_n) = \frac{P(a_1, a_2, \ldots, a_n / c_j) \times P(c_j)}{P(a_1, a_2, \ldots, a_n)} = \alpha \times P(c_j) \times P(a_1, a_2, \ldots, a_n / c_j) \quad (2)$$

α is a regularization factor. $P(c_j)$ is a prior probability of c_j. $P(c_j/a_1, a_2, \ldots, a_n)$ is a posterior probability of c_j. The prior probability is independent of the training sample data and the posterior probability reflects the sample data to the effects of class C. Therefore, it can be derived:

$$P(c_j / p,q) = \frac{P(q / c_j, p)}{P(q / p)} \times P(c_j / p) = \beta \times P(q / c_j, p) \times P(c_j / p) \quad (3)$$

$$\beta \times \gamma \times \prod_{t=1}^{k} P(a_t / c_j, a_1, a_2, \ldots, a_{t-1}) \times \prod_{s=1}^{n-k} P(a_s / c_j, a_1, a_2, \ldots, a_k) = \beta \times \gamma \times P(c_j) \times \prod_{t=1}^{k} P(a_t / c_j, K(a_t)) \times \prod_{s=1}^{n-k} P(a_i / c_j, K(a_s)) = \beta \times \gamma \times P(c_j) \times \prod_{l=1}^{n} P(a_l / c_j, K(a_l)) \quad (4)$$

β and γ are the Regularization factors. $K(a_l)$ is the value of $K(A_l)$ which is the parent node set of A_l.

So, we can see, if we can find the maximum dependencies between attributes and clearly show up, $P(c_j)$ can be more accurately. So, it can make:

$$P(c_j) \times \prod_{l=1}^{n} P(a_l / c_j, K(a_l))$$

increase, at the end, we can get a more accurate classification.

Steps of the improved TANC method: The most important step is how to define A_p and A_q. When A_p and A_q are defined, then we can make the mode through adding arc that can weaken independence assumption between attributes. There a certain interdependent relationship exists between attributes and it is in different dependence. The improved model as shown:

The steps of the process are as follows:

- **Step 1:** Calculate the conditional mutual information between each pair of attributes by using the Eq. 5:

$$I(A_i;A_j/C)=\sum_{a_i,a_j,c}P(a_i,a_j/c)\log\frac{P(a_i,a_j,c)}{P(a_i/c)P(a_j/c)} \quad (5)$$

- **Step 2:** Calculate the mutual information mean value of A_1 and other attributes:

$$EI(A_1,A_j/C)=\frac{1}{n}\sum_{j=1}^{n}I(A_1,A_j/C) \quad (6)$$

- **Step 3:** Analyze the current performance of the classifier and save the result as a Accuracy 1
- **Step 4:** If EI (A_p)>EI (A_q), then A_p is A_q's parent; Otherwise, A_q is A_q's parent. By compared with other variables, A_p and A_q are two variables of strong correlation
- **Step 5:** Analyze the performance of the classifier and save the result in Accuracy 2. If Accuracy 2 is better than Accuracy1, then go Step 6. Otherwise, return to step 2. So, the determined A_p and A_q can make the classifier achieves optimal performance
- **Step 6:** Sort the attributes according to the size of the mutual information values. Mark the number of the attribute A_1 value as n (A_1)and mark the number of the attribute A_k value as n(A_k). If n (A_1)>n (A_k), then elicit an arc from A_1 to A_k. Otherwise, elicit an arc from A_k to A_1
- **Step 7:** Choose the relevant properties of the maximum mutual information and determine the root node, then set all direction of the edges from the root variables and construct the directed graph of maximum spanning tree
- **Step 8:** Add a node C and elicit an arc from C to each attribute, then contrast TANC

RESULTS OF EXPERIMENTS

All experiments are done on Weka system (Witten and Frank, 2000), all experimental data are come from the UCI repository (Newman *et al*., 2003). Table 1 lists the number of instances of each data set, number of classes, number of attributes and whether there are missing values such as data information. Because the method can't deal with continuous numerical data, so we use the "weka.filiters.Discretizefilter" to make continuous numerical discretized. In this way, we can make all numerical attribute values convert into numerical ordinal type values and in a data set with data missing. We regard the lost value as a single value to deal with.

The main purpose of the experiment is to compare the classification accuracy of NBC, TANC and improved TANC method.

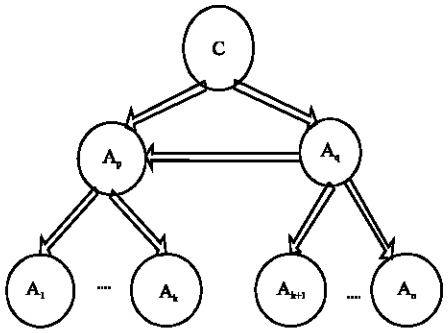

Fig. 2: An improved model of TANC

Table 1: Experimental data set

Data sets	Training sets	Classes	Attributes	Missing value
Breast	683	2	9	No
Bupa	345	2	6	No
Cleveland	303	2	13	No
Glass	214	6	10	No
Iris	150	3	4	No
Post-operative	90	3	8	Yes
Wine	178	3	13	No
Zoology	101	7	16	No

Table 2: Results of the three kinds of methods

Data sets	NBC	TANC	Improved TANC
Breast	57.22	57.24	57.24
Bupa	95.63	95.74	96.72
Cleveland	83.06	81.37	83.26
Glass	69.05	68.57	72.73
Iris	93.17	91.71	93.30
Post-operative	68.87	69.11	68.88
Wine	91.05	91.08	92.14
Zoology	94.83	95.52	96.02

The classification accuracy of each method is percentage that is examples of successful prediction for the total instances. 10 cross-validation estimation accuracy of classifier are used. Figure 2 is the results of classification accuracy of three methods. The abscissa represents the eight data sets in this experiment, the ordinate represents classification accuracy. The blue bar stands for NBC, green bar stands for TANC and red bar stands for the improved TANC method. We can see that the improved TANC method is better than the other two methods from the Fig. 3.

Table 2 is the classification accuracy results of 8 data sets of three kinds of methods. From the Table 2, we can see that the accuracy of the improved TANC is higher than that of NBC and TANC for 7 data sets, but for POST-operative data set, the accuracy is slightly lower than that of TANC, because each attribute's parent node is limited to 2, so as a result, for data sets with large numbers, classification accuracy will reduce.

Table 3 is the three kinds of classification schemes in each data set on the comparison of three methods by

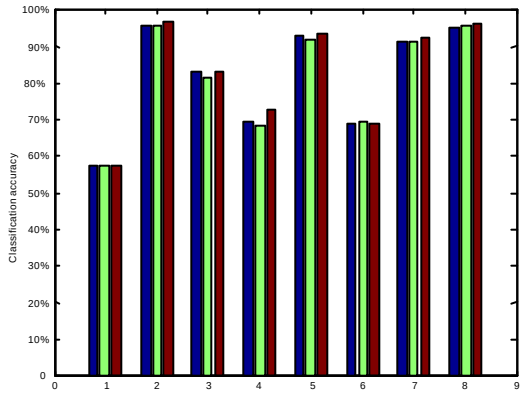

Fig. 3: Three methods of classification accuracy

Table 3: Results of three methods under double tail paired t test

	Improved algorithm	NBC	TANC
Improved algorithm	-	0	1
NBC	8	-	5
TANC	7	3	-

adopting double tail paired t test. We can see that in the case of a significant level of 0.05, the improved TANC is better than other methods.

CONCLUSION

The improved TANC method is proposed in this study according the dependence degree and the correlation between attributes. This method can set up a correspond dependencies to effectively improve the classification accuracy by selecting the appropriate attributes. Compared with NBC and TANC, experimental results showed this method is better than TANC and NBC in performance by using UCI data sets.

ACKNOWLEDGMENTS

The authors would like to thank for the support by University basic scientific research project of Gansu province (1203ZTC061). The authors also thank for the support by the Open Foundation, Technology and Research Center for Manufacturing Informatization Engineering of Gansu Province, China (No. 2012MIE01F02).

REFERENCES

Chen, Y.H., 2012. Bayesian network structure learning algorithm based on mutual information. Comput. Eng. Appl., 48: 39-43.

Deng, G., Y. Zhao, L. Liu and Y. Wang, 2012. An optimal bayes classification algorithm. Comput. Measurement Control, 20: 199-201.

Fan, K.X., 2009. Design of NB combination text classifier based on various feature selection. Comput. Eng., 35: 191-193.

Friedman, N. and D. Koller, 2003. Being bayesian about network structure: A bayesian approach to structure discovery in bayesian networks. Machine Learn., 50: 95-125.

Friedman, N. and M. Goldszmidt, 1996. Building classifier using Bayesian network. Proceedings of the 13th Nation Conference on Artificial Intelligence, August 4-8, 1996, Menlo Park, CA., pp: 1227-1284.

Hsu, C.C., Y.P. Huang and K.W. Chang, 2008. Extended Naive Bayes classifier for mixed data. Expert Syst. Appli., 35: 1080-1083.

Newman, D.J., S. Hettich and C.L. Blake, 2003. UCI repository of machine learning databases. University of California, Department of Information and Computer Science, Irvine, CA.

Perez, A., P. Larranga and I. Inza, 2009. Bayesian classifiers based on kernel density estimation: Flexible classifiers. Int. J. Approximate Reasoning, 50: 341-362.

Shi, H.B., Z.H. Wang and H.K. Huang, 2004. A restricted double-level bayesian classification model. J. Software, 15: 193-199.

Wang, Z.H. and F. Zhang, 2004. A selective tree-augmented network classifier based on rough set theory. J. Fudan Univ. (Nat. Sci.), 43: 725-728.

Witten, I.H. and E. Frank, 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, Seattle, pp: 265-314.