Pre-publication draft of a paper which appeared in the Proceedings of PAKDD'03 pp 501-512
Publisher: Springer

A New Restricted Bayesian Network Classifier

Hongbo Shi, Zhihai Wang, Geoff Webb, Houkuan Huang

[1] School of Computer and Information Technology,
Northern Jiaotong University, Beijing, 100044, China
[2] School of Computer Science and Software Engineering,
Monash University, Clayton, Victoria, 3800, Australia

Abstract. On the basis of examining the existing restricted Bayesian network classifiers, a new Bayes-theorem-based and more strictly restricted Bayesian-network-based classification model $DLBAN$ is proposed, which can be viewed as a double-level Bayesian network augmented naive Bayes classification. The experimental results show that the $DLBAN$ classifier is better than the TAN classifier in the most cases.
Keywords: Naive Bayes, Bayesian Network, Classification

1 Introduction

Many approaches and techniques have been developed to create a classification model. The naive Bayesian classifier is one of the most widely used in interactive applications due to its computational efficiency, competitive accuracy, direct theoretical base, and its ability to integrate the prior information with data sample information. However its attribute independence assumption rarely holds in real world problems. Previous research has shown that semi-naive techniques [1] and Bayesian networks [2, 3] that explicitly adjust the naive strategy to allow for violations of the independence assumption, can improve upon the prediction accuracy of the naive Bayesian classifier in many domains.

A Bayesian network classifier is a probability classification method, which can describe the probability distributions over the training data and show better classification performance in some domains. However, learning unrestricted Bayesian network is very time consuming and quickly becomes intractable as the number of attributes increases [3, 4]. Therefore, restricting the structure of Bayesian networks has become an active research area. TAN (tree augmented naive Bayes) is a tree-like Bayesian networks classifier [3, 5]. BAN (Bayesian network augmented naive Bayes) extends the structure of TAN by allowing the attributes to form an arbitrary graph, rather than just a tree [3, 6], which tends to search the whole arc space in a complete directed graph in order to select the best arc set.

2 Restricted Bayesian Network Classifiers

Bayes theorem is the theoretical basis of Bayesian network learning method, which associates the prior probabilities with posterior probabilities. Let A_1, A_2,

$\cdots, A_n$ be attribute variables, C be the class label (or variable). Bayes theorem can be expressed as follows.

$$P(C|A_1, A_2, \cdots, A_n) = \frac{P(C) \cdot P(A_1, A_2, \cdots, A_n|C)}{P(A_1, A_2, \cdots, A_n)} \tag{1}$$

$$= \alpha \cdot P(C) \cdot P(A_1, A_2, \cdots, A_n|C) \tag{2}$$

$$= \alpha \cdot P(C) \cdot \prod_{i=1}^{n} P(A_i|A_1, A_2, \cdots, A_{i-1}, C) \tag{3}$$

where α is a normalization factor. Therefore, the key issue of building a Bayesian classification model is how to estimate $P(A_i|A_1, A_2, \cdots, A_{i-1}, C)$.

The main difference among Bayesian classification models is about the different way to calculate $P(A_i|A_1, A_2, \cdots, A_{i-1}, C)$. The simplest restricted Bayesian network classifier is the naive Bayesian classifier. Each attribute node A_i in the network is just dependent to the class node C, and $P(A_i|A_1, A_2, \cdots, A_{i-1}, C)$ in equation 3 can be simplified as $P(A_i|C)$.

TAN is another restricted Bayesian network classification model. In TAN classification model, the class node is the root and has no parents, i.e. $\prod_C = \emptyset$ ($\prod_C$ represents the set of parents of C). The class variable is a parent of each attribute variables, i. e. $C \in \prod_{A_i}$ ($\prod_{A_i}$ represents the set of parents of A_i, $i = 1, 2, \cdots, n$). And except for the class node, each attribute variable node has at most one other attribute variable node as its a parent, i.e. $|\prod_{A_i}| \leq 2$. Therefore $P(A_i|A_1, A_2, \cdots, A_{i-1}, C)$ in equation 3 can be simplified as $P(A_i|C)$ or $P(A_i|A_j, C)$, where $A_j \in \{A_1, A_2, \cdots, A_{i-1}\}$. In principle, there is no any restriction on the number of parents for any attribute node in the BAN classification model and general Bayesian network classification model. According to the criterion for selecting the dependence, the attribute node A_i might associate any other attribute nodes from $\{A_1, A_2, \cdots, A_{i-1}\}$. Each attribute node A_i may have more than two parents. In the following section, we try to describe a more strictly restricted Bayesian network classification model, which shows better performance than the TAN classifier in the most cases.

3 DLBAN: A Restricted Double Level Bayesian Network

Let $\{G_1, G_2\}$ be a partition of the attribute set $\{A_1, A_2, \cdots, A_n\}$, a variant of Bayes theorem can be written as follows.

$$P(C|G_1, G_2) = \frac{P(C|G_1)P(G_2|C, G_1)}{P(G_2|G_1)} \tag{4}$$

$$= \beta \cdot P(C|G_1) \cdot P(G_2|C, G_1) \tag{5}$$

where β is a normalization factor. Assume that $G_1 = \{A_{k_1}, A_{k_2}, \cdots, A_{k_m}\}$ and $G_2 = \{A_{l_1}, A_{l_2}, \cdots, A_{l_{n-m}}\}$, if the class label C and G_1 are given, and each attribute in the subset G_2 is conditionally independent of any other attribute

in the subset G_2, then a naive-Bayes-like simplifying independence assumption can be applied to the above formula.

$$P(C|G_1, G_2) = \beta \cdot P(C|A_{k_1}, A_{k_2}, \cdots, A_{k_m}) \cdot \prod_{i=1}^{n-m} P(A_{l_i}|C, A_{k_1}, A_{k_2}, \cdots, A_{k_m}) \tag{6}$$

Zheng and Webb [7] proposed the lazy Bayesian rule (LBR) learning technique, which can be viewed as a lazy approach to classification using this variant of Bayes theorem. The more attributes in subset G_1 the weaker the assumption required. However, a counter-balancing disadvantage of adding attribute values to G_1 is that the numbers of training instances from which the required conditional probabilities are estimated decrease and hence the accuracy of estimation can be expected to also decrease. In this paper, we restrict the number of attributes belonging to the subset G_1 is less than a fixed number. If all the attributes in the subset G_1 could be found from the attribute set $\{A_1, A_2, \cdots, A_n\}$, the other attributes in the subset G_2 are dependent on them. In the Bayesian network, all the attributes in the subset G_1 would be the common parents of the other attributes in the subset G_2. Therefore, $P(A_i|A_1, A_2, \cdots, A_{i-1}, C)$ in equation 2 can be simplified as $P(A_i|K_{A_i}, C)$, where K_{A_i} is the set of parents of node A_i, then the equation 6 can be written as:

$$\gamma \cdot P(C) \cdot \prod_{i=1}^{n} P(A_i|K_{A_i}, C) \tag{7}$$

where γ is a normalization factor.

Given G_1 and C, any attribute in G_2 is conditionally independent of other attributes in G_2. The class variable C is a parent of each attribute in A. Each attribute in G_1 may be the parents of each attribute in G_2. If a Bayesian network model satisfies these conditions, it is called a $DLBAN$ model.

There might be a certain dependence between any two attributes in $\{A_1, A_2, \cdots, A_n\}$, and the degree of dependence is different from each other between two attributes for two different categories. The mutual information can measure the degree of providing information between two attributes. In this paper, we use the conditional mutual information to represent dependence between attribute A_i and attribute A_j. Given the class C, the conditional mutual information of attribute A_i and attribute A_j is written as below.

$$I(A_i, A_j|C) = \sum_{A_i, A_j, C} P(A_i, A_j|C) log \frac{P(A_i, A_j|C)}{P(A_i|C) \cdot P(A_j|C)} \tag{8}$$

The attributes in G_1 are called stronger attributes, and the attributes in G_2 are called weaker attributes.

The learning algorithm of a $DLBAN$ model is described as follows.

1) The set of stronger attributes $G_1 = \emptyset$, the set of weaker attributes $G_2 = \{A_1, A_2, \cdots, A_n\}$, the threshold ϵ is a smaller real and the number of stronger attributes at most is k;

2) Evaluate classification performance of the current classifier, and save the evaluation result $OldAccuracy$;

3) Let $i = 1$, and $ImpAccuracy[i] = 0$;

4) If A_i does not belong to G_2, go to step 7), else continue;

5) Remove A_i from G_2 into G_1, assign A_i to parents of every weaker attributes in G_2, and evaluate classification performance of the current classifier, then save the evaluation result $ImpAccuracy[i]$;

6) Remove A_i from G_1 into G_2;

7) $i = i + 1$, if $i \leq n$, go to step 4), else go to step 8);

8) For $i = 1, 2, \cdots, n$, if $ImpAccuracy[i] - OldAccuracy < 0$, go to step 10); else choose A_i as stronger attribute, where $ImpAccuracy[i] - OldAccuracy$ is the maximum for all i, and remove A_i from G_2 in to G_1;

9) If the number of stronger attributes less than k, go to step 3), else go to step 10);

10) Compute the conditional mutual information $I(A_i, A_j|C)$ according to equation 8. If $I(A_i, A_j|C) > \epsilon$, return the arc from to A_i to A_j, else remove this arc.

Table 1. Descriptions of Data

	Domain	Size♯	Classes♯	Attributes♯	Missing Value
1	Car	1728	4	6	No
2	Contact-Lenses	24	3	5	No
3	Flare-C	1389	8	10	No
4	House-Votes-84	435	2	16	No
5	Iris Classification	150	3	4	Yes
6	Chess	3196	2	36	No
7	LED	1000	10	7	No
8	Lung Cancer	32	3	56	Yes
9	Mushroom	8124	2	22	Yes
10	Nursery	12960	5	8	No
11	Post-Operative	90	3	8	Yes
12	Promoter Gene Sequences	106	2	57	No
13	Soybean Large	683	19	35	Yes
14	Tic-Tac-Toe End Game	958	2	9	No
15	Zoology	101	7	16	No

4 Experimental Methodology and Results

We chose fifteen data sets (Table 1)from the UCI machine learning repository for our experiments. In data sets with missing value, we regarded missing value as a single value besides $Post - Operative$. For $Post - Operative$ data set, we

simply removed 3 instances with missing values from the data set. Our experiments have compared $DLBAN$ classifier with the naive Bayes classifier and a TAN classifier by the classification accuracy. The classification performance was evaluated by ten-folds cross-validation for all the experiments on each data set.All the experiments were performed in the $Weka$ system [9], which provides a workbench that includes full and working implementations of many popular learning schemes that can be used for practical data mining or for research.

Table 2 shows the classification accuracies. Boldface font indicates that the accuracy of $DLBAN$ is higher than that of TAN at a significance level better than 0.05 using a two-tailed pairwise t-test on the results of the 20 trials in a domain. From Table 2, the significant advantage of $DLBAN$ over TAN in terms of higher accuracy can be clearly seen. On average over the 15 domains, $DLBAN$ increases the accuracy of TAN by 2%. In 12 out of these fifteen domains, $DLBAN$ achieves significantly higher accuracy than TAN.

Table 2. Descriptions of Data

	Domain	Naive Bayes	TAN	DLBAN
1	Car	85.57570.32	91.60010.22	**94.33020.38**
2	Contact-Lenses	**72.76673.33**	65.83334.68	**72.76673.33**
3	Flare-C	79.01370.23	83.12090.31	**83.84080.19**
4	House-Votes-84	90.06900.14	93.19540.32	**94.17240.34**
5	Iris Classification	93.16670.73	91.75001.47	**93.40001.07**
6	Chess	87.89890.12	**93.44730.12**	87.89890.12
7	Led	73.88740.34	**73.96000.24**	73.88740.34
8	Lung Cancer	**53.21573.06**	46.09383.55	**51.49993.69**
9	Mushroom	95.76800.03	99.40900.03	**99.61470.05**
10	Nursery	90.28470.05	92.53190.23	**95.51280.16**
11	Post-Operative	**68.88890.86**	66.00001.63	**68.88900.86**
12	Promoter Gene Sequences	**91.27351.76**	82.97173.26	**91.27351.76**
13	Soybean Large	**92.73060.13**	87.35850.36	**92.65740.21**
14	Tic-Tac-Toe End Game	69.73900.32	**74.43941.17**	72.84970.80
15	Zoology	94.03251.04	95.52700.84	96.02300.29

On the data sets $Chess$ and $Tic-Tac-ToeEndGame$, the $DLBAN$ classifier was inferior to the TAN. In particular, the accuracy of the $DLBAN$ classifier is lower than the TAN classifier by 6% on $Chess$. Why comes this situation? In our experiments, the most number of stronger attributes is limited to three in order to avoid making the probability estimates of the attributes unreliable. However, whether three stronger attributes is enough for higher dimension attributes is worthy researching. Debugging the learning process on $Chess$, we found that the classification accuracy is increased from 84.6996% to 87.8989% as the number of stronger attributes is added from 1 to 3. If the number of stronger attributes will continue increasing, the classification accuracy maybe will continue increasing.

5 Conclusions

The naive Bayes classifier is a simple and efficient classification algorithm, but its independence assumption makes it unable to express the real dependence among attributes in the practical data. At present, many methods and techniques are brought to improve the performance of the naive Bayes classifier. In this paper, we present a new Bayesian model $DLBAN$, which can determine the dependence relationship among attributes by selecting some suitable attributes. It could not only extend the attributes number on which one attribute depends, but also determine the dependence relationship among attributes by searching the attribute space. The experimental results show that a $DLBAN$ classifier has a bit higher classification performance than the TAN classifier. In the process of learning $DLBAN$, it is very important to select the stronger attributes. The method we use is to select attributes according to their conditional mutual information value. Additionally, the maximum number of the stronger attributes is defined to be three in our experiments. In fact, different data sets might have its maximum suitable number of the stronger attributes.

References

1. Kononenko, I.: Semi-Naive Bayesian Classifier. In: Proceedings of European Conference on Artificial Intelligence, (1991) 206-219
2. Pearl J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Francisco, CA: Morgan Kaufman Publishers. 1988
3. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning, 29 (1997) 131–163
4. Chickering D. M.: Learning Bayesian networks is NP-Hard. Technical Report MSR-TR-94-17, Microsoft Research Advanced Technology Division, Microsoft Corporation, (1994)
5. Keogh, E. J., Pazzani, M. J.: Learning Augmented Bayesian Classifiers: A Comparison of Distribution-Based and Classification-Based Approaches. In: Proceedings of the Seventh International Workshop on Artificial Intelligence and Statistics. (1999) 225-230
6. Cheng J., Greiner R.: Comparing Bayesian Network Classifiers. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence (Laskey K. B. and Prade H. Eds.). San Franscico, CA: Morgan Kaufmann Publishers.(1999): 101-108.
7. Zheng, Z., Webb, G. I.: Lazy learning of Bayesian Rules. Machine Learning. Boston: Kluwer Academic Publishers.(2000) 1-35
8. Cheng J., Bell D. A., Liu W.: Learning Belief Networks from Data: An Informaition Theory Based Approach. In: Proceedings of the Sixth ACM International Conference on Information and Knowledge Management, (1997)
9. Witten, I. H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Seattle, WA: Morgan Kaufmann Publishers. (2000)