A Fundamental Issue of Naive Bayes

Harry Zhang[1] and Charles X. Ling[2]

[1] Faculty of Computer Science,
University of New Brunswick
Fredericton, New Brunswick, Canada E3B 5A3
email: hzhang@unb.ca
[2] Department of Computer Science,
The University of Western Ontario
London, Ontario, Canada N6A 5B7
email: ling@csd.uwo.ca

Abstract. Naive Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. But the conditional independence assumption on which it is based, is rarely true in real-world applications. Researchers extended naive Bayes to represent dependence explicitly, and proposed related learning algorithms based on dependence. In this paper, we argue that, from the classification point of view, dependence distribution plays a crucial role, rather than dependence. We propose a novel explanation on the superb classification performance of naive Bayes. To verify our idea, we design and conduct experiments by extending the *ChowLiu* algorithm to use the dependence distribution to construct TAN, instead of using mutual information that only reflects the dependencies among attributes. The empirical results provide evidences to support our new explanation.

1 Introduction

Classification is a fundamental issue in machine learning and data mining. In classification, the goal of a learning algorithm is to construct a classifier given a set of training examples with class labels. Typically, an example E is represented by a tuple of attribute values $(a_1, a_2, , \cdots, a_n)$, where a_i is the value of attribute A_i. Let C represent the classification variable which takes value $+$ or $-$.

A classifier is a function that assigns a class label to an example. From the probability perspective, according to Bayes Rule, the probability of an example $E = (a_1, a_2, \cdots, a_n)$ being class C is

$$p(C|E) = \frac{p(E|C)p(C)}{p(E)}.$$

Assume that all attributes are independent given the value of the class variable (conditional independence), we obtain a classifier $g(E)$, called a naive Bayesian classifier, or simply naive Bayes (NB).

$$g(E) = \frac{p(C=+)}{p(C=-)} \prod_{i=1}^{n} \frac{p(a_i|C=+)}{p(a_i|C=-)}. \tag{1}$$

It is obvious that the conditional independence assumption is rarely true in most real-world applications. A straightforward approach to overcome the limitation of naive Bayes is to extend its structure to represent explicitly the dependencies among attributes. Tree augmented naive Bayes (TAN) is an extended tree-like naive Bayes [3], in which an attribute node can have only one parent from another attribute node. Algorithms for learning TAN have been proposed, in which detecting dependences among attributes is a major approach. The *Chowliu* algorithm is a popular one based on dependence [1, 3], illustrated below.

1. Compute $I(A_i, A_j|C)$ between each pair of attributes, $i \neq j$.
2. Build a complete undirected graph in which the nodes are the attributes A_1, $\cdots$, A_n. Annotate the weight of an edge connecting A_i to A_j by $I(A_i, A_j|C)$.
3. Build a maximum weighted spanning tree.
4. Transform the resulting undirected tree to a directed one by choosing a root attribute and setting the direction of all edges to be outward from it.
5. Construct a TAN model by adding a node labeled by C and adding an arc from C to each A_i.

Essentially, the *ChowLiu* algorithm for learning TAN is based-on the conditional mutual information $I(A_i, A_j|C)$.

$$I(A_i, A_j|C) = \sum_{A_i, A_j, C} P(A_i, A_j, C) ln \frac{P(A_i, A_j|C)}{P(A_i|C)P(A_j|C)} \tag{2}$$

2 A New Explanation on the Classification Performance of Naive Bayes

2.1 Learning TAN Based on Dependence Distribution

When we look at the essence of the *ChowLiu* algorithm, we find that Equation 2 reflects the dependences between two attributes. We can transform the form of $I(A_i, A_j|C)$ to an equivalent equation below.

$$I(A_i, A_j|C) = \sum_{A_i, A_j} (P(A_i, A_j, +) ln \frac{P(A_i|A_j, +)}{P(A_i|+)} + P(A_i, A_j, -) ln \frac{P(A_i|A_j, -)}{P(A_i|-)}) \tag{3}$$

A question arises when you think of the meaning of $I(A_i, A_j|C)$. When

$$\frac{P(A_i|A_j, +)}{P(A_i|+)} > 1$$

and

$$\frac{P(A_i|A_j, -)}{P(A_i|-)} < 1,$$

intuitively, the dependencies between A_i and A_j in both class + and − support classifying E into class +. Thus, both evidences support classifying E into class +. Therefore, from the viewpoint of classification, the information association between A_i and A_j should be the sum of them, but they actually cancel each other in Equation 3. Similarly, when

$$\frac{P(A_i|A_j,+)}{P(A_i|+)} > 1$$

and

$$\frac{P(A_i|A_j,-)}{P(A_i|-)} > 1,$$

the two evidences support different classifications. Thus, in terms of classification, they should cancel each other out, but Equation 3 reflects the opposite fact. That reminds us that we should pay more attention to dependence distribution; i.e., how the dependencies among attributes distribute in two classes.

We modify $I(A_i, A_j|C)$ and obtain a conditional mutual information as below.

$$I_D(A_i, A_j|C) = \sum_{A_i,A_j} P(A_i, A_j)(ln\frac{P(A_i|A_j,+)}{P(A_i|+)} - ln\frac{P(A_i|A_j,-)}{P(A_i|-)})^2 \quad (4)$$

Actually, $I_D(A_i, A_j|C)$ represents the dependence distribution of A_i or A_j in two classes, which reflects the influence of the dependence between A_i and A_j on classification. From the above discussion, it is more reasonable to use dependence distribution to construct a classifier, rather than dependence.

We propose an extended *ChowLiu* algorithm for learning TAN, in which $I(A_i, A_j|C)$ is replaced by $I_D(A_i, A_j|C)$. We call this algorithm *ddr-ChowLiu*.

We have conducted empirical experiments to compare our *ddr-ChowLiu* algorithm to the *ChowLiu* algorithm. We use twelve datasets from the UCI repository [4] to conduct our experiments. Table 1 lists the properties of the datasets we use in our experiments.

Our experiments follow the procedure below:

1. The continuous attributes in the dataset are discretized by the entropy-based method.
2. For each dataset, run *ChowLiu* and *ddr-ChowLiu* with the 5-fold cross-validation, and obtain the classification accuracy on the testing set unused in the training.
3. Repeat 2 above 20 times and calculate the average classification accuracy on the testing data.

Table 2 shows the experimental results of average classification accuracies of *ChowLiu* and *ddr-ChowLiu*. We conduct an unpaired two-tailed t-test by using 95% as the confidence level and the better one for a given dataset is reported in bold. Table 2 shows that *ddr-ChowLiu* outperforms *ChowLiu* in five datasets, losses in three datasets, and ties in four datasets. Overall, the experimental results show that *ddr-ChowLiu* outperforms slightly *ChowLiu*. Therefore, if we use

Table 1. Description of the datasets used in the experiments of comparing the *ddr-ChowLiu* algorithm to the *Chowliu* algorithm.

Dataset	Attributes	Class	Instances
Australia	14	2	690
breast	10	10	683
cars	7	2	700
dermatology	34	6	366
ecoli	7	8	336
hepatitis	4	2	320
import	24	2	204
iris	5	3	150
pima	8	2	392
segment	19	7	2310
vehicle	18	4	846
vote	16	2	232

directly dependence distribution, instead of using dependence, it will result in a better classifier. Further, this experiment provides evidence that it is dependence distribution that determines classification, not dependence itself.

Table 2. Experimental results of the accuracies of *ChowLiu* and *ddr-ChowLiu*.

Dataset	***ChowLiu***	***ddr-ChowLiu***
Australia	**76.7±0.32**	76.1±0.33
breast	73.3±0.37	73.3±0.33
cars	85.4±0.37	**87.1±0.28**
dermatology	97.7±0.17	97.7±0.17
ecoli	**96.1±0.23**	95.8±0.20
hepatitis	70.5±0.42	70.5±0.51
import	93.6±0.37	**95.6±0.34**
iris	91.2±0.48	91.3±0.50
pima	70.5±0.46	**71.8±0.51**
segment	82.3±0.17	**82.4±0.16**
vehicle	**89.3±0.23**	85.7±0.30
vote	78.6±0.61	**79.1±0.53**

2.2 A Novel Explanation for Naive Bayes

From Section 2.1, we observed that how dependence distributes in two classes determines classification, and the empirical experimental results provided evidence to support our claim. In fact, we can generalize this observation. In a given dataset, two attributes may depend on each other, but the dependence may distribute evenly in each class. Clearly, in this case, the conditional independence assumption is violated, but naive Bayes is still the optimal classifier. Further, what eventually affects classification is the combination of dependencies among all attributes. If we just look at two attributes, there may exist strong dependence between them that affects classification. When the dependencies among all attributes work together, however, they may cancel each other out and no longer affect classification. Therefore, we argue that it is distribution of dependencies among all attributes over classes that affects classification of naive Bayes, not merely dependencies themselves. This explains why naive Bayes still works well on the datasets in which strong dependencies among attributes do exist [2].

3 Conclusions

In this paper, we investigated the *Chowliu* algorithm and proposed an extended algorithm for learning TAN that is based on dependence distribution, rather than dependence. The experimental results showed that the new algorithm outperforms the *Chowliu* algorithm. We generalized that observation, and proposed a new explanation on the classification performance of naive Bayes. We argue that, essentially, the dependence distribution; i.e., how the local dependence of an attribute distributes in two classes, evenly or unevenly, and how the local dependencies of all attributes work together, consistently (support a certain classification) or inconsistently (cancel each other out), plays a crucial role in classification. We explain why even with strong dependencies, naive Bayes still works well; i.e., when those dependencies cancel each other out, there is no influence on classification. In this case, naive Bayes is still the optimal classifier.

References

1. Chow, C. K., Liu, C. N.: Approximating Discrete Probability Distributions with Dependence Trees. IEEE Trans. on Information Theory, Vol. 14 (1968), 462–467.
2. Domingos P., Pazzani M.: Beyond Independence: Conditions for the Optimality of the Simple Bayesian Classifier. Machine Learning **29** (1997) 103-130
3. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian Network Classifiers. Machine Learning, Vol: 29 (1997), 131–163.
4. Merz, C., Murphy, P., Aha, D.: UCI Repository of Machine Learning Databases. In: Dept of ICS, University of California, Irvine (1997). http://www.www.ics.uci.edu/mlearn/MLRepository.html.