International Journal of Pattern Recognition
and Artificial Intelligence
Vol. 19, No. 2 (2005) 183–198

World Scientific
www.worldscientific.com

EXPLORING CONDITIONS FOR THE OPTIMALITY OF NAÏVE BAYES

HARRY ZHANG

Faculty of Computer Science, University of New Brunswick
Fredericton, New Brunswick, Canada, E3B 5A3
hzhang@unb.ca

Naïve Bayes is one of the most efficient and effective inductive learning algorithms for machine learning and data mining. Its competitive performance in classification is surprising, because the conditional independence assumption on which it is based is rarely true in real-world applications. An open question is: what is the true reason for the surprisingly good performance of Naïve Bayes in classification?

In this paper, we propose a novel explanation for the good classification performance of Naïve Bayes. We show that, essentially, dependence distribution plays a crucial role. Here dependence distribution means how the local dependence of an attribute distributes in each class, evenly or unevenly, and how the local dependences of all attributes work together, consistently (supporting a certain classification) or inconsistently (canceling each other out). Specifically, we show that no matter how strong the dependences among attributes are, Naïve Bayes can still be optimal if the dependences distribute evenly in classes, or if the dependences cancel each other out. We propose and prove a sufficient and necessary condition for the optimality of Naïve Bayes. Further, we investigate the optimality of Naïve Bayes under the Gaussian distribution. We present and prove a sufficient condition for the optimality of Naïve Bayes, in which the dependences among attributes exist. This provides evidence that dependences may cancel each other out.

Our theoretic analysis can be used in designing learning algorithms. In fact, a major class of learning algorithms for Bayesian networks are conditional independence-based (or CI-based), which are essentially based on dependence. We design a dependence distribution-based algorithm by extending the *ChowLiu* algorithm, a widely used CI based algorithm. Our experiments show that the new algorithm outperforms the *ChowLiu* algorithm, which also provides empirical evidence to support our new explanation.

Keywords: Naïve Bayes; optimality; classification.

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 $(x_1, x_2, \ldots, x_n)$, where x_i is the value of attribute X_i. Let C represent the class variable, and let c be the value of C. In this paper, we assume that there are only two classes: $+$ (the positive class) and $-$ (the negative class).

A classifier is a function that assigns a class label to an example. From the probability perspective, according to the Bayes rule, the probability of an example $E = (x_1, x_2, \ldots, x_n)$ being class c is

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

E is classified as the class $C = +$ if and only if

$$f_b(E) = \frac{p(C = +|E)}{p(C = -|E)} \geq 1, \tag{1}$$

where $f_b(E)$ is called a Bayesian classifier.

Assume that all attributes are independent given the class; that is,

$$p(E|c) = p(x_1, x_2, \ldots, x_n|c) = \prod_{i=1}^{n} p(x_i|c).$$

The resulting classifier is then:

$$f_{nb}(E) = \frac{p(C = +)}{p(C = -)} \prod_{i=1}^{n} \frac{p(x_i|C = +)}{p(x_i|C = -)}. \tag{2}$$

The function $f_{nb}(E)$ is called a Naïve Bayesian classifier, or simply Naïve Bayes (NB). Figure 1 shows graphically the structure of Naïve Bayes. In Naïve Bayes, each attribute node has no parent except the class node.

Naïve Bayes is the simplest form of a Bayesian network. In Naïve Bayes, all attributes are independent of each other given the class. This assumption is called the conditional independence assumption. It is obvious that the conditional independence assumption is rarely true in most real-world applications. A straightforward approach to overcome the limitation of Naïve Bayes is to extend its structure to represent explicitly the dependences among attributes. Tree augmented Naïve Bayes (TAN) is an extended tree-like Naïve Bayes,[8] in which the class node points directly to all attribute nodes and an attribute node can have only one parent from another attribute node (in addition to the class node). Figure 2 shows an example of TAN. TAN is a specific case of general augmented Naïve Bayesian networks or

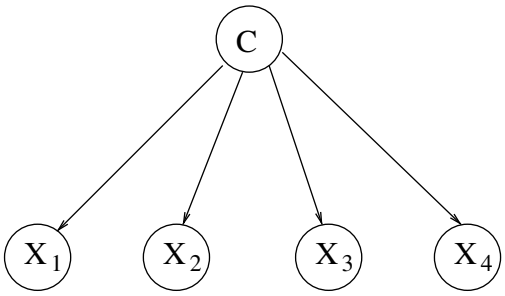


Fig. 1. An example of Naïve Bayes.

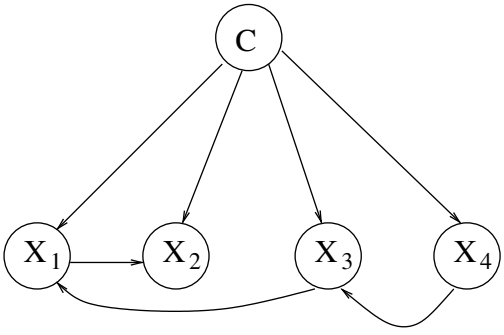


Fig. 2. An example of TAN.

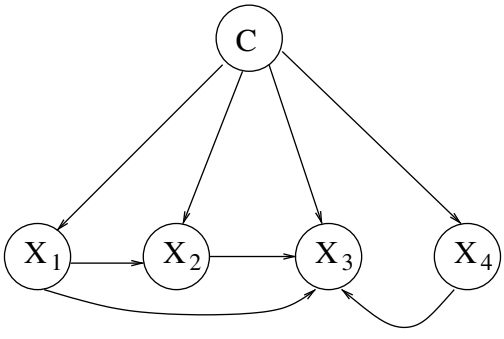


Fig. 3. An example of ANB.

simply augmented Naïve Bayes (ANB), in which the class node also points directly to all attribute nodes, but there is no limitation on the links among attribute nodes (except that they do not form any directed cycle). Figure 3 shows an example of ANB. From the view of probability, an ANB G represents a joint probability distribution, below.

$$p_G(X_1, \ldots, X_n, C) = p(C) \prod_{i=1}^{n} p(X_i | pa_i, C), \tag{3}$$

where pa_i denotes the parents of X_i from attribute nodes. ANB is a special form of Bayesian networks in which no node is specified as a class node. It has been shown that any Bayesian network can be represented by an ANB.[19] Therefore, any joint probability distribution can be represented by an ANB.

When we apply a logarithm to $f_b(E)$ in Eq. (1), the resulting classifier $\log f_b(E)$ is the same as $f_b(E)$, in the sense that an example E belongs to the positive class, if and only if $\log f_b(E) \geq 0$. f_{nb} in Eq. (2) is similar. In this paper, we assume that, given a classifier f, an example E belongs to the positive class, if and only if $f(E) \geq 0$.

It has been observed that Naïve Bayes works well in classification.[11,12,15] The reason, however, is unknown. This paper is motivated by exploring the underlying reason. The remainder of this paper is organized as follows. Section 2 introduces the related work. In Sec. 3, we propose a new explanation for the good performance of Naïve Bayes. In Sec. 4, we investigate the optimality of Naïve Bayes under the Gaussian distribution. Section 5 presents a new algorithm for learning TAN based on the idea in Sec. 3. This paper concludes with discussion and directions for future work.

2. Related Work

Many empirical comparisons between Naïve Bayes and modern decision-tree algorithms such as C4.5[16] have shown that Naïve Bayes predicts equally well as C4.5.[11,12,15] The good performance of Naïve Bayes is surprising because it makes an assumption that is almost always violated in real-world applications: given the class, all attributes are independent.

An open question is, what is the true reason for the surprisingly good performance of Naïve Bayes on most classification tasks? Intuitively, since the conditional independence assumption on which it is based almost never holds, its performance should be poor. It has been observed, however, that its classification accuracy does not depend on the dependences among attributes; i.e. Naïve Bayes may still have high accuracy on the datasets in which strong dependences exist.[4]

Domingos and Pazzani[4] present an explanation that Naïve Bayes owes its good performance to the zero-one loss function. This function defines the error as the number of incorrect classifications.[7] Unlike other loss functions, such as the squared error, the zero-one loss function does not penalize inaccurate probability estimation as long as the maximum probability is assigned to the correct class. That means that Naïve Bayes may change the posterior probabilities of each class, but the class with the maximum posterior probability is often unchanged. Thus, the classification is still correct, although the probability estimation is poor. For example, let us assume that the true probabilities $p(+|E)$ and $p(-|E)$ are 0.9 and 0.1, respectively, and that the probability estimates $p'(+|E)$ and $p'(-|E)$ produced by Naïve Bayes are 0.6 and 0.4. Obviously, the probability estimates are poor, but the classification (positive) is not affected.

Domingos and Pazzani's explanation[4] is verified by the work of Frank *et al.*,[6] which shows that the performance of Naïve Bayes is much worse when it is used for regression (predicting a continuous value). Moreover, evidence exists that Naïve Bayes produces poor probability estimates.[1,14]

In our opinion, however, Domingos and Pazzani's[4] explanation does not uncover why the strong dependences among attributes could not flip the classification. For the preceding example, why could the dependences not make the probability estimates $p'(+|E)$ and $p'(-|E)$ produced by Naïve Bayes be 0.4 and 0.6? The key point here is that we need to know how dependence affects classification, and under what conditions dependence does not affect classification.

Some other related works explore the properties of Naïve Bayes,[9,10,17,18] but none of them give an explicit condition for the optimality of Naïve Bayes. Cooper[3] suggests a strategy to avoid the need of the conditional independence assumption. In the case of only two classes, the usual conditional independence assumption can be replaced by the weaker "linked dependence" assumption below:

$$\frac{p(x_1, x_2, \ldots, x_n|+)}{p(x_1, x_2, \ldots, x_n|-)} = \prod_{i=1}^{n} \frac{p(x_i|+)}{p(x_i|-)}.$$

However, it still does not explain why Naïve Bayes works well when the conditional independence assumption is violated.

In this paper, we propose a new explanation, that the classification of Naïve Bayes is essentially affected by dependence distribution, instead of dependence. In addition, we present a sufficient condition for the optimality of Naïve Bayes under the Gaussian distribution. Further, we present a new learning algorithm for TAN based on dependence distribution, which slightly outperforms the traditional dependence-based learning algorithm.

3. A New Explanation of the Good Classification Performance of Naïve Bayes

In this section, we propose a new explanation of the good classification performance of Naïve Bayes. The basic idea comes from the following observation. In a given dataset, two attributes may depend on each other, but the dependence may distribute evenly in each class. Clearly, in that case, the conditional independence assumption is violated, but Naïve Bayes is still the optimal classifier. Further, what eventually affects classification is the combination of dependences among all attributes. If we look at just two attributes, there may exist strong dependence between them that affects classification. When the dependences among all attributes work together, however, they may cancel each other out and no longer affect classification. Therefore, we argue that it is the distribution of dependences among all attributes that affects classification, not merely the dependences themselves.

Before discussing the details, we introduce the formal definition of the equivalence of two classifiers under zero-one loss, which is used as a basic concept.

Definition 1. Given an example E, two classifiers f_1 and f_2 are said to be equal under zero-one loss on E, denoted by $f_1(E) \doteq f_2(E)$, if $f_1(E) \geq 0$ if and only if $f_2(E) \geq 0$. If for every example E in the example space, $f_1(E) \doteq f_2(E)$, f_1 and f_2 are said to be equal under zero-one loss, denoted by $f_1 \doteq f_2$.

3.1. *Local dependence distribution*

As discussed in Sec. 1, ANB can represent any joint probability distribution. Thus we choose an ANB as the underlying probability distribution. Our motivation is

to find out under what conditions Naïve Bayes classifies identically as the underlying ANB.

Assume that the underlying probability distribution is an ANB G with two classes $\{+, -\}$, and the dependences among attributes are represented by the arcs among attribute nodes. For each node, the influence of its parents is quantified by the correspondent conditional probabilities. We call the dependence between a node and its parents *local dependence* of this node. How do we measure the local dependence of a node in each class? Naturally, the ratio of the conditional probability of the node given its parents over the conditional probability of the node without its parents, reflects how strong its parents affect the node in each class. Thus we have the following definition.

Definition 2. For a node X on ANB G, the local dependence derivatives of X at $X = x$ in classes $+$ and $-$ are defined as below.

$$dd_G^+(x|pa(x)) = \frac{p(x|pa(x), +)}{p(x|+)}, \tag{4}$$

$$dd_G^-(x|pa(x)) = \frac{p(x|pa(x), -)}{p(x|-)}. \tag{5}$$

Essentially, $dd_G^+(x|pa(x))$ reflects the strength of the local dependence of node X in class $+$, which measures the influence of X's local dependence on classification in class $+$. $dd_G^-(x|pa(x))$ is similar. Further, we have the following observations.

(1) When X has no parent, then

$$dd_G^+(x|pa(x)) = dd_G^-(x|pa(x)) = 1.$$

(2) When $dd_G^+(x|pa(x)) \geq 1$, X's local dependence in class $+$ supports the classification of $C = +$. Otherwise, it supports the classification of $C = -$. Similarly, when $dd_G^-(x|pa(x)) \geq 1$, X's local dependence in class $-$ supports the classification of $C = -$. Otherwise, it supports the classification of $C = +$.

Intuitively, when the local dependences in two classes support different classifications, they partially cancel each other out, and the final classification that the local dependence supports is the class with the greater local dependence derivative. A different case is that of the local dependences in two classes supporting the same classification. Then, they work together to support that classification.

The preceding discussion shows that the ratio of the local dependence derivatives in both classes ultimately determines which classification the local dependence of a node supports. Thus we have the following definition.

Definition 3. For a node X on ANB G, the local dependence derivative ratio of X at $X = x$, denoted by $ddr_G(x)$, is defined below:

$$ddr_G(x) = \frac{dd_G^+(x|pa(x))}{dd_G^-(x|pa(x))}. \tag{6}$$

From Definition 3, $ddr_G(x)$ quantifies the influence of X's local dependence on classification. Further, we have the following observations.

(1) If X has no parents, $ddr_G(x) = 1$.
(2) If $dd^+_G(x|pa(x)) = dd^-_G(x|pa(x))$, $ddr_G(x) = 1$. This means that X's local dependence at $X = x$ distributes evenly in class + and class −. Thus, the dependence does not affect classification, no matter how strong is the dependence.
(3) If $ddr_G(x) > 1$, X's local dependence at $X = x$ in class + is stronger than that in class −. $ddr_G(x) < 1$ means the opposite.

3.2. *Global dependence distribution*

Now we investigate how the local dependences of all attributes work together, and explore under what condition an ANB works exactly the same as its correspondent Naïve Bayes. The following theorem establishes the relation of an ANB and its correspondent Naïve Bayes.

Theorem 1. *Given an ANB G and its correspondent Naïve Bayes G_{nb} (i.e. remove all the arcs among attribute nodes from G) on attributes $X_1, X_2, \ldots, X_n$, assume that f_b and f_{nb} are the classifiers corresponding to G and G_{nb}, respectively. For example $E = (x_1, x_2, \ldots, x_n)$, the equation below is true.*

$$f_b(x_1, x_2, \ldots, x_n) = f_{nb}(x_1, x_2, \ldots, x_n) \prod_{i=1}^{n} ddr_G(x_i), \tag{7}$$

where $\prod_{i=1}^{n} ddr_G(x_i)$ is called the dependence distribution factor at example E, denoted by $DF_G(E)$.

Proof. According to Eq. (3), we have:

$$\begin{aligned} f_b(x_1, \ldots, x_n) &= \frac{p(+)}{p(-)} \prod_{i=1}^{n} \frac{p(x_i|pa(x_i), +)}{p(x_i|pa(x_i), -)} \\ &= \frac{p(+)}{p(-)} \prod_{i=1}^{n} \frac{p(x_i|+)}{p(x_i|-)} \prod_{i=1}^{n} \frac{p(x_i|pa(x_i), +)p(x_i|-)}{p(x_i|pa(x_i), -)p(x_i|+)} \\ &= f_{nb}(E) \prod_{i=1}^{n} \frac{ddr^+_G(x_i|pa(x_i))}{ddr^-_G(x_i|pa(x_i))} \\ &= f_{nb}(E) \prod_{i=1}^{n} ddr_G(x_i) \\ &= DF_G(E) f_{nb}(E). \end{aligned} \tag{8}$$

□

From Theorem 1, we know that, in fact, it is the dependence distribution factor $DF_G(E)$ that determines the difference between an ANB and its correspondent

Naïve Bayes in classification. Further, $DF_G(E)$ is the product of local dependence derivative ratios of all nodes. Therefore, it reflects the global dependence distribution (how each local dependence distributes in each class, and how all local dependences work together). For example, when $DF_G(E) = 1$, G has the same classification as G_{nb} on E. However, it is not a necessary condition. The theorem below presents a sufficient and necessary condition.

Theorem 2. *Given an example $E = (x_1, x_2, \ldots, x_n)$, an ANB G is equal to its correspondent Naïve Bayes G_{nb} under zero-one loss; i.e. $f_b(E) \doteq f_{nb}(E)$, if and only if $f_b(E) \geq 1$, $DF_G(E) \leq f_b(E)$; or when $f_b(E) < 1$, $DF_G(E) > f_b(E)$.*

Proof. The proof is straightforward from Definition 1 and Theorem 1. □

From Theorem 2, if the distribution of dependences among attributes satisfies certain conditions, Naïve Bayes classifies exactly the same as the underlying ANB, even though there may exist strong dependences among attributes. Moreover, we have the following observations:

(1) When $DF_G(E) = 1$, the dependences in ANB G has no influence on classification. That is, the classification of G is exactly the same as its correspondent Naïve Bayes G_{nb}. There exist three cases for $DF_G(E) = 1$:
   - no dependence exists among attributes,
   - for each attribute X on G, $ddr_G(x) = 1$; that is, the local dependence of each node distributes evenly in two classes,
   - the influence that some local dependences support classifying E into $C = +$ is fully canceled out by the influence that other local dependences support classifying E into $C = -$.

(2) $DF_G(E) = 1$ is only a sufficient, not necessary, condition for $f_b(E) \doteq f_{nb}(E)$. Theorem 2 gives a sufficient and necessary condition, and explains why Naïve Bayes still produces accurate classification even in the datasets with strong dependences among attributes.

(3) The dependences in an ANB flip (change) the classification of its correspondent Naïve Bayes, only if the condition given by Theorem 2 is not true.

Theorem 2 represents a sufficient and necessary condition for the optimality of Naïve Bayes on example E. If for each example E in the example space, $f_b(E) \doteq f_{nb}(E)$; i.e. $f_b \doteq f_{nb}$, then Naïve Bayes is globally optimal.

4. Conditions for the Optimality of Naïve Bayes

In Sec. 3, we proposed that Naïve Bayes is optimal if the dependences among attributes cancel each other out. That is, under the circumstance, Naïve Bayes is still optimal even though dependences do exist. In this section, we investigate Naïve Bayes under the multivariate Gaussian distribution and prove a sufficient condition

for the optimality of Naïve Bayes, assuming the dependences among attributes exist. That provides us with theoretic evidence that the dependences among attributes may cancel each other out.

Let us restrict our discussion to two attributes X_1 and X_2, and assume that the class density is a multivariate Gaussian in both the positive and negative classes. That is,

$$p(x_1, x_2, +) = \frac{1}{2\pi|\sum_+|^{1/2}} e^{-\frac{1}{2}(x-\mu^+)^T \sum_+^{-1}(x-\mu^+)},$$

$$p(x_1, x_2, -) = \frac{1}{2\pi|\sum_-|^{1/2}} e^{-\frac{1}{2}(x-\mu^-)^T \sum_-^{-1}(x-\mu^-)},$$

where $x = (x_1, x_2)$, $\sum_+$ and $\sum_-$ are the covariance matrices in the positive and negative classes respectively, $|\sum_-|$ and $|\sum_+|$ are the determinants of $\sum_-$ and $\sum_+$, $\sum_+^{-1}$ and $\sum_-^{-1}$ are the inverses of $\sum_-$ and $\sum_+$; $\mu^+ = (\mu_1^+, \mu_2^+)$ and $\mu^- = (\mu_1^-, \mu_2^-)$, μ_i^+ and μ_i^- are the means of attribute X_i in the positive and negative classes respectively, $i = 1, 2$, and $(x - \mu^+)^T$ and $(x - \mu^-)^T$ are the transposes of $(x - \mu^+)$ and $(x - \mu^-)$.

We assume that two classes have a common covariance matrix $\sum_+ = \sum_- = \sum$, and X_1 and X_2 have the same variance σ in both classes. Then, when applying a logarithm to the Bayesian classifier, defined in Eq. (1), we obtain the classifier f_b below.

$$\begin{aligned} f_b(x_1, x_2) &= \log \frac{p(x_1, x_2, +)}{p(x_1, x_2, -)} \\ &= -\frac{1}{\sigma^2}(\mu^+ + \mu^-)\sum^{-1}(\mu^+ - \mu^-) + x^T \sum^{-1}(\mu^+ - \mu^-). \end{aligned}$$

Then, because of the conditional independence assumption, we have the correspondent Naïve Bayes f_{nb} below.

$$f_{nb}(x_1, x_2) = \frac{1}{\sigma^2}(\mu_1^+ - \mu_1^-)x_1 + \frac{1}{\sigma^2}(\mu_2^+ - \mu_2^-)x_2.$$

Assume that

$$\sum = \begin{pmatrix} \sigma & \sigma_{12} \\ \sigma_{12} & \sigma \end{pmatrix}.$$

X_1 and X_2 are independent if $\sigma_{12} = 0$. If $\sigma \neq \sigma_{12}$, we have

$$\sum^{-1} = \begin{pmatrix} \frac{-\sigma}{\sigma_{12}^2 - \sigma^2} & \frac{\sigma_{12}}{\sigma_{12}^2 - \sigma^2} \\ \frac{\sigma_{12}}{\sigma_{12}^2 - \sigma^2} & \frac{-\sigma}{\sigma_{12}^2 - \sigma^2} \end{pmatrix}.$$

Note that an example E is classified into the positive class by f_b, if and only if $f_b \geq 0$. f_{nb} is similar. Thus, when f_b or f_{nb} is divided by a nonzero positive constant, the resulting classifier is the same as f_b or f_{nb}. Then,

$$f_{nb}(x_1, x_2) = (\mu_1^+ - \mu_1^-)x_1 + (\mu_2^+ - \mu_2^-)x_2, \tag{9}$$

and

$$f_b(x_1, x_2) = \frac{1}{\sigma_{12}^2 - \sigma^2}(\sigma_{12}(\mu_2^+ - \mu_2^-) - \sigma(\mu_1^+ - \mu_1^-))x_1 + \frac{1}{\sigma_{12}^2 - \sigma^2}(\sigma_{12}(\mu_1^+ - \mu_1^-) - \sigma(\mu_2^+ - \mu_2^-))x_2 + a, \tag{10}$$

where $a = -\frac{1}{\sigma^2}(\mu^+ + \mu^-)\sum^{-1}(\mu^+ - \mu^-)$, a constant independent of x.

For any x_1 and x_2, Naïve Bayes has the same classification as the Bayesian classifier if

$$f_b(x_1, x_2) f_{nb}(x_1, x_2) \geq 0. \tag{11}$$

That is,

$$\begin{aligned}\frac{1}{\sigma_{12}^2 - \sigma^2}&((\sigma_{12}(\mu_1^+ - \mu_1^-)(\mu_2^+ - \mu_2^-) - \sigma(\mu_1^+ - \mu_1^-)^2)x_1^2 \\ &+ (\sigma_{12}(\mu_1^+ - \mu_1^-)(\mu_2^+ - \mu_2^-) - \sigma(\mu_2^+ - \mu_2^-)^2)x_2^2 \\ &+ (2\sigma_{12}(\mu_1^+ - \mu_1^-)(\mu_2^+ - \mu_2^-) - \sigma((\mu_1^+ - \mu_1^-)^2 + (\mu_2^+ - \mu_2^-)^2))x_1x_2) \\ &+ a(\mu_1^+ - \mu_1^-)x_1 + a(\mu_2^+ - \mu_2^-)x_2 \geq 0.\end{aligned} \tag{12}$$

Equation (12) represents a sufficient and necessary condition for $f_{nb}(x_1, x_2) \doteq f_b(x_1, x_2)$. But it is too complicated. Let $(\mu_1^+ - \mu_1^-) = (\mu_2^+ - \mu_2^-)$. Equation (12) is simplified as below.

$$w_1(x_1 + x_2)^2 + w_2(x_1 + x_2) \geq 0, \tag{13}$$

where $w_1 = \frac{(\mu_1^+ - \mu_1^-)^2}{\sigma_{12} + \sigma}$, and $w_2 = a(\mu_1^+ - \mu_1^-)$. Let $x = x_1 + x_2$, and $y = w_1(x_1 + x_2)^2 + w_2(x_1 + x_2)$. Figure 4 shows the area in which Naïve Bayes classifies identically as the Bayesian classifier.

The following theorem presents a sufficient condition that Naïve Bayes works identically as the Bayesian classifier.

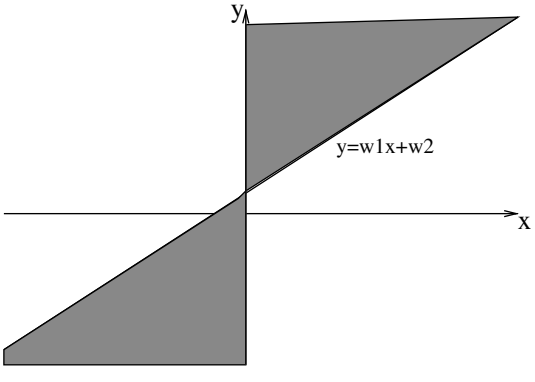


Fig. 4. Naïve Bayes classifies identically as the Bayesian classifier in the shaded areas.

Theorem 3. *$f_b \doteq f_{nb}$, if one of the following two conditions is true*:

(1) $\mu_1^+ = -\mu_2^-$, $\mu_1^- = -\mu_2^+$, *and* $\sigma_{12} + \sigma > 0$.
(2) $\mu_1^+ = \mu_2^-$, $\mu_2^+ = \mu_1^-$, *and* $\sigma_{12} - \sigma > 0$.

Proof. (1) If $\mu_1^+ = -\mu_2^-$, $\mu_1^- = -\mu_2^+$, then $(\mu_1^+ - \mu_1^-) = (\mu_2^+ - \mu_2^-)$. It is straightforward to verify that $-\frac{1}{\sigma^2}(\mu^+ + \mu^-)\sum^{-1}(\mu^+ - \mu^-) = 0$. That is, for the constant a in Eq. (10), we have $a = 0$. Since $\sigma_{12} + \sigma > 0$, Eq. 13 is always true for any x_1 and x_2. Therefore, $f_b \doteq f_{nb}$.
(2) If $\mu_1^+ = \mu_2^-$, $\mu_2^+ = \mu_1^-$, then $(\mu_1^+ - \mu_1^-) = -(\mu_2^+ - \mu_2^-)$, and $a = 0$. Thus, Eq. (12) is simplified as below.

$$\frac{(\mu_1^+ - \mu_1^-)^2}{\sigma_{12} - \sigma}(x_1 + x_2)^2 \geq 0. \tag{14}$$

It is obvious that Eq. 14 is true for any x_1 and x_2, if $\sigma_{12} - \sigma > 0$. Therefore, $f_b \doteq f_{nb}$. □

Theorem 3 represents an explicit condition that Naïve Bayes is globally optimal. It shows that Naïve Bayes is still optimal under certain conditions, even though the conditional independence assumption is violated. In other words, the conditional independence assumption is not the necessary condition for the optimality of Naïve Bayes. This provides evidence that the dependence distribution may play the crucial role in classification.

5. Learning TAN Based on Dependence Distribution

Learning Bayesian networks from data has received considerable attention in recent years, and many learning algorithms have been proposed. A major class of those learning algorithms are based on conditional independence among attributes, called CI-based algorithms. In other words, those algorithms are based on dependence. For example, conditional mutual information, depicted in Eq. (15), has often been used to measure the dependence between two attributes.

$$I(X_i, X_j|C) = \sum_{x_i, x_j, c} p(x_i, x_j, c) \ln \frac{p(x_i, x_j|c)}{p(x_i|c)p(x_j|c)}. \tag{15}$$

Those dependence-based algorithms, however, emphasize the strength of dependences among attributes, not the influence of dependences on classification. For example, the conditional mutual information in Eq. (15) reflects actually the dependence between two attributes, not the influence of dependence on classification. To make this point clear, we transform Eq. (15) into an equivalent equation below.

$$I(X_i, X_j|C) = \sum_{x_i, x_j} \left(p(x_i, x_j, +) \ln \frac{p(x_i|x_j, +)}{p(x_i|+)} + p(x_i, x_j, -) \ln \frac{p(x_i|x_j, -)}{p(x_i|-)} \right). \tag{16}$$

A question arises when one thinks of the meaning of $I(X_i, X_j|C)$. When

$$\frac{p(x_i|x_j,+)}{p(x_i|+)} > 1$$

and

$$\frac{p(x_i|x_j,-)}{p(x_i|-)} < 1,$$

intuitively, the dependences between X_i and X_j at $X_i = x_i$ and $X_j = x_j$ in both class $+$ and $-$ support classifying E into class $+$. Thus, in both cases, evidence supports classifying E into class $+$. Therefore, from the viewpoint of classification, the information association between X_i and X_j should be the sum of them, but they actually cancel out each other in Eq. (16). Similarly, when

$$\frac{p(x_i|x_j,+)}{p(x_i|+)} > 1$$

and

$$\frac{p(x_i|x_j,-)}{P(x_i|-)} > 1,$$

in both cases, evidence supports different classifications. Thus, in terms of classification, they should cancel each other out, but Eq. (16) reflects the opposite fact.

When we consider the influence of dependences on classification, as discussed in Sec.3, dependence distribution plays a crucial role. We modify $I(X_i, X_j|C)$ and obtain a conditional mutual information $I_D(X_i, X_j|C)$ to reflect the dependence distribution as below.

$$I_D(X_i, X_j|C) = \sum_{x_i,x_j} p(x_i, x_j) \left(\ln \frac{p(x_i|x_j,+)}{p(x_i|+)} - \ln \frac{p(x_i|x_j,-)}{p(x_i|-)} \right)^2. \tag{17}$$

Actually, $I_D(X_i, X_j|C)$ reflects the dependence derivative ratio (see Definition 3) of X_i or X_j; i.e. the influence of dependence between X_i and X_j on classification. From the preceding discussion, it is reasonable to use dependence derivative ratio to construct a classifier, rather than dependence.

To verify that dependence distribution plays a more important role in classification, we extend the *ChowLiu* algorithm[8] that is based on Eq. (15) to a dependence distribution-based algorithm using Eq. (17). Our new algorithm is identical to *ChowLiu*, except $I(X_i, X_j|C)$ is replaced by $I_D(X_i, X_j|C)$. We call this algorithm *ddr-ChowLiu*, depicted below.

Algorithm *ddr-ChowLiu*

(1) Compute $I_D(X_i, X_j|C)$ between each pair of attributes, $i \neq j$.
(2) Build a complete undirected graph in which the nodes are the attributes $X_1, \ldots, X_n$. Annotate the weight of an edge connecting X_i to X_j by $I_D(X_i, X_j|C)$.
(3) Build a maximum weighted spanning tree.

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

(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 X_i.

We have conducted empirical experiments to compare our *ddr-ChowLiu* algorithm to the *ChowLiu* algorithm. We use twelve datasets from the UCI repository[13] to conduct our experiments. Table 1 lists the properties of the datasets that we use in our experiments. For the datasets with more than two classes, we extend Eq. (17) to the following:

$$I_D(X_i, X_j|C) = \sum_{x_i,x_j,c} p(x_i, x_j) \left(\ln \frac{p(x_i|x_j, c)}{p(x_i|c)} - \text{Avg}(X_i, X_j, C) \right)^2, \tag{18}$$

where $\text{Avg}(X_i, X_j, C)$ is defined below.

$$\text{Avg}(X_i, X_j, C) = \sum_{x_i,x_j,c} \frac{\ln \frac{p(x_i|x_j,c)}{p(x_i|c)}}{|c|}, \tag{19}$$

where $|C|$ is the number of classes.

Our experiments follow the procedure below:

(1) The continuous attributes in the dataset are discretized by Fayyad and Irani's entropy-based method.[5]
(2) For each dataset, run *ChowLiu* and *ddr-ChowLiu* with the five-fold cross-validation, and obtain the classification accuracy on the testing set unused in the training.
(3) Repeat Step 2 twenty times and calculate the average classification accuracy on the testing data.

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

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

Table 2 shows the experimental results of average classification accuracies of *ChowLiu* and *ddr-ChowLiu*.

We conduct an unpaired two-tailed t-test 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, loses in three datasets, and ties in four datasets. Overall, the experimental results show that *ddr-ChowLiu* slightly outperforms *ChowLiu*. Therefore, if we use dependence distribution directly, instead of using dependence, it will result in a better classifier. This experiment also provides evidence that it is dependence distribution that affects classification, not dependence merely.

6. Conclusions

In this paper, we proposed a new explanation of the classification performance of Naïve Bayes. We showed that, essentially, dependence distribution, i.e. how the local dependence of an attribute distributes in each class, evenly or unevenly, and how the local dependences of all attributes work together, consistently (support a certain classification) or inconsistently (cancel each other out), play a crucial role in classification. This explains why, even with strong dependences, Naïve Bayes still works well; i.e. when those dependences cancel each other out, there is no influence on classification. In this case, Naïve Bayes is still the optimal classifier. In addition, we investigated the optimality of Naïve Bayes under the Gaussian distribution, and presented the explicit sufficient condition under which Naïve Bayes is globally optimal, even though the conditional independence assumption is violated.

We extended the *ChowLiu* algorithm by using dependence distribution to construct TAN, instead of using mutual information that only reflects the dependences among attributes merely. The extended algorithm outperforms the *ChowLiu* algorithm. This provides empirical evidence to support our explanation.

Ideally, a simple, sufficient and necessary condition for the optimality of Naïve Bayes is desirable. Our work is just a beginning toward this goal. Another interesting direction for future work is how to incorporate dependence distribution into the traditional dependence-based learning algorithms for Bayesian networks. As shown in the paper, to study a classifier, it is more reasonable to consider the influence of dependences (dependence distribution) on classification than it is to consider merely dependences.

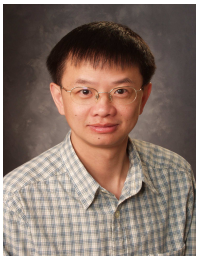
Acknowledgment

We thank the reviewers of FLAIR04 for their precious suggestions.

References

1. P. N. Bennett, Assessing the calibration of Naïve Bayes' posterior estimates, Technical Report No. CMU-CS00-155 (2000).
2. C. K. Chow and C. N. Liu, Approximating discrete probability distributions with dependence trees, *IEEE Trans. Inform. Th.* **14** (1968) 462–467.
3. W. S. Cooper, Some inconsistencies and misidentified modelling assumptions in probabilistic information retrieval, *ACM Trans. Inform. Sci.* **13**(1) (1995) 100–111.
4. P. Domingos and M. Pazzani, Beyond independence: conditions for the optimality of the simple Bayesian classifier, *Mach. Learn.* **29** (1997) 103–130.
5. U. Fayyad and K. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in *Proc. Thirteenth Int. Joint Conf. Artificial Intelligence*, (Morgan Kaufmann, 1993), pp. 1022–1027.
6. E. Frank, L. Trigg, G. Holmes and I. H. Witten, Naïve Bayes for regression, *Mach. Learn.* **41**(1) (2000) 5–15.
7. J. Friedman, On bias, variance, 0/1-loss, and the curse of dimensionality, *Data Min. Knowl. Disc.* **1**(1) (1997) 55–77.
8. N. Friedman, D. Geiger and M. Goldszmidt, Bayesian network classifiers, *Mach. Learn.* **29** (1997) 131–163.
9. A. Garg and D. Roth, Understanding probabilistic classifiers, in *Proc. 12th Eur. Conf. Machine Learning* (Springer, 2001), pp. 179–191.
10. D. J. Hand and Y. Yu, Idiots Bayes — not so stupid after all? *Int. Statist. Rev.* **69** (2001) 385–389.
11. I. Kononenko, Comparison of inductive and Naïve Bayesian learning approaches to automatic knowledge acquisition, *Current Trends in Knowledge Acquisition* (IOS Press, 1990).
12. P. Langley, W. Ibam and K. Thomas, An analysis of Bayesian classifiers, in *Proc. Tenth National Conf. Artificial Intelligence* (AAAI Press, 1992), pp. 223–228.
13. C. Merz, P. Murphy and D. Aha, UCI repository of machine learning databases, Department of ICS, University of California, Irvine (1997), http://www.ics.uci.edu/mlearn/MLRepository.html.
14. F. Monti and G. F. Cooper, A Bayesian network classifier that combines a finite mixture model and a Naïve Bayes model, in *Proc. 15th Conf. Uncertainty in Artificial Intelligence* (Morgan Kaufmann, 1999), pp. 447–456.
15. M. Pazzani, Search for dependencies in Bayesian classifiers, *Learning from Data: Artificial Intelligence and Statistics* (Springer Verlag, 1996).

16. J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann, San Mateo, CA, 1993).
17. J. R. Rachlin, S. Kasif and D. W. Aha, Toward a better understanding of memory-based reasoning systems, in *Proc. Eleventh Int. Machine Learning Conference* (Morgan Kaufmann, 1994), pp. 242–250.
18. D. Roth, Learning in natural language, in *Proc. IJCAI'99* (Morgan Kaufmannn, 1999), pp. 898–904.
19. H. Zhang and C. X. Ling, Learnability of augmented Naïve Bayes in nominal domains, in *Proc. Eighteenth Int. Conf. Machine Learning* (Morgan Kaufmann, 2001), pp. 617–623.



Harry Zhang is an assistant professor of computer sciences at the University of New Brunswick, Canada. He received his Ph.D. from the University of Western Ontario, Canada.

His current research interests include machine learning and data mining.