Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2014, Article ID 470821, 11 pages
http://dx.doi.org/10.1155/2014/470821

Research Article

Extracting Credible Dependencies for Averaged One-Dependence Estimator Analysis

LiMin Wang,[1,2] ShuangCheng Wang,[3] XiongFei Li,[1] and BaoRong Chi[4]

[1] *Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China*
[2] *State Key Laboratory of Computer Science, Beijing 100080, China*
[3] *School of Mathematics and Information, Shanghai Lixin University of Commerce, Shanghai 210620, China*
[4] *Medical College, Jilin University, Changchun 130021, China*

Correspondence should be addressed to XiongFei Li; lxf@jlu.edu.cn and BaoRong Chi; chibr@jlu.edu.cn

Received 8 April 2014; Revised 25 May 2014; Accepted 26 May 2014; Published 17 June 2014

Academic Editor: Yang Xu

Of the numerous proposals to improve the accuracy of naive Bayes (NB) by weakening the conditional independence assumption, averaged one-dependence estimator (AODE) demonstrates remarkable zero-one loss performance. However, indiscriminate superparent attributes will bring both considerable computational cost and negative effect on classification accuracy. In this paper, to extract the most credible dependencies we present a new type of seminaive Bayesian operation, which selects superparent attributes by building maximum weighted spanning tree and removes highly correlated children attributes by functional dependency and canonical cover analysis. Our extensive experimental comparison on UCI data sets shows that this operation efficiently identifies possible superparent attributes at training time and eliminates redundant children attributes at classification time.

1. Introduction

Bayesian networks (BNs) are a key research area of knowledge discovery and machine learning. A BN consists of two parts: a qualitative part and a quantitative part. The qualitative part denotes the graphical structure of the network, while the quantitative part consists of the conditional probability tables (CPTs) in the network. Although BNs are considered efficient inference algorithms, the quantitative part is considered a complex component and learning an optimal BN structure from existing data has been proven to be an NP-hard problem. The graphical structure of naive Bayes (NB) is simple and definite because of the conditional independence assumption between attributes, making NB efficient and effective [1, 2]. However, violations of this conditional independence assumption can make the classification of NB suboptimal. Numerous algorithms have been proposed to retain the desirable simplicity and efficiency of NB while alleviating the problems of the independence assumption. Averaged one-dependence estimator (AODE) [3, 4] utilizes a restricted class of one-dependence estimators (ODEs) and aggregates the predictions of all qualified estimators within this class. A superparent attribute is indiscriminately selected from an attribute set as the parent of all the other attributes in each ODE. By averaging the estimates of all of the three-dimensional estimators, AODE makes a weaker conditional independence assumption than NB. Previous studies that compared different variations of NB techniques prove that AODE is significantly better than other NB techniques in terms of zero-one loss reduction [5]. Since its introduction in 2005, AODE has enjoyed considerable popularity because of its capability to improve the accuracy of NB [5].

Another strategy to remedy violations of the attribute independence assumption is to eliminate highly correlated attributes. Backward sequential elimination (BSE) [6] uses a simple heuristic wrapper approach that selects a subset of the available attributes to minimize zero-one loss on the training set. BSE is effective especially for data sets with highly correlated attributes. Forward sequential selection (FSS) [7] uses the reverse search direction to BSE. However, both FSS

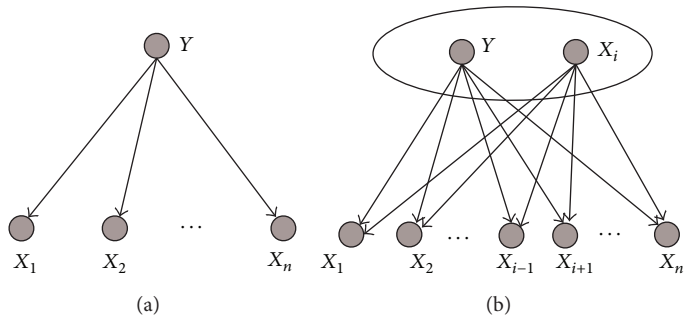


Figure 1: The network structure of NB and AODE.

and BSE have high computational overheads, especially on learning algorithms with high classification time complexity, because they apply the algorithms repeatedly until no accuracy improvement occurs. Subsumption resolution (SR) [8] identifies pairs of attribute values, such that one appears to subsume (be a generalization of) the other and deletes the generalization. Near-subsumption resolution (NSR) [8] is a variant of SR; it extends SR by deleting not only generalizations but also near-generalizations. For different instances, SR and NSR may find different attributes to remove, making them much more flexible than BSE and FSS. Since generalization mainly deals with pairs of attribute, there is no solution for more complicated situation and loop relationship. In this paper, we present a new type of seminaive Bayesian operation, which selects parent (SP) attributes by building maximum weighted spanning tree and removes children (RC) attributes by functional dependency and canonical cover analysis. Thus this algorithm has the advantages of BSE, FSS, and SR.

The remainder of the paper is organized as follows. Section 2 introduces the basic ideas of NB, AODE, and related background theory. Section 3 introduces the SP and RC techniques for attribute selection and elimination with AODE and presents the theoretical justification. Section 4 shows the experimental results on UCI data sets and a detailed analysis of different attribute selection techniques. The final section concludes the paper.

2. Related Research Work

2.1. NB and AODE. The aim of supervised learning is to predict from a training set the class of a testing instance $x = \{x_1, \ldots, x_n\}$, where x_i is the value of the ith attribute. We estimate the conditional probability of $P(y \mid x)$ by selecting $\arg\max_y P(y \mid x)$, where $y \in \{c_1, \ldots, c_k\}$ are the k classes. From Bayes theorem, we have

$$P(yx) = \frac{P(y,x)}{P(x)} \propto P(y,x) \propto P(x \mid y)\,P(y). \quad (1)$$

NB simplified the estimation of $P(x|y)$ by conditional independence assumption

$$P(xy) = \prod_{i=1}^{n} P(x_i \mid y). \quad (2)$$

Then, the following equation is often calculated in practice rather than (1):

$$P(yx) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y). \quad (3)$$

The corresponding network structure is depicted in Figure 1(a). One advantage of NB is avoiding model selection because selecting between alternative models can be expected to increase variance and allow a learning system to overfit the training data [3]. In consequence, changes in the training data will not lead to any change in NB, which leads in turn to lower variance [4]. By contrast, the underlying conditional probability tables will change correspondingly for those approaches (e.g., NB) with a definite model form when the training data changes, resulting in relatively gradual changes in the pattern of classification.

Numerous techniques have sought to enhance the accuracy of NB by relaxing the conditional independence assumption while attaining the efficiency and efficacy of one-dependence classifiers. Among them, averaged one-dependence estimator (AODE) [3, 4] utilizes a restricted class of one-dependence estimators (ODEs) and aggregates the predictions of all qualified estimators within this class. A superparent attribute (e.g., x_i) is selected as the parent of all the other attributes in each ODE, since

$$\begin{aligned} P(x,y) &= P(x_1, \ldots, x_i, \ldots, x_n, y) \\ &= P(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n, x_i, y) \\ &= P(x_i, y)\,P(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n \mid x_i, y). \end{aligned} \quad (4)$$

Independence is assumed among the remaining attributes given x_i and y. Consider

$$P(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n x_i, y) = \prod_{j=1, j\neq i}^{n} P(x_j \mid x_i, y). \quad (5)$$

Hence, x can be classified by selecting

$$\arg\max_y P(x,y) = \arg\max_y P(x_i, y) \prod_{j=1, j\neq i}^{n} P(x_j \mid x_i, y). \quad (6)$$

The corresponding network structure of AODE is depicted in Figure 1(b). AODE maintains the robustness and much of the efficiency of NB and at the same time exhibits significantly higher classification accuracy for many data sets. Therefore, it has the potential to be a valuable substitute for NB over a considerable range of classification tasks.

2.2. Related Background Theory. In the following discussion, Greek letters ($\alpha, \beta, \gamma, \ldots$) are used to denote sets of attributes. Lowercase letters represent the specific values used by corresponding attributes (e.g., x_i represents $X_i = x_i$). $P(\cdot)$ denotes the probability and $\widehat{P}(\cdot)$ denotes the probability estimation of $P(\cdot)$. Given a relation R (in a relational database), attribute Y of R is functionally dependent on attribute X of R and X of R functionally determines Y of R (in symbols $X \rightarrow Y$). Armstrong (1974) proposed in [9] a set of axioms (or, more precisely, inference rules) to infer all the functional dependencies (FDs) on a relational database, which represent the expert knowledge of the organizational data and their interrelationships. The axioms mainly include the following rules.

(i) Augmentation rule: if $\alpha \rightarrow \beta$ holds and γ is a set of attributes, then $\alpha\gamma \rightarrow \beta\gamma$.

(ii) Transitivity rule: if $\alpha \rightarrow \beta$ holds and $\beta \rightarrow \gamma$ holds, then $\alpha \rightarrow \gamma$.

(iii) Union rule: if $\alpha \rightarrow \beta$ holds and $\alpha \rightarrow \gamma$ holds, then $\alpha \rightarrow \beta\gamma$.

(iv) Decomposition rule: if $\alpha \rightarrow \beta\gamma$ holds, then $\alpha \rightarrow \beta$, $\alpha \rightarrow \gamma$.

(v) Pseudotransitivity rule: if $\alpha \rightarrow \beta$ holds and $\gamma\beta \rightarrow \delta$ holds, then $\alpha\gamma \rightarrow \delta$ holds.

Based on the aforementioned rules, we use the FD rules of probability in [10, 11] to link FD and probability theory. The following rules are included in the FD-probability theory link.

(i) Representation equivalence rule of probability: suppose data set S consists of two attribute sets $\{\alpha, \beta\}$ and β can be inferred by α; that is, the FD $\alpha \rightarrow \beta$ holds; then the following joint probability distribution holds:

$$P(\alpha) = P(\alpha, \beta). \tag{7}$$

(ii) Augmentation rule of probability: if $\alpha \rightarrow \beta$ holds and γ is a set of attributes, then the following joint probability distribution holds:

$$P(\alpha, \gamma) = P(\alpha, \beta, \gamma). \tag{8}$$

(iii) Transitivity rule of probability: if $\alpha \rightarrow \beta$ and $\beta \rightarrow \gamma$ hold, then the following joint probability distribution holds:

$$P(\alpha) = P(\alpha, \gamma). \tag{9}$$

(iv) Pseudotransitivity rule of probability: if $\alpha \rightarrow \beta$ and $\gamma\beta \rightarrow \delta$ hold, then the following joint probability distribution holds:

$$P(\alpha, \gamma) = P(\alpha, \gamma, \delta). \tag{10}$$

In the 1940s, Claude E. Shannon introduced information theory, the theoretical foundation of modern digital communication. Although Shannon was principally concerned with the problem of electronic communications, the theory has much broader applicability. Two commonly used definitions of information theory are described as follows.

Definition 1. Mutual information $I(Y; X)$ measures the information quantity that is transferred between attributes X and Y:

$$I(Y; X) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log_2 \frac{P(x, y)}{P(x) P(y)}, \tag{11}$$

where $P(x, y)$ is the joint probability distribution function of X and Y, and $P(x)$ and $P(y)$ are the marginal probability distribution functions of X and Y, respectively. High mutual information indicates a great relationship between X and Y; and zero mutual information between two random variables means they are independent.

Definition 2. Conditional mutual information (CMI) $I(X_1; X_2 \mid Y)$ measures the dependence between each pair of attributes $\{X_1, X_2\}$ given Y, which is shown as follows:

$$\begin{aligned} I(X_1; X_2 \mid Y) &= \sum_{y \in Y} P(y) \sum_{x_1 \in X_1} \sum_{x_2 \in X_2} P(x_1, x_2 \mid y) \\ &\quad \times \log_2 \frac{P(x_1, x_2 \mid y)}{P(x_1 \mid y) P(x_2 \mid y)}. \end{aligned} \tag{12}$$

Theorem 3. *Given instance $x = \{x_1, \ldots, x_n\}$ and class label y, if there exists FD $x_2 \rightarrow x_1$, then x_1 is extraneous for classification. That means $P(y \mid x) = P(y \mid x \rightharpoonup x_1)$, where "$\rightharpoonup$" represents the set difference.*

Proof. By applying the augmentation rule and the decomposition rule, from $x_2 \rightarrow x_1$ we can obtain

$$\begin{aligned} \{x \rightharpoonup x_1\} &\longrightarrow x_1, \\ \{x \rightharpoonup x_1, y\} &\longrightarrow x_1. \end{aligned} \tag{13}$$

By applying the representation equivalence rule of probability and the augmentation rule of probability, we can obtain

$$\begin{aligned} P(x) &= P(x \rightharpoonup x_1), \\ P(x, y) &= P(x \rightharpoonup x_1, y). \end{aligned} \tag{14}$$

Then,

$$\begin{aligned} P(y \mid x) &= \frac{P(x, y)}{P(x)} = \frac{P(x \rightharpoonup x_1, y)}{P(x \rightharpoonup x_1)} \\ &= P(y \mid x \rightharpoonup x_1). \end{aligned} \tag{15}$$

We can also prove Theorem 3 from the viewpoint of information theory, since

$$\begin{aligned} I(Y;X) &= -\sum_{x\in X}\sum_{y\in Y}\log_2 P(x,y)\,\frac{P(x,y)}{P(x)\,P(y)} \\ &= -\sum_{x\rightharpoonup x_1\in X\rightharpoonup X_1}\sum_{y\in Y}\log_2 P(x\rightharpoonup x_1,y) \\ &\qquad\times\frac{P(x\rightharpoonup x_1,y)}{P(x\rightharpoonup x_1)\,P(y)} \\ &= I(Y;X\rightharpoonup X_1). \end{aligned} \tag{16}$$

End of the proof. □

As for AODE, (4) will turn to be

$$\begin{aligned} &P(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n\mid x_i,y) \\ &\quad= P(x_2,\ldots,x_{i-1},x_{i+1},\ldots,x_n\mid x_i,y) \\ &\quad= \prod_{j=2,j\neq i}^{n} P(x_j\mid x_i,y). \end{aligned} \tag{17}$$

Thus, if FD $x_2 \rightarrow x_1$ is neglected, the contribution of x_1 to classification will be calculated repeatedly for each ODE and the classification result may be wrong.

3. Attribute Selection and Elimination

AODE makes a weaker attribute conditional independence assumption than that of NB. It selects one attribute as superparent in turn for each ODE submodel, and the other attributes are supposed to be conditionally independent. Previous studies have demonstrated that AODE has a considerably lower bias than that of NB with moderate increases in variance and time complexity [5]. The same attribute may play different roles (either parent or child) in different ODE submodels. In the following discussion, we will repair harmful interdependencies from two viewpoints: (1) select parent (p) attributes (SP) by building maximum weighted spanning tree; (2) remove children (c) attributes (RC) by functional dependency analysis.

3.1. How to Select Parent Attributes. SP selects the branch nodes as the p attributes from maximum weighted spanning tree (MST), the learning procedure of which can be summarized as follows.

(1) Use CMI to measure the weights of edges between each pair of attributes. Sort the edges into descending order by CMI. Let T be the set of edges comprising the MST. Set $T = \{\Phi\}$.

(2) Find the edge with the greatest weight and add this edge to T if and only if it does not form a cycle in T. If no remaining edges exist, exit and report MST to be disconnected.

(3) If T has $n-1$ edges (where n is the number of vertices in MST) stop and output T. Otherwise go to step (2).

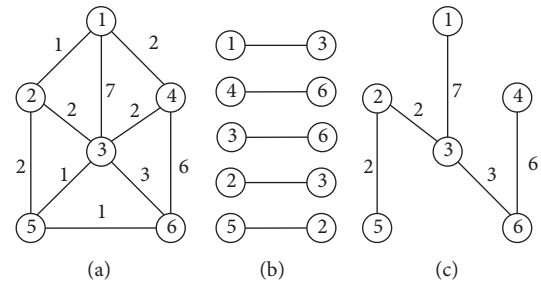


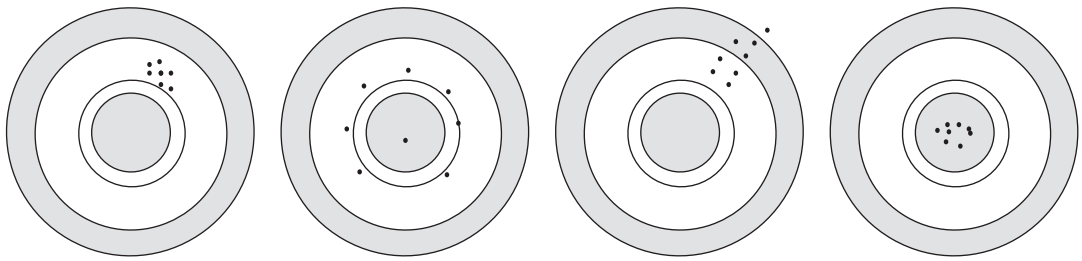
Figure 2: Example of building MST.

The p attributes selected must satisfy the criterion that they either appear as the branch nodes in MST or as the leaf nodes but with stronger relationship with other attributes. Figures 2(a), 2(b), and 2(c) show the original spanning tree, procedure of selecting edges, and final MST, respectively. As shown in Figure 2(c), attributes ②, ③, and ⑥ are branch nodes and can be used as p attributes. In addition, ①, ④, and ⑤ are leaf nodes with corresponding CMIs of 7, 6, and 2, respectively. The CMIs are then sorted into descending order. In this paper, if the sum of CMIs of the first k leaf nodes is greater than 85% of the sum of CMIs of all leaf nodes, we suppose that they represent the most important marginal relationships and can also be selected as p attributes. For example, since

$$\frac{7+6}{7+6+2} = 86.7\% > 85\%, \tag{18}$$

then ① and ④ can be used as p attributes. This criterion helps to ensure that strong, and only strong, relationships among attributes will be retained. By contrast, AODE [4] indiscriminately uses each attribute as superparent even if some attributes may be independent of others. Besides, SP supports incremental learning because it may reselect the subset of attributes when a new training instance becomes available.

At training time SP needs only to form the tables of joint attribute value, class frequencies to estimate the probabilities $\hat{P}(y)$, $\hat{P}(y,x_i)$, and $\hat{P}(y,x_i,x_j)$, which are required for estimating $\hat{P}(x_i \mid y)$, $\hat{P}(x_i,x_j \mid y)$, and $\hat{P}(x_j \mid y,x_i)$ in turn. Calculating the estimates requires a simple scan through the data, an operation of time complexity $O(tn^2)$, where t is the number of training instances and n is the number of attributes. To build maximum weighted spanning tree, SP must first calculate the CMI, requiring consideration for each pair of attributes, every pairwise combination of their respective values in conjunction with each class value. The time complexity to build a MST is $O(n^2\log n)$. The resulting time complexity is $O(tn^2 + k(nv)^2 + n^2\log n)$ and space complexity is $O(k(nv)^2)$, where k is the number of classes and v is the average number of values for an attribute. At classification time SP needs only to store the probability tables, space complexity $O(k(nv)^2)$. This compression over the table required at training time is achieved by storing probability estimates for each attribute value conditioned by



(a) High bias, low variance (b) Low bias, high variance (c) High bias, high variance (d) Low bias, low variance

Figure 3: Bias and variance in shooting arrows at a target.

the parent selected for that attribute and the class. The time complexity of classifying a single instance is $O(kn^2)$.

3.2. How to Eliminate Children Attributes. Kohavi and Wolpert [12] presented a bias-variance decomposition of expected misclassification rate, which is a powerful tool from sampling theory statistics for analyzing supervised learning scenarios. Suppose y and $\hat{y}$ are the true class label and that class generated by a learning algorithm, respectively; the zero-one loss function is defined as

$$\xi(y, \hat{y}) = 1 - \delta(y, \hat{y}), \tag{19}$$

where $\delta(y, \hat{y}) = 1$ if $\hat{y} = y$ and 0 otherwise. The bias term measures the squared difference between the average output of the target and the algorithm. This term is defined as follows:

$$\text{bias} = \frac{1}{2} \sum_{\hat{y}, y \in Y} \left[P(\hat{y} \mid x) - P(y \mid x)\right]^2, \tag{20}$$

where x is the combination of any attribute value. The variance term is a real valued nonnegative quantity and equals zero for an algorithm that always makes the same guess regardless of the training set. The variance increases as the algorithm becomes more sensitive to changes in the training set. It is defined as follows:

$$\text{variance} = \frac{1}{2}\left[1 - \sum_{\hat{y} \in Y} P(\hat{y} \mid x)^2\right]. \tag{21}$$

Moore and McCabe [13] illustrated bias and variance through shooting arrows at a target, as described in Figure 3. The perfect model can be regarded as the bull's eye on a target and the learned classifier as an arrow fired at the bull's eye. Bias and variance describe what happens when an archer fires many arrows at the target. Bias means that the aim is off and the arrows land consistently off the bull's eye in the same direction. Variance means that the arrows are scattered. Large variance means that repeated shots are widely scattered on the target. They do not give similar results but differ widely among themselves.

It is reported that removing redundant children attributes from within ODEs can help to decrease both bias and zero-one loss [3, 14]. Subsumption resolution (SR) [8] identifies pairs of attribute values such that one can replace the other.

Table 1: A loop relationship example.

X_1	X_2	X_3	X_4	Y
a_1	b_1	c_1	d_1	y_1
a_2	b_2	c_1	d_2	y_2
a_1	b_1	c_2	d_1	y_2
a_2	b_1	c_2	d_3	y_1
a_2	b_2	c_1	d_3	y_2

Deleting x_j from a Bayesian classifier should not be harmful when x_j is a generalization of x_i; that is, $P(x_j \mid x_i) = 1.0$. Only the attribute value x_i is necessary for classification; that is, $P(y \mid x_1, \ldots, x_n) = P(y|x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n)$. Such deletion may improve a classifier's estimate if it makes unwarranted assumptions about the relationship of x_j to the other attributes when estimating intermediate probability values, such as NB's independence assumption. Since $P(x_j|x_i) = 1.0$ can be represented as FD: $x_i \rightarrow x_j$, SR and FD have the same meaning but from different viewpoints.

SR mainly considers pair relationship of one-one. However, four basic relationships exist in the real world: one-one, one-many, many-many, and many-one. These four relationships can be grouped into two sets: one-one and many-one. Thus, SR cannot resolve interdependencies when a loop appears in the many-one relationship. The data presented in Table 1 shows a loop example with four attributes $\{X_1, X_2, X_3, X_4\}$ and class label Y. For the first instance $\{X_1 = a_1, X_2 = b_1, X_3 = c_1, X_4 = d_1\}$, suppose $\{X_2 = b_1\}$, $\{X_4 = d_1\}$, and $\{X_1 = a_1\}$ are generalizations of $\{X_1 = a_1\}$, $\{X_2 = b_1, X_3 = c_1\}$, and $\{X_4 = d_1\}$, respectively. The loop relationship is described in Figure 4(a), where "$\rightarrow$" represents the one-one relationship and "$\dashrightarrow$" represents the many-one relationship. After SR, only attributes X_3 are used for classification and NB will misclassify the first instance as "$Y = y_2$", even though it occurs in the training data. For different testing instances, different correlated attributes will be deleted. These instances will be illustrated from the viewpoint of FD. $\{X_1 = a_1, X_2 = b_1, X_3 = c_1, X_4 = d_1\}$ can be replaced by three FDs:

(1) $\{X_1 = a_1\} \rightarrow \{X_2 = b_1\}$,

(2) $\{X_2 = b_1, X_3 = c_1\} \rightarrow \{X_4 = d_1\}$,

(3) $\{X_4 = d_1\} \rightarrow \{X_1 = a_1\}$.

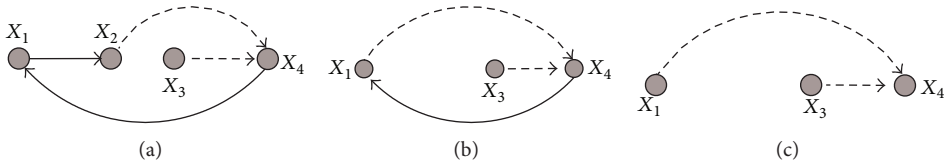


Figure 4: Loop relationship between attributes of training instance.

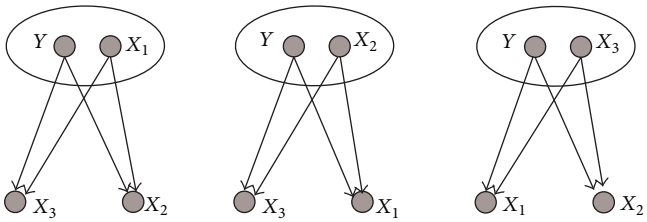


Figure 5: Original network structures of AODE submodels.

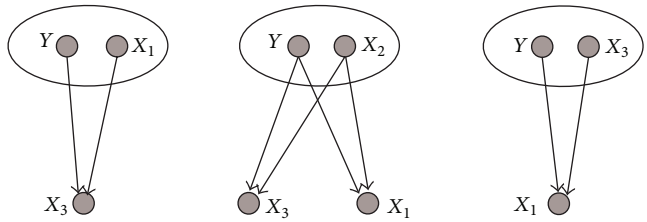


Figure 6: Network structures corresponding to instance-1.

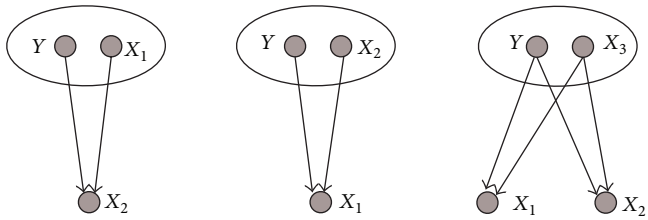


Figure 7: Network structures corresponding to instance-2.

The following results can be generated. We can obtain $\{X_1 = a_1\} \rightarrow \{X_1 = a_1, X_2 = b_1\}$ from FD (1) by applying union rule. As shown in Figure 4(b), X_2 disappears and the arc that once connects X_2 and X_4 now extends to connect X_1 and X_4.

We can obtain $\{X_1 = a_1, X_3 = c_1\} \rightarrow \{X_1 = a_1, X_2 = b_1, X_3 = c_1, X_4 = d_1\}$ from FD (2) by applying augmentation rule. As Figure 4(c) shows, the arc from X_4 to X_1 is removed to avoid a loop relationship. Thus, we can infer and obtain the other two attribute values from two attribute values of $\{X_1 = a_1, X_3 = c_1\}$. Correspondingly the first instance will be correctly classified.

It should be noted that SP selects p attributes from the probabilistic viewpoint by calculating CMI while RC selects c attributes from the logical viewpoint by inferring FDs from the training data. That is, the learning procedure of SP + RC is divided into two parts; SP roughly describes the basic structure of each submodel which uses p attributes as superparents and other attributes as the children; and for different instances, RC further refines the model by deleting redundant children attributes and thus makes the final model much more flexible and robust. Suppose an extreme instance, the CMIs of all attributes are all small and equal, and then all the attributes will be selected as p attributes. The structure will be just the same as AODE after applying SP. But for different testing instances, different FDs can help to make each submodel express the key dependencies. For example, suppose $\mathrm{FD}_1 = \{x_1 \rightarrow x_2\}$ holds for instance-1 and $\mathrm{FD}_2 = \{x_2 \rightarrow x_3\}$ holds for instance-2; Figures 5, 6, and 7 show the original AODE structure after applying SP and corresponding structures for instance-1 and instance-2, respectively.

Discovering FD from existing databases is an important issue. This issue has long been investigated and has been recently addressed with a data mining viewpoint in a novel and efficient way. Rather than exhibiting the set of all functional dependencies which hold in a relation, related work aims to discover a smaller cover equivalent to this set. This problem is known as FD inference. Association rules can be used to discover the relationships and potential associations of items or attributes among huge data. These rules can be effective in uncovering unknown relationships, thereby providing results that can be the basis of forecast and decision. They have proven to be useful tools for an enterprise as they strive to improve their competitiveness and profitability.

4. Experimental Study

We expect AODE with SP and RC to exhibit low zero-one loss and low bias. Thus, we compare the performance of the system with the following attribute selection methods. First is p attribute addition (PAA), which starts with p and c initialized to the empty and full sets, respectively. It adds one p attribute to each ODE at each step. Second is c attribute addition (CAA), which begins with p and c initialized to the full and empty sets, respectively. It adds one c attribute to every ODE at each step. Third is p attribute elimination (PAE), which starts with p and c initialized to the full set and deletes one p attribute from every ODE at each step. Fourth is c attribute elimination (CAE), which deletes one c attribute from every ODE at each step. Fifth and sixth are SR and NSR, respectively.

Table 2 summarizes the characteristics of each data set, including the numbers of instances, attributes, and classes. Missing values for qualitative attributes are replaced with modes and those for quantitative attributes are replaced

Table 2: Data sets.

Number	Data set	Number of instances	Attribute	Class
1	Abalone	4177	8	3
2	Adult	48842	14	2
3	Anneal	898	38	6
4	Audio	226	69	24
5	Car	1728	7	4
6	Chess	551	39	2
7	Cleveland	303	13	2
8	Connect 4	67557	42	3
9	Contact lenses	24	4	3
10	Donation	5749132	11	2
11	Heart	303	13	2
12	Hepatitis	155	19	2
13	Hungarian	294	13	2
14	Hypothyroid	3163	25	2
15	Iris	150	4	3
16	kr versus kp	3196	36	2
17	Labor	57	16	2
18	Mushroom	8124	22	2
19	Nursery	12960	8	5
20	Optdigits	5620	64	10
21	Satellite	6435	37	6
22	Shuttle	58000	9	7

with means from the training data. We estimate the base probabilities $P(y)$, $P(y, x_i)$, and $P(y, x_i, x_j)$ using the Laplace estimate as follows [15]:

$$\widehat{P}(y) = \frac{F(y) + 1}{K + k},$$
$$\widehat{P}(y, x_i) = \frac{F(y, x_i) + 1}{K_i + k_i}, \quad (22)$$
$$\widehat{P}(y, x_i, x_j) = \frac{F(y, x_i, x_j) + 1}{K_{ij} + k_{ij}},$$

where $F(\cdot)$ is the frequency with which a combination of terms appears in the training data, K is the number of training instances for which the class value is known, K_i is the number of training instances for which both the class and attribute X_i are known, and K_{ij} is the number of training instances for which all of the class and attributes X_i and X_j are known. k is the number of attribute values of class Y, k_i is the number of attribute value combinations of Y and X_i, and k_{ij} is the number of attribute value combinations of Y, X_j, and X_i. As NB and AODE require discrete valued data, all data were discretized using minimum description length (MDL) discretization [16]. Classifier is formed from data set and bias, variance, and zero-one loss estimated from the performance of those classifiers on the same data set. Experiments are performed on a dual-processor 3.1 GHz Windows XP computer with 3.16 Gb RAM. All algorithms are applied to the 22 data sets described in Table 2.

4.1. Zero-One Loss, Bias, and Variance Results. Table 3 presents for each data set the zero-one loss, which is estimated by 50 runs of twofold cross-validation to give an accurate estimation of the average performance of an algorithm. The advantage of this technique is that it uses the full training data as the training set and the testing set. Moreover, every case in the training data is used for the same number of times in each of the roles of training and testing data. Tables 4 and 5 provide the bias and variance results, respectively. The zero-one loss, bias, or variance across multiple data sets provides a gross measure of relative performance.

The basic relationships among attributes can be clearly observed by building MST. If one attribute is connected with several other attributes, the attribute is supposed to have crossed functional zones and will be selected as p attribute to retain complementarity. If one attribute is connected with only one other attribute, its independence characteristics may be obvious and will be reconsidered by the weight of CMI. Besides, RC helps to detect the situation in which the relationships that hold in MST need to be refined. Table 3 shows that the advantage of SP + RC is significant compared with SR and NSR in terms of zero-one loss. However, SR and NSR have a significant advantage over CAA, PAA, CAE, and PAE. The disappointing performances of CAA, PAA, CAE, and PAE can be ascribed to their susceptibility to getting trapped into poor selections by local minima during the first several additions or deletions.

The records in Table 4 show that all the c attribute selection algorithms applying SR, NSR, or FDs have a significant advantage in bias over CAE and PAE. In addition, CAE and PAE outperform CAA and PAA. However, comparing SP + RC with SP again does not show obvious difference. This result indicates that SP takes the main role for classification and its effect differs greatly in different data sets. The same result can also be inferred by comparing SP + RC with SR. The training sets containing only 25% of each data set for bias-variance evaluation are small because the data sets are primarily small. The bias of SP + RC decreases as training set size increases because more data will lead to more accurate probability distribution estimates and hence to more appropriate p attribute selection. Of these algorithms, RC, SR, and NSR have the weakest sensitivity to the changes in training data because they can utilize the testing set to infer rules for c attribute elimination. By contrast PAA, PAE, CAA, and CAE perform model selection and their biases differ greatly with different training data.

With respect to variance, as Table 5 shows, the variance of SP + RC does not show obvious advantage to other algorithms. Low-variance algorithms tend to enjoy an advantage with small data sets, whereas low-bias algorithms tend to enjoy an advantage with large data sets. Cross data set experimental studies of the traditional form presented above also support this hypothesis. The main reason may be that the relationship inferred based on MST may overfit the training data because SP needs to calculate CMI to construct MST. This requires enough instances to achieve precise probability estimation.

In the following discussion, canonical cover analysis [17], which can use limited number (e.g., 100) of instances to

Table 3: Experimental results of zero-one loss.

Data set	PAE	CAE	PAA	CAA	SR	NSR	SP	SP + RC
Abalone	0.512	0.528	0.491	0.519	0.489	0.493	0.396	0.402
Adult	0.212	0.187	0.182	0.158	0.149	0.152	0.137	0.131
Anneal	0.017	0.014	0.013	0.013	0.012	0.011	0.017	0.008
Audio	0.272	0.263	0.247	0.258	0.258	0.217	0.261	0.158
Car	0.236	0.222	0.117	0.212	0.132	0.249	0.127	0.082
Chess	0.147	0.135	0.121	0.129	0.099	0.142	0.128	0.102
Cleveland	0.259	0.262	0.251	0.252	0.247	0.196	0.247	0.178
Connect 4	0.277	0.268	0.289	0.271	0.275	0.267	0.264	0.238
Contact lenses	0.422	0.415	0.409	0.411	0.395	0.445	0.401	0.375
Donation	0.002	0.002	0.002	0.001	0.000	0.001	0.000	0.000
Heart	0.192	0.187	0.182	0.189	0.191	0.197	0.189	0.174
Hepatitis	0.216	0.213	0.209	0.211	0.198	0.208	0.187	0.163
Hungarian	0.193	0.192	0.187	0.190	0.194	0.196	0.217	0.163
Hypothyroid	0.031	0.026	0.021	0.029	0.024	0.021	0.037	0.012
Iris	0.118	0.128	0.125	0.121	0.113	0.102	0.077	0.087
kr versus kp	0.070	0.071	0.091	0.075	0.068	0.067	0.061	0.072
Labor	0.062	0.051	0.047	0.060	0.059	0.061	0.057	0.053
Mushroom	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Nursery	0.088	0.091	0.085	0.081	0.085	0.083	0.089	0.072
Optdigits	0.051	0.055	0.053	0.052	0.051	0.049	0.049	0.030
Satellite	0.122	0.119	0.117	0.132	0.112	0.117	0.116	0.114
Shuttle	0.001	0.001	0.001	0.000	0.001	0.001	0.001	0.000

Table 4: Experimental results of bias.

Data set	PAE	CAE	PAA	CAA	SR	NSR	SP	SP + RC
Abalone	0.332	0.351	0.327	0.312	0.312	0.307	0.291	0.316
Adult	0.147	0.151	0.148	0.152	0.118	0.120	0.155	0.161
Anneal	0.235	0.187	0.236	0.224	0.211	0.200	0.188	0.179
Audio	0.181	0.201	0.183	0.191	0.184	0.127	0.207	0.212
Car	0.152	0.164	0.155	0.148	0.134	0.138	0.127	0.131
Chess	0.106	0.110	0.092	0.114	0.104	0.112	0.113	0.112
Cleveland	0.071	0.068	0.072	0.069	0.064	0.048	0.052	0.049
Connect 4	0.201	0.165	0.199	0.215	0.192	0.197	0.202	0.204
Contact lenses	0.132	0.129	0.133	0.142	0.154	0.112	0.127	0.126
Donation	0.207	0.219	0.211	0.212	0.214	0.206	0.183	0.179
Heart	0.133	0.131	0.135	0.140	0.159	0.162	0.130	0.132
Hepatitis	0.112	0.120	0.114	0.121	0.113	0.109	0.115	0.113
Hungarian	0.205	0.211	0.206	0.198	0.203	0.211	0.204	0.219
Hypothyroid	0.013	0.013	0.012	0.015	0.009	0.007	0.008	0.008
Iris	0.090	0.088	0.090	0.101	0.094	0.067	0.087	0.085
kr versus kp	0.067	0.070	0.071	0.085	0.069	0.075	0.061	0.058
Labor	0.102	0.099	0.098	0.101	0.104	0.107	0.103	0.099
Mushroom	0.002	0.002	0.002	0.003	0.001	0.002	0.001	0.001
Nursery	0.050	0.049	0.050	0.048	0.052	0.054	0.051	0.049
Optdigits	0.028	0.029	0.028	0.031	0.029	0.028	0.027	0.025
Satellite	0.077	0.075	0.074	0.078	0.080	0.078	0.081	0.078
Shuttle	0.003	0.004	0.003	0.005	0.002	0.003	0.002	0.002

Table 5: Experimental results of variance.

Data set	PAE	CAE	PAA	CAA	SR	NSR	SP	SP + RC
Abalone	0.127	0.134	0.142	0.152	0.165	0.161	0.151	0.153
Adult	0.045	0.043	0.046	0.043	0.044	0.042	0.040	0.048
Anneal	0.137	0.146	0.139	0.144	0.152	0.155	0.142	0.145
Audio	0.103	0.101	0.092	0.085	0.082	0.078	0.091	0.087
Car	0.101	0.108	0.118	0.122	0.132	0.133	0.112	0.115
Chess	0.108	0.112	0.103	0.089	0.093	0.088	0.102	0.096
Cleveland	0.152	0.162	0.172	0.179	0.192	0.187	0.207	0.198
Connect 4	0.110	0.105	0.108	0.112	0.110	0.107	0.098	0.092
Contact lenses	0.202	0.213	0.192	0.182	0.178	0.180	0.175	0.172
Donation	0.048	0.054	0.051	0.052	0.064	0.061	0.069	0.062
Heart	0.128	0.133	0.137	0.128	0.142	0.138	0.156	0.152
Hepatitis	0.055	0.058	0.059	0.061	0.063	0.060	0.058	0.053
Hungarian	0.207	0.217	0.196	0.180	0.183	0.177	0.195	0.191
Hypothyroid	0.004	0.004	0.005	0.005	0.004	0.006	0.004	0.004
Iris	0.092	0.096	0.090	0.094	0.092	0.088	0.085	0.081
kr versus kp	0.032	0.032	0.031	0.031	0.023	0.021	0.020	0.020
Labor	0.037	0.034	0.036	0.037	0.038	0.036	0.027	0.024
Mushroom	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Nursery	0.025	0.026	0.027	0.028	0.029	0.031	0.037	0.032
Optdigits	0.021	0.020	0.021	0.022	0.023	0.023	0.022	0.019
Satellite	0.045	0.047	0.049	0.050	0.049	0.045	0.051	0.048
Shuttle	0.003	0.003	0.004	0.003	0.002	0.002	0.002	0.002

Table 6: Variance results on 5 large data sets normalized with respect to AODE.

Data set	PAE	CAE	PAA	CAA	SR	NSR	SP	SP + RC
Adult	0.8653	0.8269	0.8846	0.8269	0.8461	0.8076	0.7692	0.7115
Connect 4	0.9091	0.8677	0.8925	0.9256	0.9078	0.8842	0.8099	0.7603
Donation	0.6575	0.7397	0.6986	0.7123	0.8767	0.8356	0.9452	0.7393
Nursery	0.6756	0.7027	0.7297	0.7567	0.7837	0.8378	0.9875	0.5945
Shuttle	0.7142	0.7142	0.9523	0.7142	0.4761	0.4761	0.4761	0.4761

infer credible FD set, is applied in tandem with functional dependency analysis. Let F_c be a canonical cover for a set of simple FDs and the procedure of computing F_c is described as in Algorithm 1.

The chosen canonical cover is the set of minimal dependencies. Such a cover is information lossless and is considerably smaller than the set of all valid dependencies. These qualities are particularly important to decrease variance while zero-one loss and bias are not affected negatively because of the provision of relevant knowledge wherein redundancy is minimized and extraneous information is discarded. Five large data sets with number of instances >10,000 are selected for variance comparison. The experimental results are shown in Table 6, and we can see that canonical cover analysis can help to project AODE to be competitive with other attribute selection strategies from the viewpoint of variance comparison.

4.2. Elimination Ratio. Statistically a win/draw/loss record (W/D/L) is calculated for each pair of competitors A and B with regard to a performance measure M. The record represents the number of data sets in which A, respectively, beats, loses to, or ties with B on M. Small improvements in leave-one-out error may be attributable to chance. Consequently, it may be beneficial to use a statistical test to assess whether an improvement is significant. A standard binomial sign test, assuming that wins and losses are equiprobable, is applied to these records. A difference is considered to be significant when the outcome of a two-tailed binomial sign test is less than 0.05. To observe the effect of p attribute elimination of SP on zero-one loss, we used the following criterion:

$$E^p_{\text{Ratio}} = \frac{E_p}{n}, \tag{23}$$

where E_p is the number of p attributes eliminated and n is the number of attributes. Table 7 shows the comparison results. The E^p_{Ratio} of SP is much higher than that of the other two p attribute elimination algorithms, PAA and PAE. The statistical records in Table 3 show that the corresponding zero-one loss of SP is lower generally. We suggest that

F_c = FDs
repeat
  Use the union rule to replace any dependencies in F_c of the form $\alpha_1 \rightarrow \beta_1$ and $\alpha_1 \rightarrow \beta_2$ with $\alpha_1 \rightarrow \beta_1\beta_2$.
  Find a functional dependency $\alpha \rightarrow \beta$ in F_c with an extraneous attribute either in α or in β.
  If an extraneous attribute is found, delete it from $\alpha \rightarrow \beta$ in F_c.
until (F_c does not change)

Algorithm 1

Table 7: Win/draw/loss comparison of elimination ratio of p attribute.

W/D/L	PAE	PAA
PAA	8/10/4	
SP	9/9/4	6/11/5

Table 8: Win/draw/loss comparison of elimination ratio of c attribute.

W/D/L	CAE	CAA	SR	NSR
CAA	6/12/4			
SR	9/9/4	6/12/4		
NSR	9/8/5	6/11/5	6/11/5	
RC	11/7/4	11/9/2	10/8/4	9/9/4

the reason for SP's outstanding performance on zero-one loss reduction is that it greatly utilizes the probabilistic dependency relationship on training data.

For example, the elimination ratio of SP is as high as 61.5% for data set "anneal." However, the corresponding zero-one loss is lower than that of PAA and PAE. The main reason may be that, for as many as 38 attributes and only 894 instances, some attributes may have cross-functional zones, and only a few attributes may play the decisive role. After calculating and comparing the sum of CMI between one attribute and all the other attributes, most of the eliminated attributes have weak relationships to other attributes or are even nearly independent of them.

SP selects p attributes based on MST; attributes with a strong relationship among them will be selected first. If any attribute is removed by mistake, the classification results will not be affected greatly. However, for different training sets, especially when their sizes are very small, the conditional distribution estimates may differ greatly and different structures of MST may be obtained. Different p attributes will be selected for classification. The number of FDs extracted from RC is less than that from SR because numerous p attributes are eliminated during SP, especially for data set "audio," with fewer p attributes and much more complicated FDs to remove c attributes. In the W/D/L records, the advantage in zero-one loss is significant with respect to SP versus PAE or SP versus PAA, but not SP + RC versus SP. This result shows that the advantage of SP + RC is from SP but not from RC.

With the increasing number of attributes, more RAM is needed to store joint probability distributions. An important restriction of our algorithm is that the number of the left side of FD should be no more than 2. To observe the effect of SP + RC and SR on each data set, we calculate the c attribute elimination ratios by the following criterion:

$$E^c_{\text{Ratio}} = \frac{\sum_{i=1}^{N} E^i_c}{N}, \tag{24}$$

where E^i_c is the number of c attributes eliminated for the ith instance, and N is the size of data set. Table 8 shows the comparison results of E^c_{Ratio} of SP + RC with the other three c attribute elimination algorithms, CAA, CAE, and SR. Table 3 shows that SP + RC has a significant advantage in zero-one loss over SR and NSR while SR and NSR outperform CAA and CAE. Comparing Table 3 with Table 8 reveals that both RC and SR can help to decrease zero-one loss. However, the effectiveness of RC relies greatly on SP while SR can always improve the performance of AODE. If SP removes p attributes by mistake, some valuable FDs will not be extracted by RC. However, if just redundant p nodes are eliminated, RC can extract more reliable FDs than SR because RC has considered all possible situations of SR.

For example, on data set "hypothyroid" the E^c_{Ratio} of RC is 32%, which indicates that RC just uses approximately 30% of all attributes as c attribute. The reason for this high ratio is that SP has eliminated 21% of the attributes from the data set. For data set "anneal," the E^c_{Ratio} is also as high as 34%, and the zero-one loss is much higher than that of the other three algorithms. This result means SP has removed some p attributes by mistake; RC cannot extract FDs that are dependent on those deleted p attributes. Hence, the experimental results of SP + RC can still be improved if we can find other methods to keep more valuable p attributes for classification.

5. Conclusion and Future Work

AODE provides an attractive framework by averaging all models from a restricted class of one-dependence classifiers. The class of all such classifiers that have all other attributes depends on a common attribute and the class attribute. The current work aims to improve the accuracy derived by MST and FD from weakening the attribute independence assumption without high computational overheads.

Overall, this study developed a classification learning technique that retains the simplicity and direct theoretical foundation of AODE while reducing computational overhead without incurring a negative effect on classification performance. The p attribute of AODE can also be considered the parent of the class attribute. Therefore, we hypothesize that the success of AODE and its variations may be attributed to the fact that AODE not only aggregates all other restricted classes of models but also extends NB to handle the parent of class attribute. If this hypothesis can be proven, we can design a novel and perhaps more effective Bayesian classifier than AODE by constructing the Markov blanket of class attributes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Science Foundation of China (Grants nos. 61272209 and 61300145) and Postdoctoral Science Foundation of China (Grants nos. 20100481053 and 2013M530980).

References

[1] D. Dash and G. F. Cooper, "Exact model averaging with naive Bayesian classifiers," in *Proceedings of the 19th International Conference on Machine Learning*, pp. 91–98, Sydney, Australia, July 2002.

[2] E. Frank, M. Hall, and B. Pfahringer, "Locally weighted naive Bayes," in *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence*, pp. 249–256, Acapulco, NM, USA, August 2003.

[3] F. Zheng and G. I. Webb, "Finding the right family: parent and child selection for averaged one dependence estimators," in *Proceedings of the 18th European Conference on Machine Learning*, pp. 490–501, Warsaw, Poland, September 2007.

[4] F. Zheng and G. I. Webb, "Efficient lazy elimination for averaged one-dependence estimators," in *Proceedings of the 23rd International Conference on Machine Learning*, pp. 1113–1120, Pittsburgh, Pa, USA, June 2006.

[5] A. Z. Nayyar, C. Jesus, and G. I. Webb, "Alleviating naive bayes attribute independence assumption by attribute weighting," *The Journal of Maching Learning Research*, vol. 14, no. 6, pp. 1113–1120, 2006.

[6] P. Langley and S. Sage, "Induction of selective Bayesian classifiers," in *Proceedings of the 10th Conference on Uncertainty in Artificial Intelligence*, pp. 399–406, Seattle, Wash, USA, July 1994.

[7] M. J. Pazzani, "Constructive induction of Cartesian product attributes," in *Proceedings of the Information, Statistics and Induction in Science Conference*, pp. 66–77, July 1996.

[8] F. Zheng, G. I. Webb, P. Suraweera, and L. Zhu, "Subsumption resolution: an efficient and effective technique for semi-naive Bayesian learning," *Machine Learning*, vol. 87, no. 1, pp. 93–125, 2012.

[9] W. W. Armstrong, "Dependency structures of data base relationships," in *Proceedings of the IFIP Congress*, pp. 580–583, 1974.

[10] L. M. Wang, G. F. Yao, and X. Li, "Extracting logical rules and attribute subset from confidence domain," *Information*, vol. 15, no. 1, pp. 173–180, 2012.

[11] L. M. Wang and G. F. Yao, "Bayesian network inference based on functional dependency mining of relational database," *Information*, vol. 15, no. 6, pp. 2441–2446, 2012.

[12] R. Kohavi and D. Wolpert, "Bias plus variance decomposition for zero-one loss functions," in *Proceedings of the 13th European Conference on Machine Learning*, pp. 275–283, June 1996.

[13] D. S. Moore and G. P. McCabe, *Introduction to the Practice of Statistics*, Michelle Julet, San Francisco, Calif, USA, 4th edition, 2002.

[14] A. Z. Nayyar and G. I. Webb, "Fast and effective single pass Bayesian learning," in *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, vol. 7818 of *Lecture Notes in Computer Science*, pp. 149–160, Gold Coast, Australia, April 2013.

[15] B. Cestnik, "Estimating probabilities: a crucial task in machine learning," in *Proceedings of the 9th European Conference on Artificial Intelligence*, pp. 147–149, Pitman, Boston, Mass, USA, August 1990.

[16] U. M. Fayyad and K. B. Irani, "Multi-interval discretization of continuous-valued attributes for classification learning," in *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pp. 1022–1029, August 1993.

[17] L. M. Wang and G. F. Yao, "Learning NT Bayesian classifier based on canonical cover analysis of relational database," *Information*, vol. 15, no. 1, pp. 165–172, 2012.