Entropy **2015**, *17*, 4134-4154; doi:10.3390/e17064134

OPEN ACCESS
entropy
ISSN 1099-4300
www.mdpi.com/journal/entropy

Article

General and Local: Averaged k-Dependence Bayesian Classifiers

Limin Wang [1,*], Haoyu Zhao [2], Minghui Sun [1] and Yue Ning [1]

[1] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, ChangChun 130012, China; E-Mails: smh@jlu.edu.cn (M.S.); ningyue@jlu.edu.cn (Y.N.)
[2] School of Software, Jilin University, ChangChun 130012, China; E-Mail: zhaohw@jlu.edu.cn

* Author to whom correspondence should be addressed; E-Mail: wanglim@jlu.edu.cn; Tel.: +86-431-85626892.

Academic Editors: Carlos Alberto de Bragança Pereira and Adriano Polpo

Received: 4 May 2015 / Accepted: 9 June 2015 / Published: 16 June 2015

Abstract: The inference of a general Bayesian network has been shown to be an NP-hard problem, even for approximate solutions. Although k-dependence Bayesian (KDB) classifier can construct at arbitrary points (values of k) along the attribute dependence spectrum, it cannot identify the changes of interdependencies when attributes take different values. Local KDB, which learns in the framework of KDB, is proposed in this study to describe the local dependencies implicated in each test instance. Based on the analysis of functional dependencies, substitution-elimination resolution, a new type of semi-naive Bayesian operation, is proposed to substitute or eliminate generalization to achieve accurate estimation of conditional probability distribution while reducing computational complexity. The final classifier, averaged k-dependence Bayesian (AKDB) classifiers, will average the output of KDB and local KDB. Experimental results on the repository of machine learning databases from the University of California Irvine (UCI) showed that AKDB has significant advantages in zero-one loss and bias relative to naive Bayes (NB), tree augmented naive Bayes (TAN), Averaged one-dependence estimators (AODE), and KDB. Moreover, KDB and local KDB show mutually complementary characteristics with respect to variance.

Keywords: k-dependence Bayesian classifier; substitution-elimination resolution; functionaldependency rules of probability

---

1. Introduction

Bayesian networks (BNs), which were introduced by Pearl [1], can encode dependencies among all variables. Their success has led to a recent flurry of algorithms for learning BNs from data [2–5]. $BN =< N, A, \Theta >$ is a directed acyclic graph with a conditional probability distribution for each node, collectively represented by Θ that quantifies how much a node depends on its parents. Each node $n \in N$ represents a domain variable, and each arc $a \in A$ between nodes represents a probabilistic dependency. A BN can be used as a classifier that characterizes the joint distribution $P(\mathbf{x}, y)$ (In the following discussion, lower-case letters denote specific values taken by corresponding attributes. For instance, x_i represents the event that $X_i = x_i$) of class variable Y and a set of attributes $\mathbf{X} = \{X_1, X_2, \cdots, X_n\}$, and predicts the class label with the highest conditional probability. Denoting the parent nodes of x_i by $Pa(x_i)$, the joint distribution $P_B(\mathbf{x}, y)$ can be represented by factors over the network structure B, as follows:

$$P_{B(\mathbf{x},y)} = \prod_{i=1}^{n} P(x_i|Pa(x_i)). \tag{1}$$

The inference of a general BN has been shown to be an NP-hard problem [6] even for approximate solutions [7]. However, learning unrestricted BNs does not necessarily lead to a classifier with good performance. For example, naive Bayes (NB) [8] is the simplest BN, which considers only the dependence between each attribute X_i and the class variable Y. However, Friedman *et al.* [9] observed that unrestricted BN classifiers do not outperform NB in a large sample of benchmark data sets. Many BN classifiers have been proposed to overcome the limitation of NB. One practical approach for structure learning is to impose some restrictions on the structures of BNs, for example, learning tree-like structures. Sahami [10] proposed to describe the limited dependence among variables within a general framework, which is called k-dependence Bayesian (KDB) classifier. Friedman *et al.* [9] proposed tree augmented naive Bayes (TAN), a structure-learning algorithm that learns a maximum spanning tree from the attributes. Conditional mutual information is applied in these two algorithms to measure the weight of arcs between predictive attributes. When data size becomes larger, the superiority in high-dependence representation helps KDB obtain better classification performance than TAN.

The key differences between Bayesian classifiers are their structure-learning algorithms. Many criteria, such as Bayesian scoring function [11], minimal description length (MDL) [12] and Akaike Information Criterion (AIC) [13], have been proposed to find out one global graph structure B_G that best characterizes the true distribution of given data. Considering the time and space complexity overhead, only a limited number of conditional probabilities can be encoded in BN. All credible dependencies must be represented to obtain a more accurate estimation of the true joint distribution. However, these criteria can only approximately measure the overall interdependencies between attributes, but cannot identify the change of interdependencies when attributes take different values. Thus the candidate graph structures may have very close score values and are non-negligible in the posterior sense [14]. To extend the limited representation of B_G, some researchers proposed to aggregate several candidate BNs together. Averaged one-dependence estimators (AODE), which were proposed by Webb *et al.* [15], aggregate the predictions of all qualified restricted class of one-dependence estimators. Zheng *et al.* [16] proposed subsumption resolution (SR), to efficiently identify occurrences of the specialization-generalization relationship and eliminate generalizations at classification time. By introducing Functional Dependency (FD) analysis

into the learning procedures, the model interpretability and robustness of different Bayesian classifiers can be improved greatly. After eliminating highly dependent attribute values by applying FD analysis, the maximal spanning tree (MST) of TAN is rebuilt with the rest of the attribute values for each test instance. Correspondingly the extraneous effect caused by logical relationships between attribute values will be mitigated [17]. To evaluate the feasibility of integrating probabilistic reasoning and logical reasoning into the framework of AODE, we first select the branch nodes of MST as the super parents, then refine AODE by applying FD analysis to delete redundant children attribute [18].

In this paper, local mutual information and conditional local mutual information, which are deduced from classical information theory, are applied to build the local graph structure B_L. B_L can be considered a complementary part of B_G, to describe local causal relationships. To construct classifiers at arbitrary points (values of k) along the attribute dependence spectrum, both B_L and B_G are built in the framework of KDB model. Substitution-elimination resolution (SER), a new type of semi-naive Bayesian operation is proposed to substitute or eliminate generalization to achieve accurate estimation of conditional probability distribution while reducing computational complexity. SER deals only with specific values and only in the context of other specific values. We prove that this adjustment is theoretically correct and demonstrate experimentally that it can considerably improve zero-one loss, bias and variance.

The remainder of this paper is organized as follows: Section 2 first proposes the background theory—information theory and functional dependency rules of probability, and then clarifies the rationality of SER. Section 3 introduces the basic ideas of KDB, local KDB and the proposed algorithm, averaged k-dependence Bayesian classifiers (AKDB), which averages the output of KDB and local KDB. Section 4 compares various approaches on data sets from the UCI Machine Learning Repository. Finally, Section 5 presents possible future work.

2. Background Theory and Related Research Work

2.1. Information Theory

In the 1940s, Claude E. Shannon introduced information theory [19], the theoretical basis of modern digital communication. Although Shannon was principally concerned with the problem of electronic communications, the theory has a broader applicability. Many commonly used measures are based on the entropy of information theory and used in a variety of classification algorithms.

Definition 1. [19] *Entropy of an attribute (or random variable) is a function that attempts to characterize its unpredictability. When given a discrete random variable X with any possible value x and probability distribution function $P(\cdot)$, entropy is defined as follows,*

$$H(X) = -\sum_{x \in X} P(x) log_2 P(x) \tag{2}$$

Deterministic attributes have zero entropy as entropy measures the amount of uncertainty with which they take some values. Similar to the concept of conditional probability, conditional entropy $H(X|Y)$ may be understood as the amount of randomness in the random variable X when the value of Y is known.

Definition 2. [19] *Given discrete random variables X and Y and their possible values x and y, conditional entropy is defined as follows:*

$$H(X|Y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)log_2P(x|y) \tag{3}$$

Using the definition of entropy and conditional entropy, we can calculate the amount of information shared between two attributes. The stronger the correlation, the higher the value of mutual information will be.

Definition 3. [19] *The mutual information (MI) $I(X;Y)$ of two random variables is a measure of the mutual dependence of the variables and is defined as follows:*

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)log_2\frac{P(x,y)}{P(x)P(y)} \tag{4}$$

Mutual information $I(X;Y)$ between two attributes X and Y measures the expected reduction in entropy and is nonnegative, *i.e.*, $I(X;Y) \geq 0$. $I(X;Y) = 0$ if they are independent and is maximized if $H(X|Y) = 0$. Similar to the definition of conditional entropy, conditional mutual information $I(X;Y|Z)$ indicates the amount of information shared between two attributes X and Y when all the values of attribute Z are known.

Definition 4. [19] *Conditional mutual information (CMI) $I(X;Y|Z)$ is defined as follows:*

$$I(X;Y|Z) = \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} P(x,y,z)log_2\frac{P(x,y|z)}{P(x|z)P(y|z)} \tag{5}$$

Definition 5. *Local mutual information (LMI) $I(X;y)$ is defined to measure the reduction of entropy about variable X after observing that $Y = y$, as follows:*

$$I(X;y) = \sum_{x\in X} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \tag{6}$$

Definition 6. *Conditional local mutual information (CLMI) $I(x;y|Z)$ is defined to measure the amount of information shared between two attribute values x and y when all the values of attribute Z are known, as follows:*

$$I(x;y|Z) = \sum_{z\in Z} P(x,y,z)\log\frac{P(x,y|z)}{P(x|z)P(y|z)} \tag{7}$$

2.2. Functional Dependency Analysis and Substitution-Elimination Resolution

Given a data set D, attribute value y is functionally dependent on attribute value x, and x functionally determines y (in symbols $x \rightarrow y$). We demonstrated functional dependency rules of probability in [17,18] to build a linkage between probabilistic inference and logical inference, and the following rules are mainly included:

- Representation equivalence of probability: Suppose two attribute values $\{x,y\}$ and y can be inferred by x, *i.e.*, the FD $x \rightarrow y$ holds, then the following joint probability distribution holds:

$$P(x) = P(x,y) \tag{8}$$

- Augmentation rule of probability: If FD $x \rightarrow y$ holds and z is another attribute value, then the following joint probability distribution holds:

$$P(x, z) = P(x, y, z) \tag{9}$$

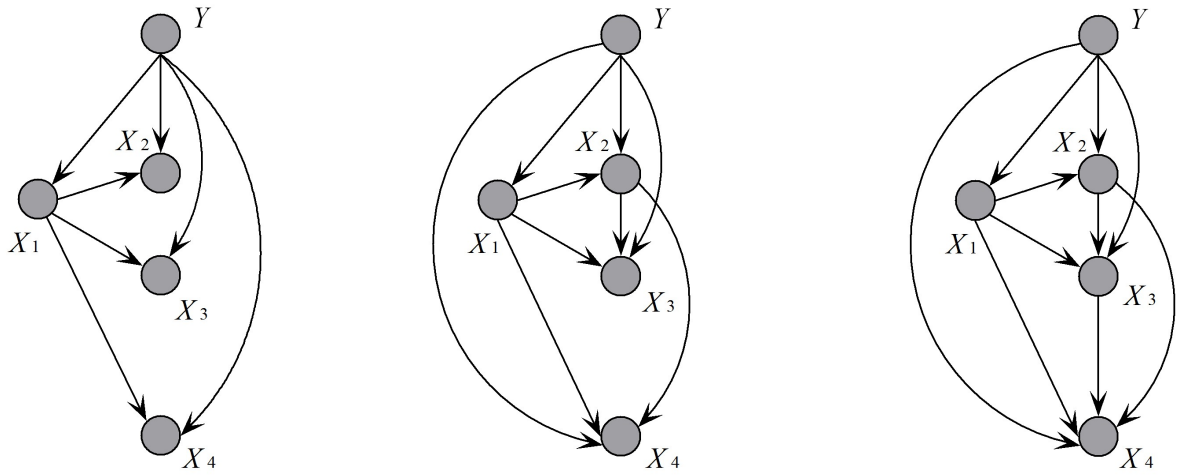
- Transitivity rule of probability: If FDs $x \rightarrow y$ and $y \rightarrow z$ hold, then the following joint probability distribution holds:

$$P(x) = P(x, z) \tag{10}$$

- Pseudo-transitivity rule of probability: If $yz \rightarrow \delta$ and $x \rightarrow y$ hold, then the following joint probability distribution holds:

$$P(x, z) = P(x, z, \delta) \tag{11}$$

Definition 7. *A k-dependence Bayesian classifier (k-DBC) is a BN that contains the structure of the naive Bayesian classifier and allows each attribute X_i to have a maximum of k attribute nodes as parents. Given attribute order $\{X_1, \cdots, X_n\}$, $Pa(X_i) = \{Y, X_{d_i}\}$ where X_{d_i} is a subset of $\{X_1, \cdots, X_{i-1}\}$, $|d_i| = min\{i-1, k\}$ and $Pa(y) = \varnothing$.*



(*a*) *k*=1 (*b*) *k*=2 (*c*) *k*=3

Figure 1. The k-dependence relationships between attributes inferred from KDB.

Learning the structure of a k-DBC actually means learning an order of variables and then adding arcs from a variable to all the variables ranked after it. In fact, given the order of variables, learning a k-DBC is relatively easier.

We consider a hypothetical example with four predictive attributes {*Pregnant*, *Gender*, *Familial Inheritance*, and *Breast Cancer*} and class variable {*Normal*}. When given different k values, the corresponding k-DBC models are shown in Figure 1, where $X = \{X_1, X_2, X_3, X_4\}$ and Y denote {*Pregnant*, *Gender*, *Familial Inheritance*, *Breast Cancer*} and class *Normal*, respectively.

Subsumption is a central concept in Inductive Logic Programming [20], where it is used to identify generalization-specialization relationships between clauses and to support the process of unifying clauses.

Definition 8. *(Generalization and specialization) For two attribute values x_i and x_j, if $P(x_j|x_i) = 1.0$ then x_j is a generalization of x_i and x_i is a specialization of x_j.*

Suppose that *Gender* has two values: *female* and *male*, and *Pregnant* has two values: *yes* and *no*. If *Pregnant* = *yes*, it follows that *Gender* = *female*. Therefore, *Gender* = *female* is a generalization of *Pregnant* = *yes*, *i.e.*, FD: $\{Gender = female\} \rightarrow \{Pregnant = yes\}$ holds.

Theorem 1. *Substitution resolution: Suppose that for k-DBC the attribute order is $\{X_1, X_2, \cdots, X_n\}$ and $X_i (i > k)$ should select k attributes as parents. If x_p is a generalization of x_q ($x_p, x_q \in Pa_i$), then for $\forall x_t \notin Pa_i (t < i)$, x_t as a substitute of x_p will help achieve a more accurate approximate estimation of probability distribution.*

Proof. For k-DBC, conditional probability $P(x_i|Pa_i, y)$ can be considered an approximate estimation of $P(x_i|x_1, \cdots, x_{i-1}, y)$. Evidently, $P(x_i|Pa_i, x_t, y)$ will be more accurate than $P(x_i|Pa_i, y)$, where $x_t \notin Pa_i$ and $t < i$. If x_p is a generalization of x_q ($x_p, x_q \in Pa_i$), by applying the augmentation rule of probability we will have $P(x_i|Pa_i, y) = P(x_i|Pa_i - x_p, y)$, where " $-$ " denotes set difference. To retain the k-dependence restriction, we need to select x_t as a substitute of x_p. □

The example presented in Figure 1 illustrates this relationship. The joint probability distribution of the full Bayesian classifier, as shown in Figure 1c, is expressed as follows:

$$P(y, \mathbf{x}) = P(y)P(x_1|y)P(x_2|y, x_1)P(x_3|y, x_1, x_2)P(x_4|y, x_1, x_2, x_3) \tag{12}$$

In addition, the joint probability distribution of 2-DBC, as Figure 1b shows, is as follows,

$$P(y, \mathbf{x}) = P(y)P(x_1|y)P(x_2|y, x_1)P(x_3|y, x_1, x_2)P(x_4|y, x_1, x_2) \tag{13}$$

By comparing Equation (12) and Equation (13) we can observe that, Equation (13) uses $P(x_4|y, x_1, x_2)$ to obtain an approximate estimation of $P(x_4|y, x_1, x_2, x_3)$. *Gender* = *female* is a generalization of *Pregnant* = *yes*. Thus, $P(Gender = female|Pregnant = yes) = 1$ or $P(x_2|x_1) = 1$. By applying the augmentation rule of probability, we derive the following equations:

$$\begin{aligned} P(x_4|y, x_1, x_2, x_3) &= \frac{P(x_4, y, x_1, x_2, x_3)}{P(y, x_1, x_2, x_3)} \\ &= \frac{P(x_4, y, x_1, x_3)}{P(y, x_1, x_3)} \\ &= P(x_4|y, x_1, x_3) \end{aligned} \tag{14}$$

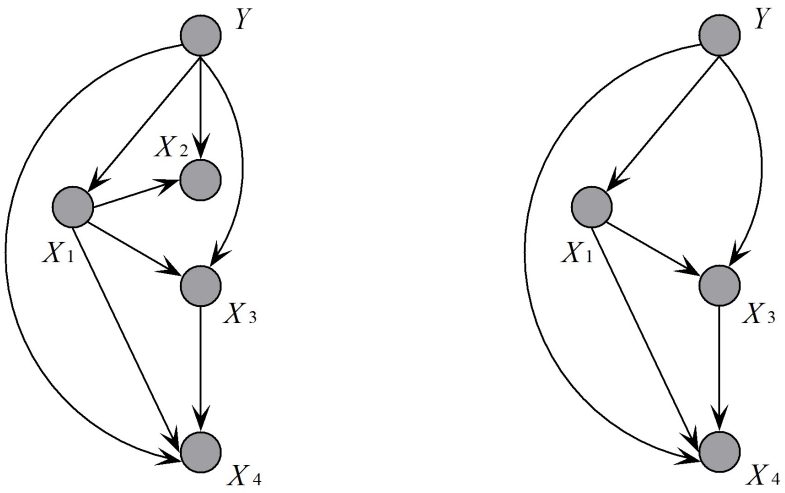
and

$$\begin{aligned} P(x_3|y, x_1, x_2) &= \frac{P(x_3, y, x_1, x_2)}{P(y, x_1, x_2)} \\ &= \frac{P(x_3, y, x_1)}{P(y, x_1)} \\ &= P(x_3|y, x_1) \end{aligned} \tag{15}$$

Equation (12) will change to be,

$$P(y, \mathbf{x}) = P(y)P(x_1|y)P(x_2|y, x_1)P(x_3|y, x_1)P(x_4|y, x_1, x_3) \tag{16}$$

Thus for Equation (13), if we use x_3 to substitute x_2 in Pa_4 and $\varnothing$ to substitute x_2 in Pa_3, corresponding 2-DBC as shown in Figure 2a is just the same as the full Bayesian classifier for the instances where $Pregnant = yes$ holds. Thus we can obtain a more accurate estimation of $P(y, \mathbf{x})$ and the *2*-dependence restriction still retained.



(a) After substitution resolution (b) After elimination resolution

Figure 2. The 2-dependence relationships between attributes after substitution-elimination resolution.

Theorem 2. *Elimination resolution: For* $\forall x_p \in Pa_i$*, if* x_i *is a generalization of* x_p*, then* $P(x_i|Pa_i) = 1.0$ *and the factor* $P(x_i|Pa_i)$ *will be eliminated from the joint probability distribution.*

For example, Equations (12) and (13) both use the factor $P(x_2|x_1, y)$. If x_2 is a generalization of x_1, then, by applying the augmentation rule of probability, we derive the following equation:

$$\begin{aligned} P(x_2|x_1, y) &= \frac{P(x_2, y, x_1)}{P(y, x_1)} \\ &= \frac{P(y, x_1)}{P(y, x_1)} = 1 \end{aligned} \tag{17}$$

Thus, Equation (16) can be rewritten as follows:

$$P(y, \mathbf{x}) = P(y)P(x_1|y)P(x_3|y, x_1)P(x_4|y, x_1, x_3) \tag{18}$$

The corresponding Bayesian structure is shown in Figure 2b. When two attributes are strongly related, the classifier may overweigh the inference from the two attributes, which results in prediction bias. FDs will help avoid this situation, and the high-dimensional representation or even entire classification model is simplified and improved.

3. KDB, Local KDB and AKDB

KDB allows us to construct classifiers at arbitrary points (values of k) along the feature dependence spectrum, while also capturing most of the computational efficiency of the naive Bayesian model. Thus KDB presents an alternative to the general trend in BN learning algorithms that conducts an expensive search through the space of network structures.

KDB is supplied with both a database of pre-classified instances, DB, and the k value for the maximum allowable degree of feature dependence. The KDB outputs a k-dependence Bayesian classifier with conditional probability tables determined from the input data. The algorithm is as follows:

Algorithm 1 KDB.

1. For each attribute X_i, compute *MI*, $I(X_i;Y)$, where Y is the class.
2. Compute class *CMI* $I(X_i;X_j|Y)$ for each pair of attributes X_i and X_j, where $i \neq j$.
3. Let the used variable list, *S*, be empty.
4. Let the Bayesian network being constructed, *BN*, begin with a single class node, *Y*.
5. Repeat until *S* includes all domain attributes
   5.1. Select feature X_{max} which is not in *S* and has the highest value $I(X_{max};Y)$.
   5.2. Add a node to *BN* representing X_{max}.
   5.3. Add an arc from *Y* to X_{max} in *BN*.
   5.4. Add $m = min(|S|,k)$ arcs from *m* distinct attributes X_j in S with the highest value for $I(X_{max};X_j|Y)$.
   5.5. Add X_{max} to S.
6. Compute the conditional probability tables inferred by the structure of *BN* by using counts from *DB*, and output *BN*.

From Definitions 3–6, we can obtain the following results:

$$\begin{cases} I(X_i;Y) = \sum_{X_i} I(x_i;Y) \\ I(X_i;X_j|Y) = \sum_{X_i}\sum_{X_j} I(x_i;x_j|Y) \end{cases} \tag{19}$$

MI and *CMI* are commonly applied to roughly measure the direct or conditional relationships between predictive attributes and class variable Y. However, in the real world, the relationships between attributes may differ significantly as the situation changes. For some instances attributes *A* and *B* are highly related. Meanwhile, for other instances, *A* is independent of *B*, but highly related to attribute Y. Considering the relationships among attributes *Gender*, *Pregnant* and *Breast Cancer*, *Gender = female* and *Breast Cancer=yes* are highly related. By contrast, if *Gender = female*, then we cannot make any definite conclusion about the value of *Pregnant*, nor about the value of *Gender* if *Breast Cancer= no*. Note that traditional Bayesian classifier, e.g., KDB, which is learned based on classical information theory, cannot describe such interdependencies. However, *LMI* and *CLMI* can be used to identify the dynamic changes, thus making the final model much more flexible.

As shown in Figure 3, for the first instance, the attribute value x_2 is independent of other attribute values and the local relationship between $\{x_1, x_3\}$ and class variable Y is just like a triangle. For the i_{th} instance, $\{x_2, x_3\}$ are independent of Y, and the local relationship between x_1 and Y is just like an

oval. For the last instance, x_3 is independent of other attribute values and the local relationship between $\{x_1, x_2\}$ and Y is just like a broken line. If all situations are considered together, then the overall relationship between attributes $\{X_1, X_2, X_3\}$ and class variable Y is just like a rectangle.

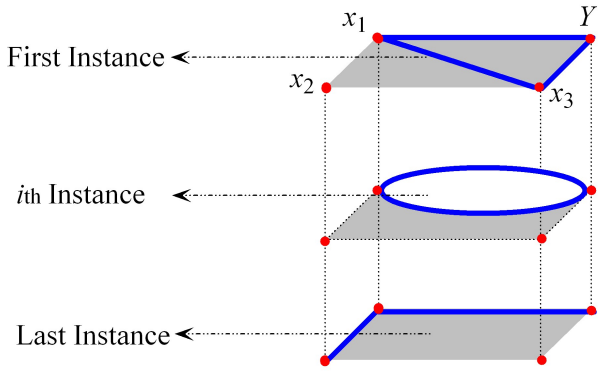


Figure 3. The basic and local relationships among $\{X_1, X_2, X_3\}$ and Y.

KDB learns the basic relationships of full BN. In the first two learning steps of KDB, if $I(Y; X)$ and $I(X_i; X_j|Y)$ are replaced by $I(Y; x)$ and $I(x_i; x_j|Y)$ respectively, then the local KDB that describes the local relationships of each test instance can be inferred. On the basis of this, FD analysis is introduced into the learning procedure to improve model robustness.

The learning procedure of the local KDB is described as follows:

Algorithm 2 Local KDB.

For each test instance x= $\{x_1, x_2, \cdots, x_n\}$

1. For each attribute value $x_i \in$**x**, compute *LMI*, $I(x_i; Y)$, where Y is the class.
2. Compute *CLMI* $I(x_i; x_j|Y)$ for each pair of attribute values x_i and x_j, where $i \neq j$ and $x_i, x_j \in$**x**.
3. Let the used variable list, *S*, be empty.
4. Let the Bayesian network being constructed, *BN*, begin with a single class node, *Y*.
5. Repeat until *S* includes all attribute values
   5.1. Select attribute value x_{max} which is not in *S* and has the highest value $I(x_{max}; Y)$.
   5.2. Add a node to *BN* representing x_{max}.
   5.3. Add an arc from *Y* to x_{max} in *BN*.
   5.4. Add $m = min(|S|, k)$ arcs from *m* distinct attribute values x_j in S with the highest value for $I(x_{max}; x_j|Y)$.
   5.5. Add x_{max} to S.
6. Apply SER to substitute generalization or eliminate redundant conditional probability.
7. Compute the conditional probability tables inferred by the structure of *BN* by using counts from *DB*, and output *BN*.

The final classifier, AKDB, estimates the class membership probabilities by averaging KDB and local KDB classifiers. The basic idea of AKDB can be explained from the perspective of medical diagnosis. KDB describes the basic relationships between different symptoms that can be explained

by domain knowledge learned from book or in school. Meanwhile, the local KDB describes the possible relationships between different symptoms of a specific patient. To make a definite diagnosis, rich experience (which corresponds to robust KDB model) and flexible mind (which corresponds to dynamic local KDB model) are both necessary and important.

FDs require a method for inferring from the training data whether one attribute value is a generalization of another. FDs use the following criterion:

$$|T_{x_i}| = |T_{x_i,x_j}| \geq l$$

to infer that x_j is a generalization of x_i, where $|T_{x_i}|$ is the number of training cases with value x_i, $|T_{x_i,x_j}|$ is the number of training cases with both values, and l is a user-specified minimum frequency. A large number of deterministic attributes, which are on the left side of the FD, will increase the risk of incorrect inference, and at the same time need more computer memory to store credible FDs. Consequently, only the one-one FDs are selected in our current work. Besides, as no formal method has been used to select an appropriate value for l, we use the same setting as that proposed by Webb [15], *i.e.*, $l = 100$, which is achieved from empirical studies.

The learning framework of local KDB is described as follows. During training time, FD analysis is applied to detect all possible specialization-generalization relationships. During classification time, local KDB first builds the basic network structure for each test instance t; then selects the specialization-generalization relationships that hold in t, and applies SER to refine the network structure. From the definitions of local mutual information and FD we can see that, they both deal with attribute values rather than attributes. In the real world the interdependencies may be varied when attributes take different values. As for some test instances attributes X_i and X_j are independent, for other test instances X_i may be dependent on X_j. Classical Bayesian classifiers, e.g., TAN and KDB, which build the network structure by computing mutual information and conditional mutual information, cannot resolve such situations. Whereas local KDB helps to remedy this limitation.

Another feature of our algorithm which makes it very suitable for data mining domains is its relatively small computational complexity. Computing the actual network structure of KDB requires $O(n^2mcv^2)$ time (dominated by Step 2) and that of Local KDB only requires $O(n^2mc)$ time, where n is the number of attributes, m is the number of training instances, c is the number of classes, and v is the average number of discrete values that an attribute may take. Moreover, classifying an instance both KDB and local KDB require $O(nck)$ time. Forming the additional two-dimensional probability estimate table SER requires $O(mn^2v^2)$ time. Classification of a single instance requires considering each pair of attributes to detect dependencies and is of time complexity $O(cn)$.

4. Experiments

4.1. Bias and Variance

The classification of each case in the test set is done by choosing, as class label, the value of the class variable that has the highest posterior probability. Classification accuracy is measured by the percentage of correct predictions on the test sets (*i.e.*, using a zero-one loss function). Kohavi and Wolpert presented a bias-variance decomposition of the expected misclassification rate [21], 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 generated by a learning algorithm, respectively, the zero-one loss function is defined as:

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

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

$$bias = \frac{1}{2} \sum_{\hat{y}, y \epsilon Y} [P(\hat{y}|\mathbf{x}) - P(y|\mathbf{x})]^2,$$

where $\mathbf{x}$ is the combination of any attribute value. The variance term is a real valued non-negative 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:

$$variance = \frac{1}{2}[1 - \sum_{\hat{y} \epsilon Y} P(\hat{y}|\mathbf{x})^2].$$

4.2. Statistical Results on UCI Data Sets

In order to verify the efficiency and effectiveness of the proposed AKDB, we conduct experiments on 41 data sets from the UCI machine learning repository. Table 1 summarizes the characteristics of each data set, including the number of instances, attributes and classes. Large data sets with an instance number greater than 3000 are annotated with the symbol "*". Missing values for qualitative attributes are replaced with modes, and those for quantitative attributes are replaced with means from the training data. For each benchmark data set, numeric attributes are discretized using Minimum Description Length discretization [22]. The following techniques are compared:

- NB, standard naive Bayes.
- TAN, tree-augmented naive Bayes.
- AODE, AODE with subsumption resolution.
- KDB, standard k-dependence Bayesian classifier.
- TAN-FDA [17], a variation of TAN that rebuilds MST for each testing instance.
- AODE-SR [18], a variation of AODE that selects super parent and delete extraneous children attributes.
- LKDB (Local KDB), a variation of KDB that describes the local dependencies among attributes.
- AKDB, a combination of KDB and local KDB.

All algorithms were coded in MATLAB 7.0 on a Pentium 2.93 GHz/2GB RAM computer. Base probability estimates $P(y)$ and $P(x_j|y)$ with Laplace correction are defined as follows,

$$\begin{cases} \hat{P}(y) = \dfrac{\sum_{i=1}^{N} \delta(y_i, y) + 1}{N + t} \\ \hat{P}(x_j|y) = \dfrac{\sum_{i=1}^{N} \delta(x_{ij}, x_j)\delta(y_i, y) + 1}{\delta(y_i, y) + t_j} \end{cases} \tag{20}$$

where N is the number of training instances, t is the number of classes, y_i is the class label of the i_{th} training instance, t_j is the number of values of the j_{th} attribute, x_{ij} is the j_{th} attribute value of the i_{th} training instance, x_j is the j_{th} attribute value of the test instance, and $\delta(\cdot)$ is a binary function, which is one if its two parameters are identical and zero otherwise. Thus, $\sum_{i=1}^{N}\delta(y_i,y)$ is the frequency that the class label y occurs in the training data and $\sum_{i=1}^{N}\delta(x_{ij},x_j)\delta(y_i,y)$ is the frequency that the class label y and the attribute value x_j occurs simultaneously in the training data.

Table 2 presents for each data set the average zero-one loss, which is estimated by 10-fold cross-validation to obtain an accurate estimation of the average performance of an algorithm. The average bias and variance results are shown in Tables 3 and 4, respectively, in which only 15 large data sets are selected because of statistical significance. The average zero-one loss, bias or variance across multiple data sets provides a gross measure of relative performance. Statistically, a win/draw/loss (W/D/L) record is calculated for each pair of competitors A and B with respect to performance measure M. The record represents the number of data sets in which A either beats, loses to or ties with B on M. We assess a difference as significant if the outcome of a one-tailed binomial sign test is less than 0.05. Tables 5, 6 and 7 show the W/D/L records that correspond to zero-one loss, bias and variance, respectively.

Allowing more dependencies for KDB reduces zero-one loss significantly more often than it increases. As more attributes are utilized for classification, the increase in the value of k will help ensure that more causal relationships will appear and be expressed in the joint probability distribution. By contrast, the AODE family, e.g., AODE and AODE-SR, which utilize a restricted class of one-dependence estimators (ODEs), aggregates the predictions of all qualified estimators within this class. The superiority of AODE family over single-structure classifiers, e.g., TAN, TAN-FDA and KDB, that are learned on the basis of classical information theory, can be attributed to the superiority of the aggregating mechanism. From Table 5 we observed that, AKDB, which is the combination of KDB and local KDB, enjoys a significant zero-one loss advantage relative to other algorithms. Moreover, we applied Friedman test (FT) [23,24], which is a non-parametric measure, to rank and compare the algorithms. FT helps to compare and evaluate the overall prediction performance of different learning algorithms when dealing with numerous data sets. The best performing algorithm getting the rank of 1, the second best rank 2, $\cdots$. In case of ties, average ranks are assigned. Let r_i^j be the rank of the j-th of k algorithms on the i-th of N data sets. FT compares the average ranks of algorithms, $R_j = \frac{1}{N}\sum_i r_i^j$. The experimental results of FT are shown in Table 8, from which we can see that, the order of these algorithms is {AKDB, AODE-SR, AODE,TAN-FDA, LKDB, KDB, TAN, NB} when comparing the experimental results on all data sets. Thus the effectiveness of AKDB is proved from the perspective of FT.

We need to further evaluate whether local KDB works as an effective complementary part of AKDB. The relative zero-one loss/bias/variance ratio $\varrho(\cdot)$ is proposed to measure the extent to which local KDB helps to improve the performance of KDB. A higher value of the ratio $\varrho(\cdot)$ corresponds to a smaller ratio of AKDB and KDB, which indicates the better performance of AKDB.

$$\begin{cases} \varrho(Z) & = 1 - \dfrac{zero-one\ loss\ of\ AKDB}{zero-one\ loss\ of\ KDB} \\ \varrho(B) & = 1 - \dfrac{bias\ of\ AKDB}{bias\ of\ KDB} \\ \varrho(V) & = 1 - \dfrac{variance\ of\ AKDB}{variance\ of\ KDB} \end{cases} \tag{21}$$

Table 1. Data sets.

No.	Data Set	# Instance	Attribute	Class
1	Abalone *	4,177	9	3
2	Adult *	48,842	15	2
3	Anneal	898	39	6
4	Audio	226	70	24
5	Balance Scale (Wisconsin)	625	5	3
6	Breast-cancer-w	699	10	2
7	Car	1,728	8	4
8	Chess	551	40	2
9	Connect-4 *	67,557	43	3
10	Contraceptive-mc	1,473	10	3
11	Credit	690	16	2
12	Cylinder-bands	540	40	2
13	Dermatology	366	35	6
14	Glass Identification	214	10	3
15	Heart Disease	303	14	2
16	Hepatitis	155	20	2
17	Hungarian	294	14	2
18	Iris	150	5	3
19	Kr-vs-kp *	3,196	36	2
20	Labor	57	17	2
21	LED	1,000	8	10
22	Letter-recog*	20,000	16	26
23	Localization*	164,860	5	11
24	Lung Cancer	32	57	3
25	Lymphography	148	19	4
26	Magic *	19,020	11	2
27	Mushroom *	8,124	22	2
28	Nursery *	12,960	8	5
29	Optdigits *	5,620	64	10
30	Poker-hand*	1,025,010 *	11	10
31	Primary Tumor	339	18	22
32	Satellite *	6,435	37	6
33	Segment	2,310	20	7
34	Shuttle *	58,000	9	7
35	Sick *	3,772	30	2
36	Spambase *	4,601	58	2
37	Teaching-ae	151	6	3
38	Vehicle	846	19	4
39	Vowel	990	13	11
40	Wine Recognition	178	14	3
41	Zoo	101	17	7

Table 2. Experimental results of 0–1 loss.

Dataset	NB	TAN	AODE	TAN-FDA	AODE-SR	KDB	LKDB	AKDB
Abalone	0.472	0.459	0.448	0.448	0.449	0.467	0.456	0.462
Adult	0.158	0.138	0.149	0.133	0.130	0.138	0.132	0.129
Anneal	0.038	0.009	0.009	0.009	0.008	0.009	0.013	0.009
Audio	0.239	0.292	0.204	0.325	0.284	0.323	0.331	0.302
Balance Scale	0.285	0.280	0.298	0.289	0.243	0.293	0.289	0.292
Breast-cancer-w	0.026	0.042	0.036	0.038	0.046	0.074	0.036	0.041
Car	0.140	0.057	0.082	0.037	0.037	0.038	0.061	0.028
Chess	0.113	0.093	0.100	0.082	0.075	0.100	0.111	0.079
Connect-4	0.278	0.235	0.242	0.215	0.209	0.228	0.241	0.227
Contraceptive-mc	0.504	0.489	0.494	0.485	0.484	0.500	0.482	0.481
Credit-a	0.141	0.151	0.139	0.161	0.149	0.146	0.149	0.143
Cylinder-bands	0.215	0.283	0.189	0.261	0.243	0.226	0.181	0.190
Dermatology	0.019	0.033	0.016	0.045	0.048	0.066	0.044	0.029
Glass Identification	0.262	0.220	0.252	0.215	0.160	0.220	0.199	0.201
Heart	0.178	0.193	0.170	0.220	0.192	0.211	0.196	0.190
Hepatitis	0.194	0.168	0.181	0.177	0.172	0.187	0.148	0.169
Hungarian	0.160	0.170	0.167	0.160	0.160	0.180	0.184	0.166
Iris	0.087	0.080	0.087	0.085	0.081	0.087	0.080	0.080
Kr-vs-kp	0.121	0.078	0.084	0.045	0.078	0.042	0.047	0.037
Labor	0.035	0.053	0.053	0.069	0.052	0.035	0.051	0.052
Led	0.267	0.266	0.268	0.257	0.263	0.262	0.264	0.265
Letter-recog	0.253	0.130	0.088	0.086	0.091	0.099	0.118	0.077
Localization	0.496	0.358	0.360	0.291	0.297	0.296	0.319	0.277
Lung-cancer	0.438	0.594	0.500	0.735	0.666	0.594	0.621	0.560
Lymphography	0.149	0.176	0.169	0.212	0.209	0.237	0.169	0.172
Magic	0.224	0.168	0.175	0.158	0.155	0.164	0.171	0.154
Mushrooms	0.020	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Nursery	0.097	0.065	0.073	0.028	0.041	0.029	0.065	0.039
Optdigits	0.077	0.041	0.031	0.034	0.029	0.037	0.049	0.031
Poker-hand	0.499	0.330	0.481	0.311	0.207	0.196	0.050	0.053
Primary-tumor	0.546	0.543	0.575	0.555	0.557	0.572	0.579	0.561
Satellite	0.181	0.121	0.115	0.112	0.109	0.108	0.139	0.102
Segment	0.079	0.039	0.034	0.040	0.043	0.047	0.033	0.033
Shuttle	0.004	0.002	0.001	0.001	0.001	0.001	0.001	0.001
Sick	0.031	0.026	0.027	0.024	0.023	0.022	0.031	0.022
Spambase	0.102	0.067	0.067	0.065	0.058	0.064	0.069	0.057
Teaching-ae	0.497	0.550	0.490	0.519	0.507	0.536	0.467	0.460
Vehicle	0.392	0.294	0.290	0.248	0.279	0.294	0.307	0.291
Vowel	0.424	0.130	0.150	0.212	0.126	0.182	0.247	0.132
Wine	0.497	0.034	0.023	0.033	0.028	0.023	0.048	0.022
Zoo	0.030	0.010	0.030	0.031	0.030	0.050	0.027	0.009

Table 3. Experimental results of bias.

Dataset	NB	TAN	AODE	TAN-FDA	AODE-SR	KDB	LKDB	AKDB
Abalone	0.403	0.320	0.330	0.321	0.321	0.328	0.311	0.312
Adult	0.140	0.107	0.113	0.113	0.104	0.107	0.108	0.101
Connect-4	0.232	0.182	0.192	0.170	0.250	0.178	0.177	0.181
Kr-vs-kp	0.111	0.066	0.069	0.035	0.071	0.038	0.040	0.032
Letter-recog	0.229	0.175	0.182	0.164	0.167	0.165	0.168	0.159
Localization	0.382	0.326	0.321	0.308	0.302	0.314	0.311	0.302
Magic	0.198	0.135	0.161	0.133	0.139	0.132	0.131	0.133
Mushrooms	0.039	0.000	0.000	0.001	0.002	0.000	0.000	0.000
Nursery	0.073	0.051	0.051	0.046	0.061	0.041	0.041	0.040
Optdigits	0.065	0.031	0.029	0.027	0.035	0.028	0.029	0.026
Poker-hand	0.326	0.226	0.263	0.276	0.254	0.331	0.290	0.244
Satellite	0.166	0.094	0.080	0.086	0.081	0.081	0.078	0.090
Shuttle	0.085	0.045	0.039	0.032	0.032	0.039	0.031	0.031
Sick	0.006	0.002	0.002	0.002	0.001	0.003	0.002	0.002
Spambase	0.096	0.065	0.067	0.058	0.068	0.051	0.045	0.051

Table 4. Experimental results of variance.

Dataset	NB	TAN	AODE	TAN-FDA	AODE-SR	KDB	LKDB	AKDB
Abalone	0.093	0.161	0.137	0.158	0.158	0.153	0.171	0.160
Adult	0.027	0.066	0.044	0.071	0.055	0.073	0.081	0.051
Connect-4	0.095	0.088	0.110	0.097	0.040	0.104	0.109	0.086
Kr-vs-kp	0.025	0.019	0.023	0.012	0.010	0.010	0.008	0.005
Letter-recog	0.142	0.165	0.142	0.167	0.173	0.174	0.177	0.160
Localization	0.191	0.256	0.219	0.254	0.258	0.271	0.281	0.249
Magic	0.041	0.079	0.061	0.076	0.031	0.082	0.080	0.072
Mushrooms	0.008	0.001	0.000	0.001	0.001	0.001	0.000	0.001
Nursery	0.027	0.038	0.030	0.039	0.043	0.045	0.052	0.034
Optdigits	0.025	0.028	0.024	0.030	0.025	0.032	0.032	0.029
Poker-hand	0.209	0.231	0.214	0.279	0.216	0.328	0.287	0.262
Satellite	0.021	0.041	0.049	0.046	0.051	0.053	0.049	0.040
Shuttle	0.004	0.001	0.002	0.002	0.000	0.002	0.001	0.001
Sick	0.006	0.007	0.005	0.007	0.005	0.006	0.006	0.007
Spambase	0.010	0.017	0.014	0.020	0.007	0.024	0.025	0.016

Table 5. W/D/L comparison results of 0–1 loss on all data sets.

W/D/L	NB	TAN	AODE	TAN-FDA	AODE-SR	KDB	LKDB
TAN	25/5/11						
AODE	24/12/5	14/13/14					
TAN-FDA	24/8/9	18/12/11	12/15/14				
AODE-SR	25/7/9	22/13/6	18/13/10	15/19/7			
KDB	22/10/9	15/13/13	14/12/15	9/20/12	7/11/23		
LKDB	24/6/11	12/14/15	10/15/16	14/11/16	11/11/19	13/12/16	
AKDB	27/7/7	21/18/2	20/14/7	23/14/4	15/19/7	23/15/3	22/17/2

Table 6. W/D/L comparison results of bias on large data sets.

W/D/L	NB	TAN	AODE	TAN-FDA	AODE-SR	KDB	LKDB
TAN	14/1/0						
AODE	14/1/0	2/10/3					
TAN-FDA	14/1/0	9/3/3	8/5/2				
AODE-SR	14/0/1	4/5/6	6/5/4	4/5/6			
KDB	13/2/0	8/5/2	8/5/2	5/6/4	7/5/3		
LKDB	14/1/0	7/7/1	8/6/1	4/9/2	7/6/2	3/11/1	
AKDB	14/1/0	8/6/1	10/4/1	6/8/1	6/7/2	4/8/3	4/8/3

Table 7. W/D/L comparison results of variance on large data sets.

W/D/L	NB	TAN	AODE	TAN-FDA	AODE-SR	KDB	LKDB
TAN	4/1/10						
AODE	4/4/7	10/1/4					
TAN-FDA	3/1/11	2/6/7	3/1/11				
AODE-SR	7/2/6	10/3/2	5/3/7	9/4/2			
KDB	3/2/10	2/3/10	2/2/11	2/4/9	0/5/10		
LKDB	3/2/10	3/2/10	2/3/10	4/1/10	2/2/11	4/10/1	
AKDB	4/1/10	3/9/3	4/1/10	8/6/1	5/2/8	12/1/2	12/1/2

The data sets, in which AODE performs better than KDB, are selected for comparison. Figures 4 and 5 show the experimental results of $\varrho(\cdot)$ with respect to zero-one loss, bias and variance. The index numbers of data sets in Figures 4 and 5, which correspond to that described in Table 1. From Figure 4, we can see clearly that, local KDB works for all the data sets regardless of whether the data size is small or large.

Table 8. Ranks of different classifiers.

Dataset	NB	TAN	AODE	TAN-FDA	AODE-SR	KDB	LKDB	AKDB
Abalone	5.5	5.5	1.5	1.5	5.5	5.5	5.5	5.5
Adult	8.0	4.5	7.0	4.5	1.5	4.5	4.5	1.5
Anneal	8.0	4.0	4.0	4.0	1.0	4.0	7.0	4.0
Audio	2.0	4.5	1.0	7.0	3.0	7.0	7.0	4.5
Balance Scale	5.5	2.0	5.5	5.5	1.0	5.5	5.5	5.5
Breast-cancer-w	1.0	5.5	2.5	4.0	7.0	8.0	2.5	5.5
Car	8.0	5.0	7.0	3.0	3.0	3.0	6.0	1.0
Chess	7.5	4.0	5.5	2.5	1.0	5.5	7.5	2.5
Connect-4	8.0	6.0	6.0	1.5	1.5	3.5	6.0	3.5
Contraceptive-mc	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
Credit-a	2.0	5.5	2.0	8.0	5.5	5.5	5.5	2.0
Cylinder-bands	4.5	8.0	2.0	7.0	6.0	4.5	2.0	2.0
Dermatology	2.0	4.0	1.0	5.5	7.0	8.0	5.5	3.0
Glass Identification	7.5	5.0	7.5	5.0	1.0	5.0	2.5	2.5
Heart	1.5	4.5	1.5	7.5	4.5	7.5	4.5	4.5
Hepatitis	7.5	2.5	5.0	5.0	5.0	7.5	1.0	2.5
Hungarian	2.0	5.0	5.0	2.0	2.0	7.5	7.5	5.0
Iris	6.5	2.5	6.5	6.5	2.5	6.5	2.5	2.5
Kr-vs-kp	8.0	5.5	7.0	3.5	5.5	2.0	3.5	1.0
Labor	1.5	5.0	5.0	8.0	5.0	1.5	5.0	5.0
Led	4.5	4.5	4.5	4.5	4.5	4.5	4.5	4.5
Letter-recog	8.0	7.0	3.5	2.0	3.5	5.0	6.0	1.0
Localization	8.0	6.5	6.5	3.0	3.0	3.0	5.0	1.0
Lung-cancer	1.0	5.0	2.0	8.0	7.0	5.0	5.0	3.0
Lymphography	1.0	3.5	3.5	6.5	6.5	8.0	3.5	3.5
Magic	8.0	6.0	6.0	3.5	1.5	3.5	6.0	1.5
Mushrooms	8.0	7.0	6.0	5.0	4.0	3.0	2.0	1.0
Nursery	8.0	5.5	7.0	1.5	3.5	1.5	5.5	3.5
Optdigits	8.0	6.0	2.5	4.0	1.0	5.0	7.0	2.5
Poker-hand	7.5	6.0	7.5	5.0	4.0	3.0	1.0	2.0
Primary-tumor	1.5	1.5	5.5	5.5	5.5	5.5	5.5	5.5
Satellite	8.0	5.5	5.5	3.0	3.0	3.0	7.0	1.0
Segment	8.0	4.5	2.0	4.5	6.0	7.0	2.0	2.0
Shuttle	8.0	7.0	3.5	3.5	3.5	3.5	3.5	3.5
Sick	7.5	5.5	5.5	3.5	3.5	1.5	7.5	1.5
Spambase	8.0	6.0	6.0	3.5	1.5	3.5	6.0	1.5
Teaching-ae	5.0	7.5	2.5	5.0	5.0	7.5	2.5	1.0
Vehicle	8.0	6.0	3.0	1.0	3.0	6.0	6.0	3.0
Vowel	8.0	2.0	4.0	6.0	2.0	5.0	7.0	2.0
Wine	8.0	5.5	2.0	5.5	4.0	2.0	7.0	2.0
Zoo	5.5	2.0	5.5	5.5	5.5	8.0	3.0	1.0
Avg	5.8	5.0	4.4	4.5	3.8	4.9	4.8	2.8

Bias can be used to evaluate the extent to which the final model learned from training data fits the entire data set. From Table 6, we can see that the fitness of NB is the poorest because its structure is

definite regardless of the true data distribution. AKDB still performs the best, although the advantage is not significant. By calculating CMI from the global viewpoint and calculating CLMI from the local point, the aggregating mechanism can help AKDB make full use of the information that is supplied by the training data and test instance. The complicated relationship among attributes are measured and depicted from the viewpoint of information theory. Thus, performance robustness can be achieved. In two data sets, the KDB performs more poorly than AODE. From Figure 5, we observed that AKDB works in one data set.

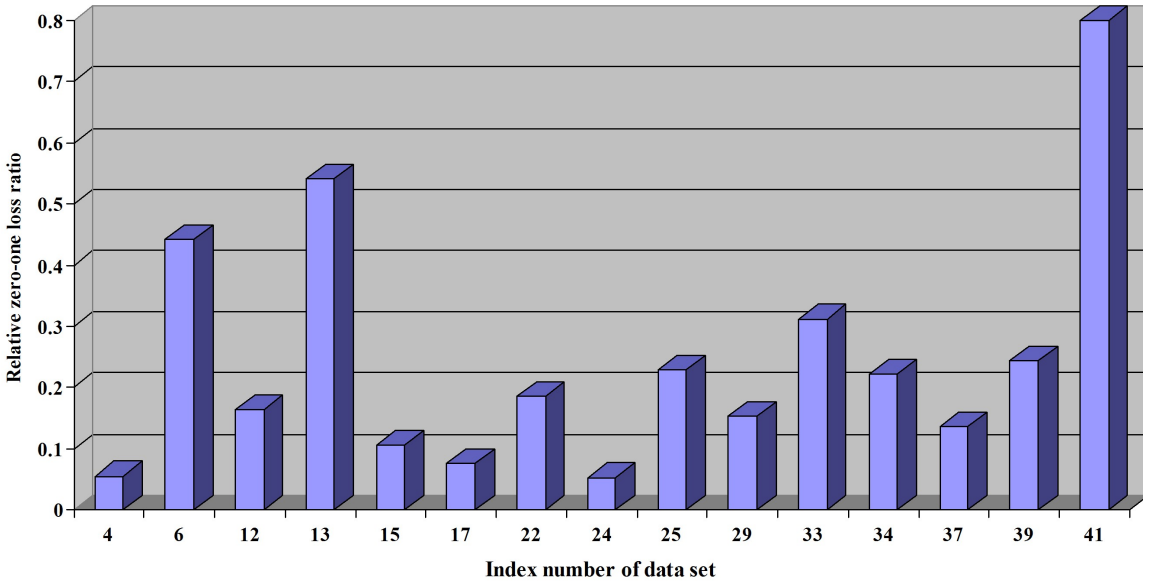


Figure 4. The comparison results of relative zero-one loss ratio.

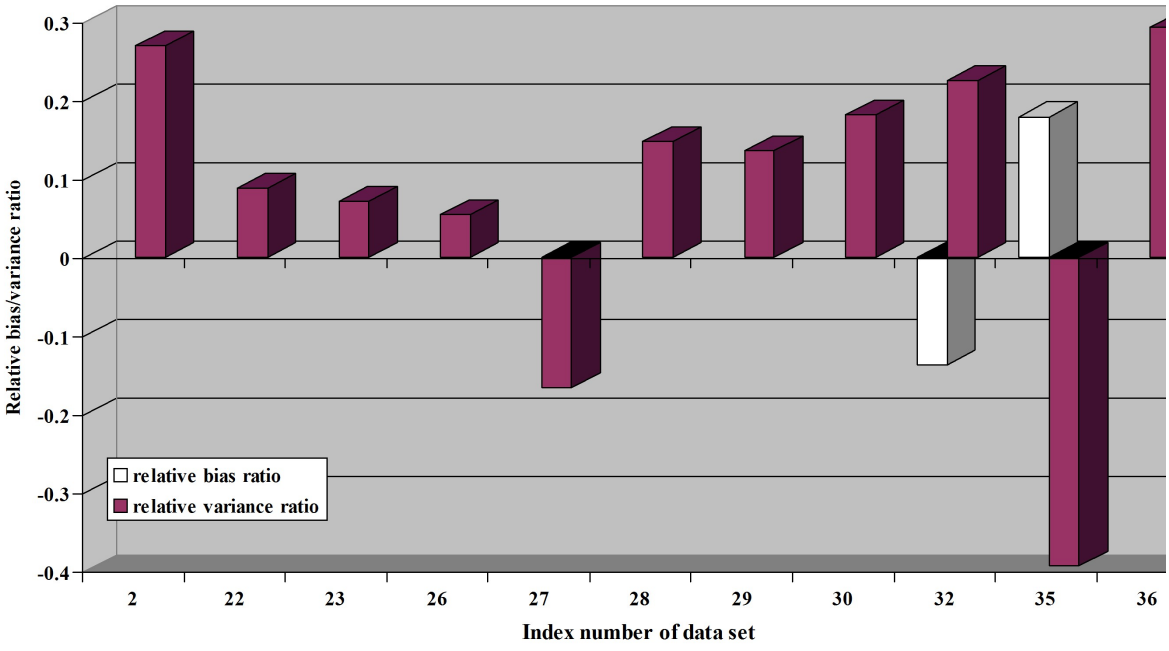


Figure 5. The comparison results of relative bias/variance ratio.

With respect to variance, NB performs the best among these algorithms because its network structure is definite and is therefore insensitive to changes in the training set, as shown in Table 7. By contrast, KDB performs the worst. When k increases, the resulting network tends to have a complex structure. Thus, the network has high variance because of the inaccurate probability estimation caused by the limited amount of training data. Meanwhile, for the local KDB, only the attribute values in the test instance are needed to compute the CLMI. The negative effect caused by probability distribution will be mitigated significantly. Moreover, FDs are extracted from the entire data set and entirely unrelated to the training set. From Figure 5, we observed that the local KDB expresses significant complementary characteristics. Moreover, in 9 of 11 data sets, AKDB performs better than KDB.

5. Conclusions

We propose to build a local KDB as a complementary part of KDB to describe some specific situations to retain the high-dependence representation characteristic of KDB and aggregating mechanism of AODE. The final model, AKDB, has shown its superiority from the comparison results of zero-one loss, bias, and variance. The local KDB is trained in the framework of KDB. Similarly, applying the basic idea of the current work to other high-dependence Bayesian classifiers is possible.

Acknowledgments

This work was supported by the National Science Foundation of China (Grant No. 61272209, 61300145) and the Postdoctoral Science Foundation of China (Grant No. 2013M530980), Agreement of Science & Technology Development Project, Jilin Province (No. 20150101014JC).

Author Contributions

All authors have contributed to the study and preparation of the article. The 1st author conceived the idea, derived equations and wrote the paper. The 2nd author and the 3rd author did the analysis. The 4th author finish the programming work. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Morgan Kaufmann: Burlington, MA, USA, 1988.
2. Cheng, J.; Greiner, R.; Kelly, J.; Bell, D.; Liu, W. Learning Bayesian Networks from Data: An Information-Theory Based Approach. *Artif. Intell.* **2002**, *137*, 43–90.
3. Jiang, L.X.; Cai, Z.H.; Wang, D.H. Improving Tree Augmented Naive Bayes for Class Probability Estimation. *Knowl. Base. Syst.* **2012**, *26*, 239–245.

4. Francisco, L.; Anderson, A. Bagging k-Dependence Probabilistic Networks: An Alternative Powerful Fraud Detection Tool. *Expert. Syst. Appl.* **2012**, 11583–11592.
5. Bielza, C.; Larranaga, P. Discrete Bayesian Network Classifiers: A Survey. *ACM Comput. Surv.* **2014**, *47*, 1–43.
6. Cooper, G.F. The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks. *Artif. Intell.* **1990**, *42*, 393–405.
7. Dagum, P.; Luby, M. Approximating Probabilistic Inference in Bayesian Belief Networks is NP-Hard. *Artif. Intell.* **1993**, *60*, 141–153.
8. Langley, P.; Iba, W.; Thompson. K. An Analysis of Bayesian Classifiers. In Proceedings of the Tenth National Conference on Artificial Intelligence, San Jose, CA, USA, 12–16 July 1992; pp. 223–228.
9. Friedman, N.; Geiger, D.; Goldszmidt, M. Bayesian Network Classifiers. *Mach. Learn.* **1997**, *29*, 131–163.
10. Sahami, M. Learning Limited Dependence Bayesian Classifiers. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*; AAAI Press: Palo Alto, CA, USA, 1996; pp. 335–338.
11. Watanabe, S. A Widely Applicable Bayesian Information Criterion. *J. Mach. Learn. Res.* **2013**, *14*, 867–897.
12. Chaitankar, V.; Ghosh, P.; Perkins, E. A Novel Gene Network Inference Algorithm Using Predictive Minimum Description Length Approach. *BMC. Syst. Biol.* **2010**, *4*, 107–126.
13. Posada, D.; Buckley, T.R. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches over Likelihood Ratio Tests. *Syst. Biol.* **2004**, *53*, 793–808.
14. Friedman, N.; Koller, D. Being Bayesian about Bayesian Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks. *Mach. Learn.* **2013**, *50*, 95–125.
15. Webb, G.I.; Boughton, J.; Wang, Z. Not So Naive Bayes: Aggregating One-Dependence Estimators. *Mach. Learn*. **2005**, *58*, 5–24.
16. Zheng, F.; Webb, G.I. Subsumption Resolution: An Efficient and Effective Technique for Semi-Naive Bayesian Learning. *Mach. Learn*. **2012**, *87*, 1947–1988.
17. Wang, L.M. Extraction of Belief Knowledge from a Relational Database for Quantitative Bayesian Network Inference. *Math. Probl. Eng.* **2013**, doi.org/10.1155/2013/297121.
18. Wang, L.M.; Wang, S.C.; Li, X.F.; Chi, B.R. Extracting Credible Dependencies for Averaged One-Dependence Estimator Analysis. *Math. Probl. Eng.* **2014**, doi.org/10.1155/2014 /470821.
19. Shannon C.E.; Weaver, W. *The Mathematical Theory of Communication*; University of Illinois Press: Champaign, IL, USA, 1949.
20. De Raedt, L. Logic of Generality. In *Encyclopedia of Machine Learning*; Sammut, C., Webb, G.I., Eds.; Springer: New York, NY, USA, 2010; pp. 624–631.
21. Kohavi, R.; Wolpert, D. Bias Plus Variance Decomposition for Zero-One Loss Functions. In Proceedings of the Thirteenth International Conference on Machine Learning, Bari, Italy, 3–6 July 1996; pp. 275–283.

22. Fayyad, U.M.; Irani, K.B. Multi-interval Discretization of Continuous-Valued Attributes for Classification Learning. In Proceedings of the 13th International Joint Conference on Artificial Intelligence, Chambery, France, 28 August–3 September, 1993; pp. 1022–1029.
23. Garcia, S.; Herrera, F. An Extension on "Statistical Comparisons of Classifiers over Multiple Data Sets" for all Pairwise Comparisons. *J. Mach. Learn. Res.* **2008**, *9*, 2677–2694.
24. Friedman, M. The Use of Ranks to Avoid the Assumption of Normality Implicit in the Analysis of Variance. *J. Am. Stat. Assoc.* **1937**, *32*, 675–701.