Applied Intelligence (2023) 53:29799–29817
https://doi.org/10.1007/s10489-023-05098-y

Exploring complex multivariate probability distributions with simple and robust bayesian network topology for classification

Lanni Wang[1] · Limin Wang[1,2] · Lu Guo[3] · Qilong Li[4] · Xiongfei Li[2]

Accepted: 8 October 2023 / Published online: 3 November 2023

Abstract
Bayesian network classifier (BNC) allows efficient and effective inference under condition of uncertainty for classification, and it depicts the interdependencies among random variables using directed acyclic graph (DAG). However, learning an optimal BNC is NP-hard, and complicated DAGs may lead to biased estimates of multivariate probability distributions and subsequent degradation in classification performance. In this study, we suggest using the entropy function as the scoring metric, and then apply greedy search strategy to improve the fitness of learned DAG to training data at each iteration. The proposed algorithm, called One+ Bayesian Classifier (O^+BC), can represent high-dependence relationships in its robust DAG with a limited number of directed edges. We compare the performance of O^+BC with other six state-of-the-art single and ensemble BNCs. The experimental results reveal that O^+BC demonstrates competitive or superior performance in terms of zero-one loss, bias-variance decomposition, Friedman and Nemenyi tests.

Keywords Bayesian network classifier · Multivariate probability distributions · Directed acyclic graph · Classification

1 Introduction

The study on data mining has long been an active research domain in artificial intelligence and machine learning due to its theoretical and practical significance. By analyzing data and performing inductive reasoning on the basis of data mining, researchers expect to extract valuable knowledge and information from data intelligently and automatically [1–7]. Among numerous data mining algorithms, the Bayesian networks (BNs) [8, 9] designed for inference or the Bayesian network classifiers (BNCs) [10–12] for classification encode the probability distribution over a set of random variables using directed acyclic graph (DAG). The DAG qualitatively describes the dependencies (or independencies) among random variables in the form of directed edges (or independent nodes), and then it quantitatively factorizes the joint probability into the product of a set of conditional probabilities.

However, how to learn a BNC that can fit data well is NP-hard [13–15], and the study of restricted BNC has attracted great attention especially after Naive Bayes (NB) [16, 17] is ranked as one of the top ten data mining algorithms. Restricted BNC takes the topology of NB as the skeleton, and an effective and feasible approach to structure learning is adding augmented edges to relax the attribute independence assumption of NB. The newly added directed edges should help make the multivariate probability distribution encoded in the local topology approximate the true one. However, each edge in the DAG, e.g., $X_i \rightarrow X_j$, is commonly measured by conditional mutual information (CMI) $I(X_i; X_j|Y)$ [18]. Thus CMI measures the significance of one single edge rather than the fitness of the learned topology to data.

High-dependence topology can help the learned BNC model complex multivariate probability distributions. To

✉ Limin Wang
wanglim@jlu.edu.cn

Lanni Wang
lnwang21@mails.jlu.edu.cn

1 College of Computer Science and Technology, Jilin University, ChangChun 130012, China

2 Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, ChangChun 130012, China

3 College of Software, Jilin University, ChangChun 130012, China

4 College of Instrumentation and Electrical Engineering, Jilin University, ChangChun 130012, China

achieve excellent generalization and classification performance, the learned DAG should be robust and simple, and can represent high-dependence relationships with a limited number of directed edges. Some researchers proposed to use $I(X_i; X_j|Y)$ to measure high-dependence relationships by implicitly assuming that the parent attributes impact on the children attribute independently rather than jointly [19]. Thus the learned topology is suboptimal and sometimes high-dependence BNC even performs poorer than one-dependence BNC.

Given the topology $\mathcal{B}$ learned from training data $\mathcal{D}$, the optimal value of joint probability should correspond to the maximum value of the log likelihood (LL) function $LL(\mathcal{B}|\mathcal{D})$ or the minimum value of the entropy function $H(X, Y|\mathcal{B}, \mathcal{D})$ [17], where X and Y respectively denote the predictive attribute set and class variable. Since entropy is commonly applied to measure the disorder or randomness, if it is introduced as the scoring function for measuring the uncertainty of the learned topology, then directed edges will have causal semantics and the topology robustness will be retained when 1-dependence topology is scaled up to represent high-order dependencies. The variation in training data or noise will not make the classification performance of the learned robust BNC vary greatly. The main contributions of this paper are described as follows,

- We use entropy based scoring criteria to measure the fitness of learned topology to training data, and then a novel greedy search strategy is introduced to learn high-dependence maximum weighted spanning tree from data. During the learning procedure, the newly added augmented edge should help improve the topology robustness at each iteration. The operations applied include edge addition and edge reversal. The learned $n - 1$ directed edges in the topology of $\mathcal{B}$ can represent from 1-dependence to arbitrary k-dependence relationships.
- The experimental study proves the effectiveness of the proposed entropy function on measuring dependencies and the feasibility of the application of greedy search strategy on labeled data. The proposed O$^+$BC demonstrates competitive or superior classification performance in terms of zero-one loss, bias, variance, Friedman and Nemenyi test, and it can handle different domains with distinctive properties.

The rest of the paper is structured as follows. Section 2 reviews relevant theoretical background and research work. Section 3 outlines the basic idea and details of the proposed O$^+$BC. Section 4 shows the experimental results of O$^+$BC and its comparison with six state-of-the-art BNCs. Finally, Section 5 concludes the paper.

2 Background theory and related research work

2.1 Directed acyclic graph

A succinct summary of the symbols used in this paper are listed in Table 1.

The topology of BN graphically represents the dependencies in the form of DAG, which encodes the probability distributions learned from data. Given a finite set $X = \{X_1, \ldots, X_n\}$ of discrete random variables, a BN for X contains two components, i.e., $\mathcal{G}$ and Θ. $\mathcal{G}$ is the learned DAG with vertices corresponding to the random variables and edges representing direct dependencies between the variables. Θ represents the set of probability distributions that quantifies the DAG. A BN $\mathcal{B}$ defines the joint probability distribution as follows,

$$P(x_1, \ldots, x_n|\mathcal{B}) = \prod_{i=1}^{n} P(x_i|\Pi_i) \tag{1}$$

For restricted BNC, the class variable Y is considered as the root node of DAG or the common parent of all the variables in X. The main objective of BNC learning [20] is

Table 1 Summary of the symbols used in this paper

Symbols	Description
$\mathcal{G}$	the directed acyclic graph
$\mathcal{B}$	the Bayesian network
$\mathcal{D}$	the training dataset
Θ	the set of probability distributions
X_i	the predictive attribute
x_i	the value of X_i
Y	the class variable
y	the value of Y
$X = \{X_1, \ldots, X_n\}$	the set of attributes
$\mathbf{x} =< x_1, \cdots, x_n >$	the unlabeled testing instance
Π_i	the parents of node X_i in $\mathcal{G}$
π_i	the value of Π_i
y^*	the predicted class label
N	the number of training instances
m	the number of class labels
n	the number of attributes
v	the maximum number of discrete values per attribute
k	the maximum number of parents per attribute

to build DAG for representing the dependency relationships among these variables, and then compute the posterior probability of the class given any configuration of the attributes, $P(y|\boldsymbol{x})$.

2.2 Semi-naive Baysian classifiers

NB [21] is the simplest BNC without any directed edges between predictive attributes in its fixed topology. The high-confidence estimates of the 1-order conditional probability $P(y|x_i)$ for each attribute help NB exhibit surprisingly competitive classification accuracy, and it performs significantly better than other more sophisticated BNCs on numerous datasets especially when there exists no correlation between attributes [11, 17, 22–24]. Semi-naive Baysian classifiers (SNBs) [25] take the network topology of NB as the skeleton and then relax the independence assumption by adding augmented edges. The SNBs can be grouped according to the maximum number of parent attributes.

Friedman et al. [17] proposed to fully mine one-dependence relationships from data and the resulting algorithm is the Tree-augmented naive Bayesian network (TAN) (see Fig.1(a)). Each attribute in the framework of maximum weighted spanning tree (MWST) can have no more than one parent attribute and all attributes are required to point outward from the randomly selected root attribute. This alleviates NB's independence assumption to some extent and reduces the search space at the expense of the reasonability of directionality. Considering the random selection of root attribute, Jiang et al. [26] suggested to learn an ensemble rather an individual TAN by taking each attribute as the potential root attribute in turn. Based on statistical n-gram language modeling, Feng et al. [27] proposed to employ Markov chain to model adjacent attribute dependencies.

Sahami [28] further extended NB to represent arbitrary k-dependence relationships. The resulting algorithm called k-dependence Bayesian classifier or KDB for simplicity (see Fig.1(b)) allows every attribute to be conditioned to at most k parent attributes. Sahami argued that a BNC would be expected to obtain optimal Bayesian accuracy given enough data and large k. KDB also disregards the identification of directionality. KDB compares mutual information (MI) $I(X_i; Y)$ [29] to sort the attributes in descending order, and only when $i < j$ holds the directed edge $X_i \rightarrow X_j$ possibly exists. To achieve the necessary efficiency and remove memory size as a bottleneck, Ana et al. [19] proposed to apply leave-one-out cross validation (LOOCV) to select attribute subset and appropriate parameter k. Then semi-supervised KDB (SSKDB) algorithm [30] applied heuristic search strategy to learn BNCs that can work jointly in the framework of semi-supervised learning. Jiang et al. [31] proposed to use MI as the weighting metric to assign weights to SPODE members in AODE. Duan et al. [32] considered the difference between SPODE members in AODE, and the weighting metric may vary greatly from instance to instance.

2.3 The scoring metrics and directionality identification

Information-theoretic metrics, e.g., AIC, BIC, MDL and MML [33–37], can measure the degree of fitness of a DAG to training data with distinct characteristics, thus the learned simple and robust topology may achieve excellent generalization and classification performance. From the perspective of information theory, the learned topology aims to minimize the conditional entropy of each variable given its parents and then greedy search strategy provides a feasible approach to find the parents that can give as much information as possible about this variable [38].

The MDL scoring function $\mathrm{MDL}(\mathcal{B}|\mathcal{D})$ is asymptotically correct as the size of training data increases since the learned probability distribution may approximate the true one [39], and it represents the combined length of the network descrip-

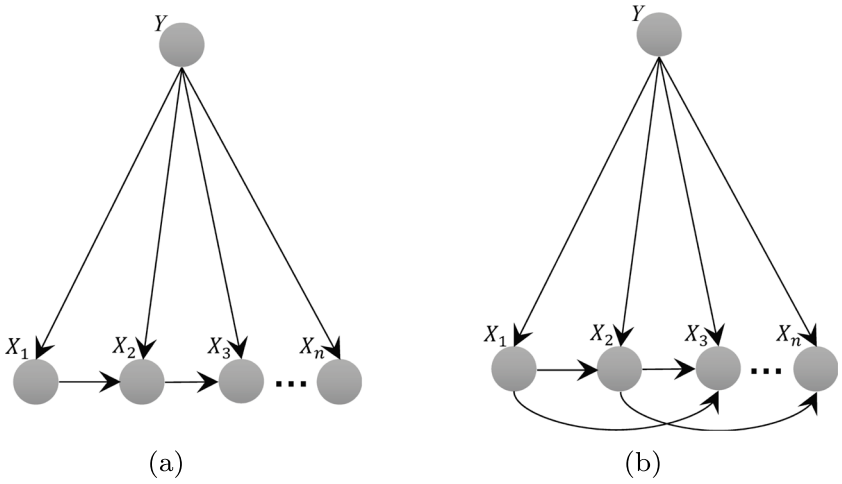


Fig. 1 Examples of (a) TAN (b) KDB with $k = 2$

tion as follows [17],

$$\mathrm{MDL}(\mathcal{B}|\mathcal{D}) = \frac{\log N}{2}|\mathcal{B}| - LL(\mathcal{B}|\mathcal{D}), \tag{2}$$

where $|\mathcal{B}|$ denotes the number of parameters for $\mathcal{B}$. Thus the first term can be considered as a constant given $\mathcal{B}$. Given the joint probability distribution $P_{\mathcal{B}}$ over the N instances $\{d_1, \cdots, d_N\}$ appearing in the training data $\mathcal{D}$, the second term in (2), i.e., $LL(\mathcal{B}|\mathcal{D})$, is commonly applied to measured the quality of $\mathcal{B}$ and can be factorized into the following form,

$$\begin{aligned} LL(\mathcal{B}|\mathcal{D}) &= \sum_{i=1}^{N} \log P(d_i|\mathcal{B}) = \sum_{X,Y} P(\boldsymbol{x}, y) \log P_{\mathcal{B}}(\boldsymbol{x}, y) \\ &= \sum_{Y} P(y) \log P(y) + \sum_{X,Y} P(\boldsymbol{x}, y) \log P_{\mathcal{B}}(\boldsymbol{x}|y) \end{aligned} \tag{3}$$

Maximizing the log likelihood function $LL(\mathcal{B}|\mathcal{D})$ is equivalent to minimizing the entropy function $H(X, Y|\mathcal{B}, \mathcal{D})$ [17], and (3) turns to be

$$\begin{aligned} H(X, Y|\mathcal{B}, \mathcal{D}) &= -LL(\mathcal{B}|\mathcal{D}) = H(Y) + H(X|Y) \\ &= H(Y) + \sum_{i=1}^{n} H(X_i|\Pi_i, Y) \end{aligned} \tag{4}$$

where

$$\begin{cases} H(Y) = -\sum\limits_{Y} P(y) \log P(y) \\ H(X_i|\Pi_i, Y) = -\sum\limits_{X_i} \sum\limits_{\Pi_i} \sum\limits_{Y} P(x_i, \pi_i, y) \log(x_i|\pi_i, y) \end{cases} \tag{5}$$

The entropy function corresponding to NB is

$$H(X, Y|\mathrm{NB}, \mathcal{D}) = H(Y) + \sum_{i=1}^{n} H(X_i|Y) \tag{6}$$

Researchers commonly apply greedy search strategy and add augmented edges between attributes for the obvious computational reasons. The learning procedure starts with the topology of NB and successively applies local operations to maximize the score until a local minima is reached. The difference between (4) and (6), i.e., $\sum_{i=1}^{n} I(X_i; \Pi_i|Y) = H(X, Y|\mathrm{NB}, \mathcal{D}) - H(X, Y|\mathcal{B}, \mathcal{D})$, measures the quality of the conditional dependencies between attributes mined from data. Due to the symmetry form of CMI, we need to make rule to sort the attributes explicitly or implicitly, and then determine the directionality of the edge $X_i - X_j$ or which one is the parent for attribute pair $\{X_i, X_j\}$.

3 Learning robust Bayesian network classifier with simple topology

For restricted BNCs, as shown in Fig. 2 conditional independence $\mathcal{B}_{ind}$ and conditional dependence $\mathcal{B}_{de}$ are the two basic local topologies. The CMI is commonly used to measure the significance of conditional dependence between attribute pair, and it can be factorized in the form of entropy functions as follows [40],

$$\begin{aligned} I(X_i; X_j|Y) &= \sum_{X_i} \sum_{X_j} \sum_{Y} P(y, x_i, x_j) \log \frac{P(y, x_i, x_j|\mathcal{B}_{de})}{P(y, x_i, x_j|\mathcal{B}_{ind})} \\ &= \sum_{X_i} \sum_{X_j} \sum_{Y} P(y, x_i, x_j) \log P(y, x_i, x_j|\mathcal{B}_{de}) \end{aligned}$$

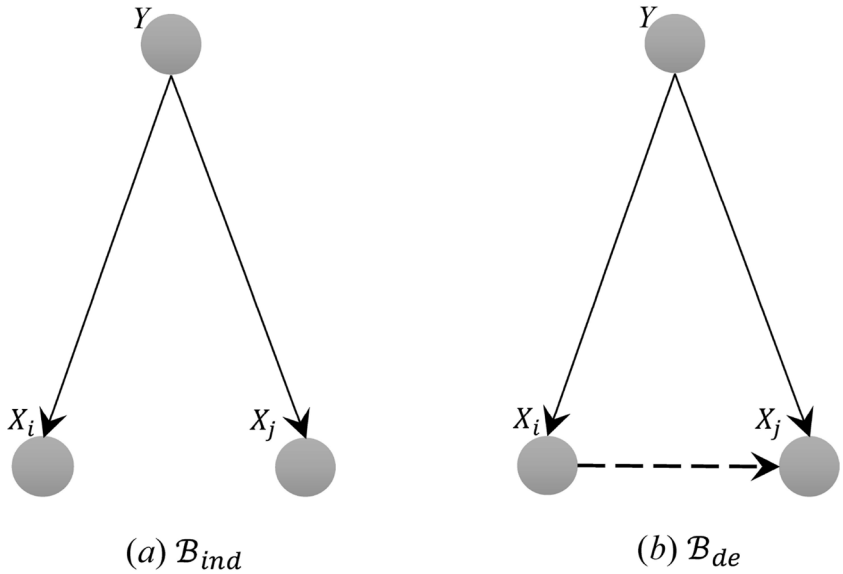


Fig. 2 Two basic topologies depicting *a*) the conditional independence and *b*) the conditional dependence respectively

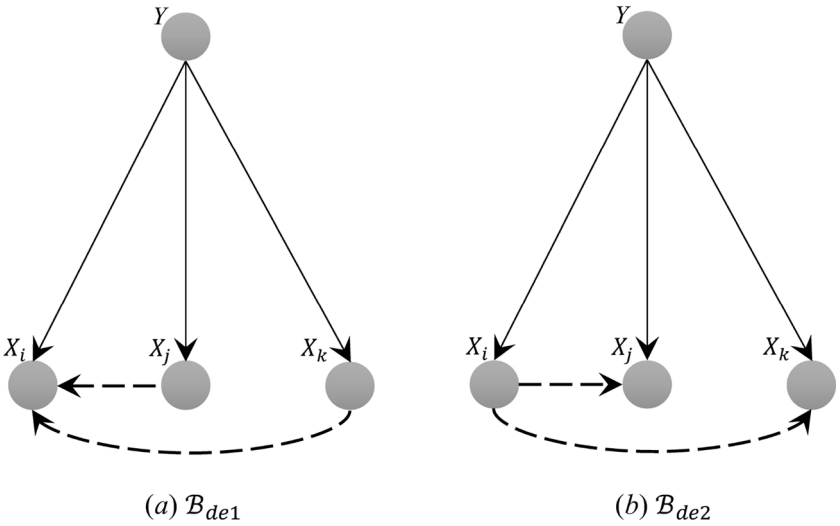


Fig. 3 Two local topologies with directed edges *a*) pointing to X_i or *b*) pointing from X_i

$(a)\ \mathcal{B}_{de1}$ $(b)\ \mathcal{B}_{de2}$

$$
\begin{aligned}
&-\sum_{X_i}\sum_{X_j}\sum_{Y} P(y, x_i, x_j)\log P(y, x_i, x_j|\mathcal{B}_{ind}) \\
&= H(\mathcal{B}_{ind}) - H(\mathcal{B}_{de}) \qquad (7)
\end{aligned}
$$

where

$$
\begin{cases}
P(y, x_i, x_j|\mathcal{B}_{de}) = P(x_i, x_j|y) = P(x_i, x_j|y)P(y) \\
P(y, x_i, x_j|\mathcal{B}_{ind}) = P(x_i|y)P(x_j|y) = P(x_i|y)P(x_j|y)P(y)
\end{cases} \qquad (8)
$$

Thus the underlying essence of $I(X_i; X_j|Y)$ is to evaluate the extent to which $\mathcal{B}_{de}$ is more reasonable than $\mathcal{B}_{ind}$, whereas $I(X_i; X_j|Y)$ cannot directly measure the extent to which $\mathcal{B}_{de}$ fits data [41]. Furthermore, the aim to learn the topology of BNC is to explore complex multivariate probability distributions from data. Although the 2-dependence topology shown in Fig. 3(a) and the 1-dependence topology shown Fig. 3(b) have the same skeleton, considering the impact of attribute X_k, the significance of $X_i \rightarrow X_j$ and $X_j \rightarrow X_i$ may vary greatly. Thus $I(X_i; X_j|Y)$ measures the significance of one single edge whereas the entropy function $H(X, Y|\mathcal{B}, \mathcal{D})$ in (4) measures the fitness of the learned topology to data, and the latter is considered more appropriate to be the scoring function for measuring the robustness of the learned topology.

3.1 The initial topology of O$^+$BC with two nodes

For high-dependence BNCs, the number of directed edges or dependency relationships increases as the topology complexity increases, and the learned high dimensional probability distributions may fit data better especially while dealing with large dataset. In contrast, TAN allows at most one parent for every attribute node, and its 1-dependence topology cannot represent high dimensional probability distributions, thus TAN commonly demonstrates superior performance while dealing with small or medium sized datasets [42]. To prove the effectiveness of the scoring function in learning high-dependence topology with a limited number of directed edges, the proposed algorithm O$^+$BC also takes $n-1$ directed edges in the topology as TAN. We will clarify the basic idea in detail in the following discussion.

O$^+$BC initializes the network topology of $\mathcal{B}$ as an empty topology, and then applies local search operations to add attribute X_i and the edges connecting to it to the topology at each iteration. During the learning procedure, the edges connecting to X_i should help achieve local optimal solution, or more precisely, the minimum value of $H(X, Y|\mathcal{B}, \mathcal{D})$. The operations applied include edge addition and edge reversal. Assuming that initially there are two attribute nodes $\{X_1, X_2\}$ and the class node Y in the topology, as shown in Fig. 4 the directed edge between X_1 and X_2 may be either $X_1 \rightarrow X_2$ or $X_2 \rightarrow X_1$. The entropy functions corresponding to one-dependence BNCs $\mathcal{B}_{12}$ and $\mathcal{B}_{21}$ are respectively

$$
\begin{cases}
H(\mathcal{B}_{12}) = H(Y) + H(X_2|X_1, Y) + H(X_1|Y) = H(X_2, X_1, Y) \\
H(\mathcal{B}_{21}) = H(Y) + H(X_1|X_2, Y) + H(X_2|Y) = H(X_2, X_1, Y).
\end{cases} \qquad (9)
$$

Thus $H(X_1, X_2, Y) = H(\mathcal{B}_{12}) = H(\mathcal{B}_{21})$ always holds, and the topologies for $\mathcal{B}_{12}$ and $\mathcal{B}_{21}$ are both reasonable and we can randomly select one of them as the candidate one-dependence topology of O$^+$BC. The corresponding pseudo codes are shown as follows,

However, the number of bits encoded in the undirected edge $X_i - \{X_j, X_k\}$ may be different from that encoded in $\{X_i, X_j\} - X_k$ for high-dependence BNC. Thus while dealing with three or more attributes, it is difficult or even an NP-

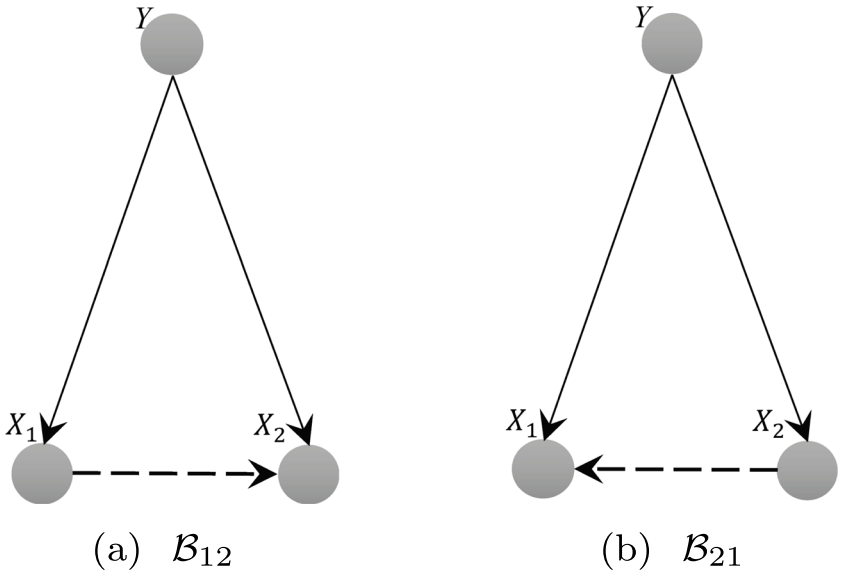


Fig. 4 Two kinds of initial topologies with directed edge a) $X_1 \rightarrow X_2$ or b) $X_2 \rightarrow X_1$

Algorithm 1 InitialModeling($\mathcal{D}$).

Input: Training set $\mathcal{D}$ with attributes $X = \{X_1, ..., X_n\}$ and class variable Y.
Output: The DAG $\mathcal{G}$ with the attribute set X as nodes.
1 Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ be a directed graph in which $\mathbf{V}$ is a set of vertices and $\mathbf{E}$ is a set of directed edges.
2 $\mathbf{V} = \{Y\}$, $\mathbf{E} = \emptyset$.
3 Calculate $H(X_i, X_j, Y)$ for each pair of attributes in $\mathcal{D}(i \neq j)$.
4 Select the attribute pair $< X_i, X_j >$ with the minimum value of $H(X_i, X_j, Y)$.
5 $\mathbf{E} = \mathbf{E} \cup \{Y \rightarrow X_i, Y \rightarrow X_j\}$, $\mathbf{V} = \mathbf{V} \cup \{X_i, X_j\}$, $X = X \setminus \{X_i, X_j\}$.
6 Randomly select $X_i \rightarrow X_j$ or $X_j \rightarrow X_i$ and add it to $\mathbf{E}$.
7 return $\mathcal{G}$.

hard problem to capture all appropriate correlations between attributes [43] and simultaneously differentiate between the parent attributes and the children attributes.

3.2 High-dependence maximum weighted spanning tree

To transform undirected tree to directed one, TAN requires that the root node point outwards, and this rule helps simplify the learning procedure whereas restricts the learning flexibility. We argue that the resulting one-dependence topology may be suboptimal since it fails to consider other possible transformations. The $n-1$ directed edges can represent from 1-dependence to arbitrary k-dependence relationships. Since the structure learning complexity increases exponentially with k, we set $k = 2$ in the following discussion and experimental study.

The augmented edges pointing to or from the newly added attribute should help minimize the entropy function at each iteration. For example, if we need to add one more attribute node, e.g., X_3, to the initial one-dependence topology of O$^+$BC learned by Algorithm 1, then the attribute set in the topology includes $\{X_1, X_2, X_3\}$. The undirected edge between X_3 and $\{X_1, X_2\}$ may be either $X_3 - X_1$ or $X_3 - X_2$, and the resulting four directed topologies are shown in Fig. 5. The corresponding entropy functions are respectively

$$\begin{cases} H(\mathcal{B}_{23}) = H(Y) + H(X_1|Y) + H(X_2|X_1, Y) + H(X_3|X_2, Y) \\ H(\mathcal{B}_{32}) = H(Y) + H(X_1|Y) + H(X_2|X_1, X_3, Y) + H(X_3|Y) \\ H(\mathcal{B}_{13}) = H(Y) + H(X_1|Y) + H(X_2|X_1, Y) + H(X_3|X_1, Y) \\ H(\mathcal{B}_{31}) = H(Y) + H(X_1|X_3, Y) + H(X_2|X_1, Y) + H(X_3|Y). \end{cases} \tag{10}$$

The topologies for $\mathcal{B}_{23}$, $\mathcal{B}_{13}$ and $\mathcal{B}_{31}$ are one-dependence, and only the topology for $\mathcal{B}_{32}$ is two-dependence since X_2 has two parent attributes $\{X_1, X_3\}$. Due to the restriction in structure complexity, k-dependence BNC can select for each attribute no more than k parents attributes. Suppose that $H(X_1, X_2, X_3, Y|\mathcal{B}_{32}) < H(X_1, X_2, X_3, Y|\mathcal{B}_{23}) < H(X_1, X_2, X_3, Y|\mathcal{B}_{13}) = H(X_1, X_2, X_3, Y|\mathcal{B}_{31})$ holds, then the topology for $\mathcal{B}_{32}$ is the optimal one among all, whereas the topology for $\mathcal{B}_{23}$ is the optimal one among one-dependence topologies. The description above gives an example to illustrate the learning procedure after X_3 is selected as the next attribute. In practice, each attribute other than $\{X_1, X_2\}$ in the attribute list will be tested to check if it is the right one to be added. So we need to test the directed edge connecting to the candidate attribute, and finally select the augmented topology which is the most helpful for improving $H(X, Y|\mathcal{B}, \mathcal{D})$. Thus we apply a greedy search to iteratively add the attributes and corresponding directed edges to the topology. The algorithm terminates until all the attributes are included in the topology of $\mathcal{B}$. The learning procedure of adding a new directed edge can be summarized as shown in

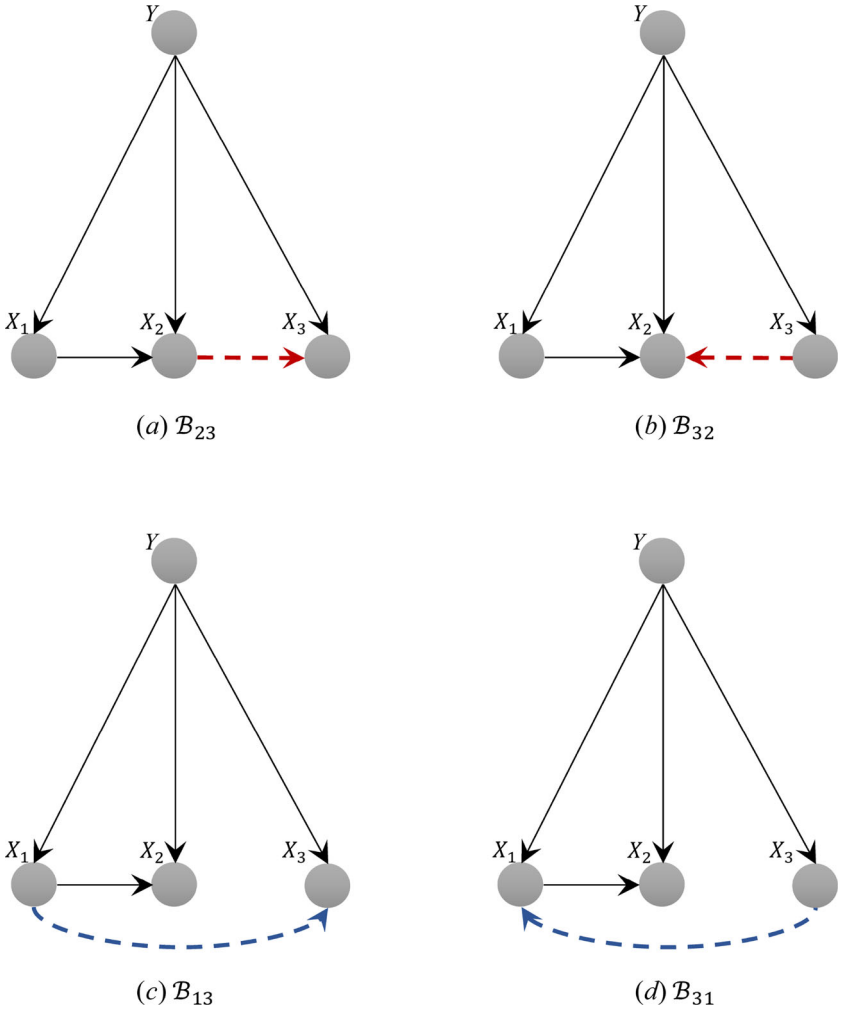


Fig. 5 Four types of directed topologies with three attribute nodes, including *a*) the topology with directed edge $X_2 \rightarrow X_3$, *b*) the topology with directed edge $X_3 \rightarrow X_2$, *c*) the topology with directed edge $X_1 \rightarrow X_3$ and *d*) the topology with directed edge $X_3 \rightarrow X_1$

Algorithm 2. The O$^+$BC algorithm is illustrated in Fig. 6 and described in Algorithm 3.

Given a testing instance $\boldsymbol{x}$, O$^+$BC calculates the joint probability distribution and assigns y^* to $\boldsymbol{x}$ using

$$y^* = \arg\max_Y P(y|\boldsymbol{x}) = \arg\max_Y \frac{P(\boldsymbol{x}, y)}{P(\boldsymbol{x})} \propto \arg\max_Y P(\boldsymbol{x}, y). \quad (11)$$

At training time, O$^+$BC generates a three-dimensional table of co-occurrence counts for each pair of attribute values and each class value, which needs $O(Nn^2)$ time. O$^+$BC selects two attributes to build an initial model by computing $H(X_i, X_j, Y)$ in (9), which has a time complexity of $O(m(nv)^2)$. Assuming that each attribute can take a maximum of k parent attributes, O$^+$BC then iteratively selects the attributes and the corresponding directed edges and adds them to the topology by calculating the corresponding entropy function $H(\mathcal{B})$ in (4). Hence, the corresponding time complexity is $O(m(nv)^{k+1})$. At classification time, O$^+$BC calculates the joint probability distribution in (11) for classification and it only requires $O(mnk)$ time.

Algorithm 2 AddingOneEdge($\mathcal{G}$, Y, k).

Input: $\mathcal{G}$ with the attribute set X as nodes, class variable Y and parameter k.
Output: $\mathcal{G}$ with the attribute set X as nodes.
Let $\mathcal{Q}$ be a $|X| \times |\mathbf{V}|$ matrix and $\mathcal{Q} = \emptyset$.
for each $X_i \in \mathbf{X}$ **do**
  for each $X_j \in \mathbf{V}$ **do**
    Calculate $H(\mathcal{B}_{ji})$, where $\mathcal{B}_{ji}$ is the BNC with topology $\mathcal{G}_{ji}$ and $\mathcal{G}_{ji} = \mathcal{G} \cup \{X_j \rightarrow X_i\}$.
    $\mathcal{Q}[j][i] = H(\mathcal{B}_{ji})$.
    Calculate $H(\mathcal{B}_{ij})$, where $\mathcal{B}_{ij}$ is the BNC with topology $\mathcal{G}_{ij}$ and $\mathcal{G}_{ij} = \mathcal{G} \cup \{X_i \rightarrow X_j\}$.
    $\mathcal{Q}[i][j] = H(\mathcal{B}_{ij})$.
  end
end
Select the minimum value of $\mathcal{Q}[i][j]$ in $\mathcal{Q}$, where $|\Pi_j| < k$.
$\mathbf{V} = \mathbf{V} \cup X_i$, $\mathbf{E} = \mathbf{E} \cup \{X_i \rightarrow X_j\}$, $X = X \setminus X_i$.
return $\mathcal{G}$.

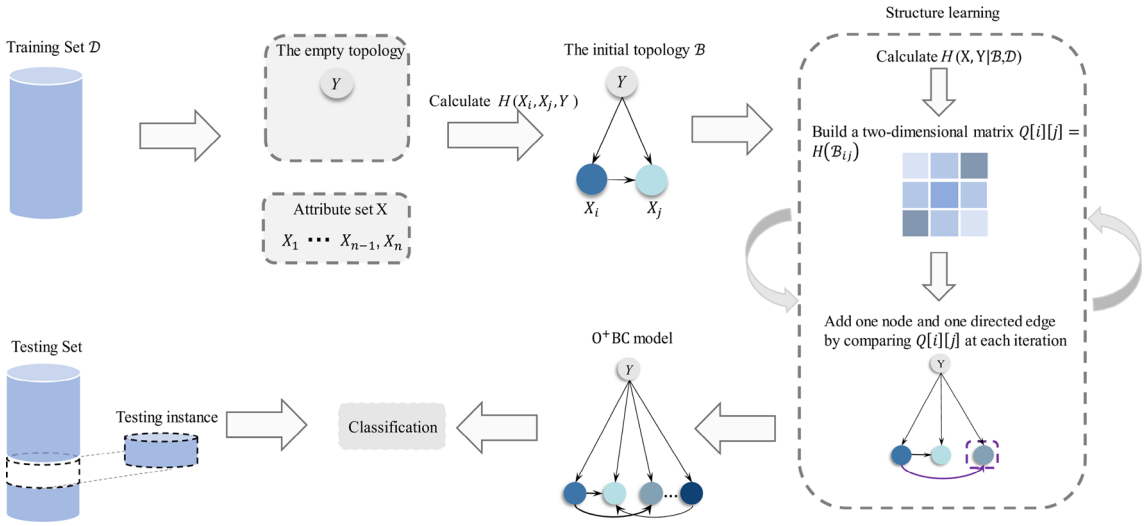


Fig. 6 The learning process of O$^+$BC

4 Experimental study

For comparison study, the state-of-the-art BNCs described as follows run on 44 benchmark datasets from the UCI machine learning repository [44]. Table 2 provides detailed characteristics of these datasets. The missing values for numerical attribute and nominal attribute are respectively replaced by means or modes from the available data. The MDL [45] discretization method is applied to handle numerical attributes for each dataset.

Table 2 The details of datasets

No.	Dataset	Inst	Att	Class	No.	Dataset	Inst	Att	Class
1	Labor	57	16	2	23	Vehicle	846	18	4
2	Zoo	101	16	7	24	Tic-tac-toe	958	9	2
3	Promoters	106	57	2	25	Led	1000	7	10
4	Iris	150	4	3	26	Diabetic-RD	1151	19	2
5	Teaching-ae	151	5	3	27	Contraceptive-mc	1473	9	3
6	Hepatitis	155	19	2	28	Yeast	1484	8	10
7	Wine	178	13	3	29	Volcanoes	1520	3	4
8	Autos	205	25	7	30	Car	1728	6	4
9	Sonar	208	60	2	31	Mfeat-mor	2000	6	10
10	Glass-id	214	9	3	32	Seismic-bumps	2584	18	2
11	Audio	226	69	24	33	Hypo	3772	29	4
12	Hungarian	294	13	2	34	Abalone	4177	8	3
13	Heart-disease-c	303	13	2	35	Electrical-Grid	10000	13	2
14	Soybean-large	307	35	19	36	Firm-Teacher	10800	19	4
15	Primary-tumor	339	17	22	37	Nursery	12960	8	5
16	House-votes-84	435	16	2	38	Magic	19020	10	2
17	Cylinder-bands	540	39	2	39	Adult	48842	14	2
18	Chess	551	39	2	40	Activity-recognition-with	75128	8	4
19	Balance-scale	625	4	3	41	Waveform	100000	21	3
20	Crx	690	15	2	42	Localization	164860	5	11
21	Breast-cancer-w	699	9	2	43	Skin-Segmentation	245057	3	2
22	Pima-ind-diabetes	768	8	2	44	Donation	5749132	11	2

Table 3 The W/D/L records of ZOL for O^+BC against 6 algorithms

O^+BC vs.	TAN	WATAN	SKDB	AODE-SR	WAODE-MI	IWAODE
W/D/L	19/22/3	18/24/2	22/17/5	20/22/2	16/25/3	19/23/2
p	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

Algorithm 3 The O^+BC algorithm.

Input: Training set $\mathcal{D}$ with attributes $X = \{X_1, ..., X_n\}$ and class variable Y, parameter k.
Output: The DAG of O^+BC.
$\{\mathcal{G}\} \leftarrow InitialModeling(\mathcal{D})$. // See Algorithm 1
repeat
  $\{\mathcal{G}\} \leftarrow AddingOneEdge(\mathcal{G}, Y, k)$. // See Algorithm 2
until $X = \emptyset$;
return $\mathcal{G}$.

- TAN [17], which extends NB by representing the one-dependence relationships in the form of MWST.
- SKDB [19], which extends KDB and uses LOOCV to select the best parameter k.
- AODE-SR [46], which eliminates generalizations at classification time by using subsumption resolution.
- WAODE-MI [31], which assigns weight to each SPODE by using MI as the weighting metric.
- IWAODE [32], which defines weights for each SPODE by using instance-based weighting metric.

Each algorithm is tested for 10 rounds on 44 datasets using the 10-fold cross-validation method.

The Win/Draw/Loss (W/D/L) records are used to perform statistical comparisons, and Tables 3, 4 and 5 summarize the number of datasets on which the proposed O^+BC obtains better, similar or worse outcomes relative to the alternative in terms of zero-one loss, bias and variance. The detailed experimental results are respectively given in Tables 7 – 9 in the Appendix. We assess a difference as significant if the outcome of a one-tailed binomial sign test is less than 0.05 [46, 47].

4.1 Zero-one loss

Zero-one loss (ZOL) is a common loss function to measure the classification accuracy. Table 3 provides statistical W/D/L records in terms of ZOL to show the comparison of O^+BC against its competitors.

We can see in Table 3 that O^+BC achieves the best classification performance among all the BNCs. For example, O^+BC enjoys significant advantages over TAN (19/22/3) and SKDB (22/17/5), since O^+BC could represent high-dependence relationships and achieve the tradeoff between data fitness and topology complexity. Among ensemble BNCs, WATAN represents the same undirected dependency relationships in its TAN members, and weighted AODEs don't differentiate significant dependencies from insignificant ones due to its unrealistic independence assumptions. Thus O^+BC outperforms WATAN (18/24/2), AODE-SR (20/22/2), WAODE-MI (16/25/3) and IWAODE (19/23/2).

To further describe the advantage of O^+BC, Fig. 7 shows scatter plots of the experimental results in terms of ZOL, with each point corresponding to one dataset. The three dotted lines in black, blue and red respectively correspond to $Y = 1.1 * X$, $Y = X$ and $Y = 0.9 * X$, which respectively denote that O^+BC performs significantly better, equally well or significantly poorer than its competitor in terms of ZOL. As shown in Fig. 7, O^+BC enjoys significant advantages over TAN, WATAN, SKDB, AODE-SR, WAODE-MI and IWAODE on 11, 11, 16, 10, 9 and 11 datasets, respectively. And O^+BC performs significantly poorer than TAN, WATAN, SKDB, AODE-SR, WAODE-MI and IWAODE only on 1, 1, 4, 1, 2 and 2 datasets, respectively. These illustrative results prove the effectiveness of our proposed O^+BC.

4.2 Bias and variance

The bias-variance decomposition [48] of ZOL provides further insights into the analysis of classification performance. Bias measures the resulting systematic error of the learner for describing the decision boundary, and variance measures the sensitivity of the learner to random variation in the training data.

As shown in Table 4, O^+BC obtains significantly better bias compared to other learners. The 0-dependence topology for NB requires only one pass through the data. The implicit independence assumptions make NB fail to represent high-dependence relationships. The numbers of directed

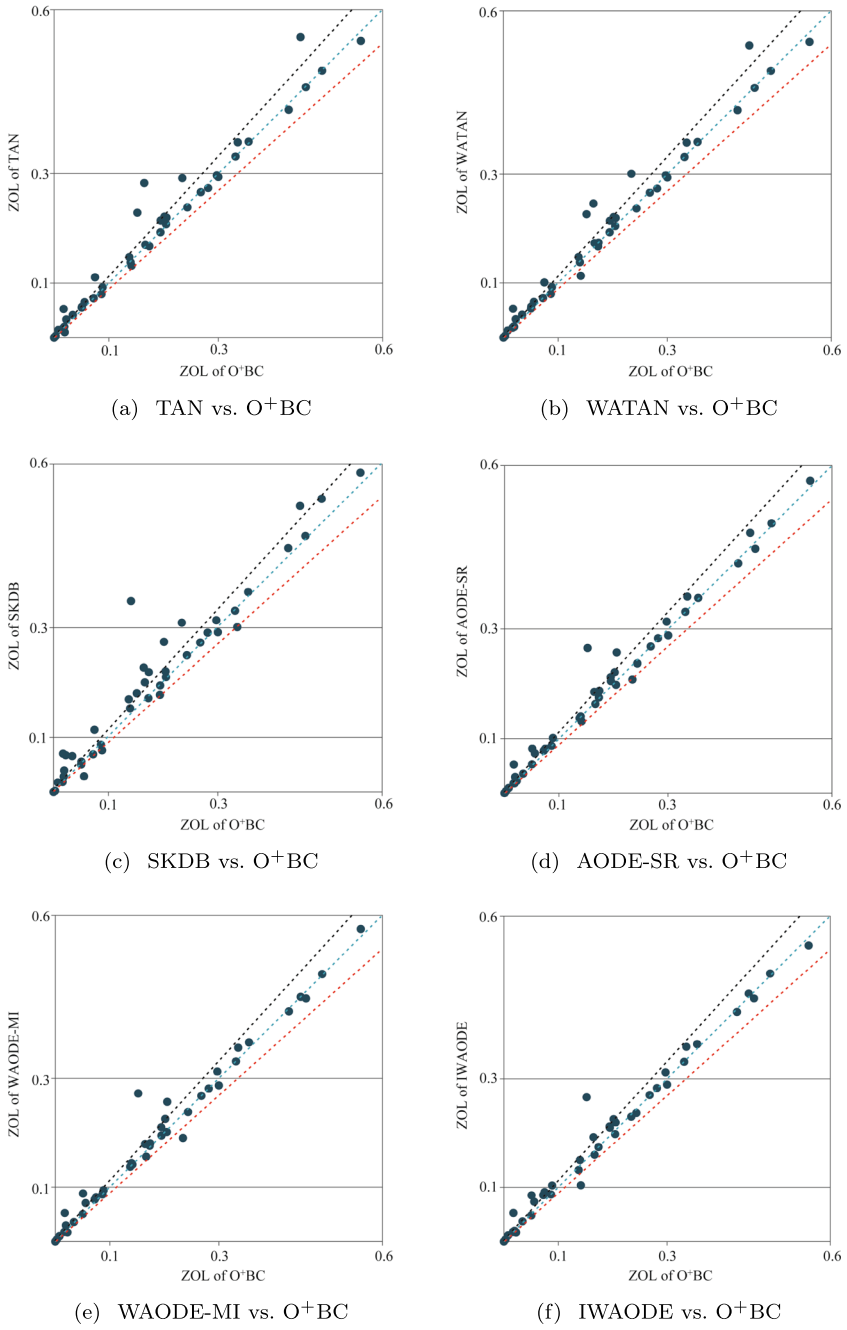
Table 4 The W/D/L records of bias for O^+BC against 6 algorithms

O^+BC vs.	TAN	WATAN	SKDB	AODE-SR	WAODE-MI	IWAODE
W/D/L	17/23/4	18/22/4	19/16/9	20/19/5	18/21/5	24/16/4
p	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05	< 0.05

Table 5 The W/D/L records of variance for O^+BC against 6 algorithms

O^+BC vs.	TAN	WATAN	SKDB	AODE-SR	WAODE-MI	IWAODE
W/D/L	38/2/4	38/2/4	39/2/3	4/11/29	4/9/31	3/8/33
p	<0.05	<0.05	<0.05	<0.05	<0.05	<0.05

edges in DAG for TAN and each member of WATAN are the same, and the same topology skeleton makes them fit training data to the same extent. Thus O^+BC performs much better than TAN (17/23/4) and WATAN (18/22/4) in terms of bias. SKDB and weighted AODE can represent more dependency relationships, whereas complex topology with insignificant dependencies may bias the estimate of probability distributions. Thus O^+BC also outperforms SKDB (19/16/9), AODE-SR (20/19/5), WAODE-MI (18/21/5) and IWAODE (24/16/4).



(a) TAN vs. O^+BC (b) WATAN vs. O^+BC

(c) SKDB vs. O^+BC (d) AODE-SR vs. O^+BC

(e) WAODE-MI vs. O^+BC (f) IWAODE vs. O^+BC

Fig. 7 Scatter plot of comparisons in terms of ZOL

Table 6 Average ranks of the algorithms

Algorithm	ZOL rank	Bias rank	Variance rank
O^+BC	3.0000	3.1023	3.8523
WATAN	3.7159	3.6136	5.5455
TAN	3.8068	4.0227	5.8409
WAODE-MI	4.0341	4.2841	2.6250
IWAODE	4.1477	4.8750	1.9886
AODE-SR	4.2159	4.2841	2.4545
SKDB	5.0795	3.8182	5.6932
Friedman statistic F_F	4.4278	3.3790	74.0856

Variance-wise, the weighting metrics and subsumption resolution can help AODE-SR, WAODE-MI and IWAODE finely tune the probability estimates whereas their independence assumptions reduce the sensitivity to variation in training data. Thus as shown in Table 5, they demonstrate significant advantages over O^+BC in terms of variance. TAN constructs MWST to represent one-dependence relationships between attributes, and WATAN uses weighting metric to mitigate the negative effect caused by random selection of the root attribute. High-dependence topology corresponds to high-order probability distributions and relatively high risk of overfitting. Thus O^+BC demonstrates significant advantages over TAN (38/2/4), WATAN (38/2/4) and SKDB (39/2/3).

4.3 Friedman and Nemenyi test

The Friedman test [49] rank the classification performance of these BNCs in terms of ZOL, bias and variance. The difference in average ranks under the null-hypothesis can help statistically investigate the differences among BNCs. As shown in Table 6, the average ranks of ZOL, bias and variance are respectively 4.4278, 3.3790 and 74.0856. Based on the significance level $\alpha = 0.05$ and the degree of freedom $(7-1) \times (44-1) = 258$, the critical value for Friedman test is 2.1338, thus we can reject the null-hypothesis and there exist significant difference among the compared algorithms.

Since the Friedman test can only conclude whether there exists difference in metrics for evaluating the classification performance of the compared algorithms, we apply the Nemenyi test [50] as a "follow-up test" to find out which algorithms have statistical differences in their performance.

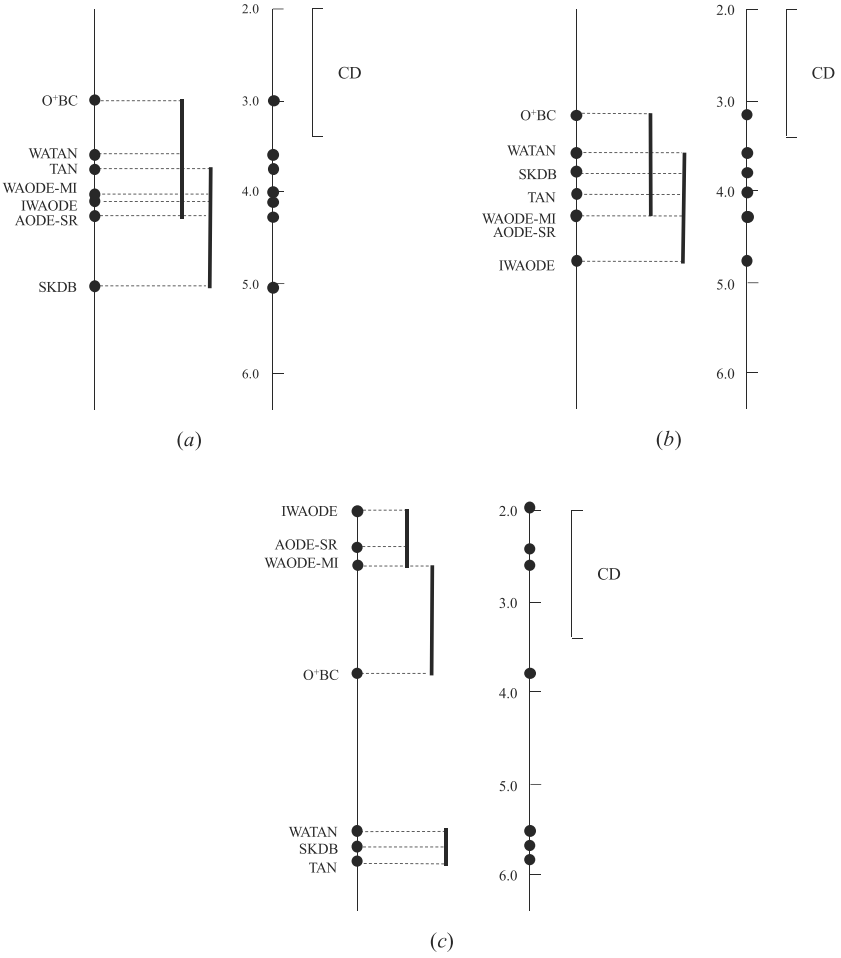


Fig. 8 Comparison of 7 algorithms against each other with the Nemenyi test (*a*) ZOL (*b*) Bias (*c*) Variance

Figure 8 depicts the experimental results of 7 algorithms using the Nemenyi test on ZOL, bias and variance. The BNCs and the matching average ranks are respectively represented by the left line and the parallel right line. If the difference is greater than the critical difference (CD) [51], the algorithm with lower average rank is statistically better than the algorithm with higher average rank. The algorithms will be connected if the differences among them are not significant. The higher position corresponds to lower rank and better performance. The CD value for $\alpha = 0.05$ is 1.3582 for 44 datasets and 7 algorithms.

In Fig. 8(a), the Nemenyi test differentiates O^+BC from other BNCs and the analysis reveals that O^+BC achieves the lowest average ZOL rank followed by WATAN (3.7159), TAN (3.8068), WAODE-MI (4.0341). The differences in ZOL ranks between TAN, WAODE-MI, IWAODE, AODE-SR and SKDB are small, ranging from 3.8068 to 5.0795. Figs. 8(b) and (c) graphically show the compared results of the Nemenyi test on bias and variance. From Fig. 8(b), the bias rank obtained by O^+BC (3.1023) is the lowest followed by WATAN (3.6136) and SKDB achieves the third lowest mean bias rank (3.8182). Although WAODE-MI, AODE-SR and IWAODE can fully represent the conditional dependencies due to ensemble learning, some conditional independencies or weak dependencies may be introduced into the topology of committee members. Thus WAODE-MI, AODE-SR and IWAODE achieve higher mean bias ranks than other BNCs (4.2841, 4.2841 and 4.8750 respectively). Variance-wise, as shown in Fig. 8(c), the top three ranked algorithms are IWAODE (1.9886), AODE-SR (2.4545) and WAODE-MI (2.6250), with no significant difference in performance among them due to their fixed structures. O^+BC achieves lower average variance rank (3.8523) when compared to WATAN (5.5455) and SKDB (5.6932). TAN has the highest mean variance rank (5.8409) since it encodes all significant dependency relationships in one topology.

4.4 Comparision of running time

We will compare the averaged training and classification time in milliseconds of 7 algorithms on 44 UCI benchmark datasets in this subsection. Each bar in Fig. 9 shows the logarithmic average running time of the corresponding algorithm in 10-fold cross validation experimental study. As shown in Fig. 9(a), AODE-SR and IWAODE require the least training time among all the algorithms, since they need no structure learning for training. In contrast, WAODE-MI needs additional time to calculate MI as the attribute weights. TAN utilizes CMI to construct MWST. WATAN constructs n TAN classifiers using n MWSTs. SKDB needs to select attribute subset and parameter k. O^+BC uses information-theoretic metrics (i.e., conditional entropy) to order attributes and determine the directions of the edges and thus it needs the most training time against other algorithms.

As shown in Fig. 9(b), WATAN uses a weighted average learning method at classification time thus it needs time to calculate the weights of MWSTs. AODE-SR and IWAODE spend more time than WAODE-MI since AODE-SR checks all attribute-value pairs for generalization relationships and IWAODE needs to compute the weighting metrics for each testing instance during classification. TAN, SKDB and O^+BC are single BNCs, thus they need less time to compute the joint probability for classification.

5 Conclusion

The magnificent bloom of the machine learning area enhances the development of Bayesian learning. Bayesian network can not only qualitatively describe the implicit knowledge hidden in data in the form of DAG, but it can also quantitatively measure the fitness to data in the form of

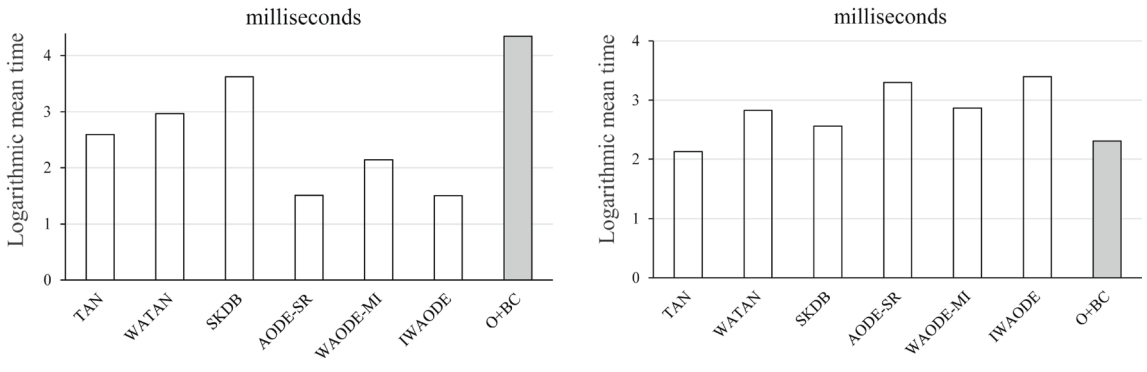


Fig. 9 Comparison of training and classification time for 7 algorithms on 44 datasets (milliseconds)

factorized joint probability. In this paper, we prove theoretically the reasonableness and feasibility of the application of entropy function as the scoring function, and the experimental results demonstrate the efficiency and effectiveness of the greedy search strategy. The proposed algorithm O^+BC allows to build high-dependence topology with a limited number of dependency relationships, and it delivers lower ZOL and bias than single learners (e.g., TAN, SKDB) and ensemble learners (e.g., WATAN, AODE-SR, WAODE-MI and IWAODE) at the cost of relatively high variance. The potential directions for future research include the study on how to scale up O^+BC to represent complex multivariate distributions and learn instantiated O^+BC to represent distinct data characteristics hidden in different instances.

Appendix

Table 7 The experiment results of ZOL

Datasets \ Learners	TAN	WATAN	SKDB	AODE-SR	WAODE-MI	IWAODE	O^+BC
Labor	0.0526	0.0526	0.0702	0.0526	0.0526	0.0526	0.0175
Zoo	0.0099	0.0198	0.0396	0.0297	0.0297	0.0198	0.0198
Promoters	0.1321	0.1132	0.3491	0.1321	0.1415	0.1038	0.1415
Iris	0.0800	0.0800	0.0867	0.0867	0.0867	0.0867	0.0867
Teaching-ae	0.5497	0.5364	0.5232	0.4768	0.4503	0.4570	0.4503
Hepatitis	0.1677	0.1742	0.2194	0.1871	0.1806	0.1742	0.1742
Wine	0.0337	0.0337	0.0674	0.0225	0.0169	0.0169	0.0225
Autos	0.2146	0.2146	0.1951	0.2049	0.1951	0.2098	0.1951
Sonar	0.2212	0.2212	0.2740	0.2212	0.2260	0.2260	0.2019
Glass-id	0.2196	0.2196	0.2103	0.2570	0.2570	0.2196	0.2056
Audio	0.2920	0.3009	0.3097	0.2080	0.1903	0.2301	0.2345
Hungarian	0.1701	0.1735	0.2007	0.1633	0.1565	0.1599	0.1667
Heart-disease-c	0.2079	0.2046	0.2211	0.1980	0.2013	0.1980	0.2046
Soybean-large	0.1107	0.1010	0.1140	0.0814	0.0814	0.0912	0.0749
Primary-tumor	0.5428	0.5428	0.5841	0.5723	0.5752	0.5457	0.5605
House-votes-84	0.0552	0.0529	0.0506	0.0529	0.0506	0.0483	0.0506
Cylinder-bands	0.2833	0.2463	0.2278	0.1852	0.1796	0.1926	0.1648
Chess	0.0926	0.0926	0.0762	0.1016	0.0944	0.1034	0.0889
Balance-scale	0.2736	0.2736	0.2912	0.2832	0.2816	0.2832	0.2816
Crx	0.1478	0.1478	0.1696	0.1377	0.1377	0.1319	0.1377
Breast-cancer-w	0.0415	0.0415	0.0658	0.0358	0.0358	0.0372	0.0343
Pima-ind-diabetes	0.2383	0.2370	0.2500	0.2370	0.2383	0.2370	0.2435
Vehicle	0.2943	0.2943	0.2920	0.2884	0.2872	0.2896	0.3002
Tic-tac-toe	0.2286	0.2265	0.1806	0.2651	0.2724	0.2662	0.1524
Led	0.2660	0.2660	0.2730	0.2680	0.2680	0.2700	0.2680
Diabetic-RD	0.3588	0.3588	0.3658	0.3571	0.3666	0.3640	0.3553
Contraceptive-mc	0.4888	0.4895	0.5363	0.4942	0.4922	0.4942	0.4895
Yeast	0.4171	0.4171	0.4461	0.4205	0.4232	0.4232	0.4286
Volcanoes	0.3316	0.3316	0.3316	0.3316	0.3316	0.3316	0.3316
Car	0.0567	0.0567	0.0556	0.0816	0.0885	0.0851	0.0509
Mfeat-mor	0.2970	0.2980	0.3140	0.3135	0.3130	0.3120	0.2975
Seismic-bumps	0.0720	0.0720	0.0689	0.0778	0.0774	0.0855	0.0724
Hypo	0.0141	0.0130	0.0175	0.0095	0.0101	0.0114	0.0080
Abalone	0.4587	0.4582	0.4680	0.4475	0.4475	0.4482	0.4599
Electrical-Grid	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002

Table 7 continued

Datasets \ Learners	TAN	WATAN	SKDB	AODE-SR	WAODE-MI	IWAODE	O^+BC
Firm-Teacher	0.1933	0.1934	0.1779	0.2126	0.2103	0.2124	0.1947
Nursery	0.0654	0.0654	0.0291	0.0730	0.0708	0.0735	0.0560
Magic	0.1675	0.1674	0.1718	0.1751	0.1762	0.1744	0.1734
Adult	0.1380	0.1380	0.1532	0.1410	0.1445	0.1502	0.1400
Activity-recognition-with	0.0178	0.0179	0.0183	0.0181	0.0176	0.0177	0.0168
Waveform	0.0202	0.0202	0.0285	0.0181	0.0181	0.0181	0.0187
Localization	0.3575	0.3575	0.3013	0.3596	0.3566	0.3593	0.3354
Skin-Segmentation	0.0030	0.0029	0.0024	0.0038	0.0039	0.0039	0.0029
Donation	0.0000	0.0000	0.0000	0.0002	0.0002	0.0002	0.0000

The cells in dark gray and in light gray indicate the performance of the corresponding model is better or equally well compared with other models on the corresponding dataset

Table 8 The experiment results of bias

Datasets \ Learners	TAN	WATAN	SKDB	AODE-SR	WAODE-MI	IWAODE	O^+BC
Labor	0.0211	0.0142	0.0316	0.0347	0.0200	0.0205	0.0242
Zoo	0.0303	0.0270	0.0585	0.0273	0.0273	0.0282	0.0358
Promoters	0.1329	0.1329	0.1251	0.4777	0.5489	0.2840	0.1371
Iris	0.0638	0.0618	0.0560	0.0586	0.0656	0.0664	0.0550
Teaching-ae	0.4566	0.4990	0.4598	0.4370	0.3984	0.4616	0.3940
Hepatitis	0.1712	0.1684	0.1727	0.1649	0.1655	0.1749	0.1627
Wine	0.0507	0.0531	0.0569	0.0346	0.0381	0.0317	0.0425
Autos	0.2356	0.2269	0.2265	0.2165	0.2115	0.2034	0.2113
Sonar	0.1646	0.1646	0.1675	0.1696	0.1722	0.1694	0.1823
Glass-id	0.2756	0.2748	0.2706	0.2785	0.2780	0.2818	0.2738
Audio	0.3617	0.3228	0.3095	0.1753	0.1799	0.2740	0.2635
Hungarian	0.1424	0.1491	0.1592	0.1582	0.1611	0.1597	0.1363
Heart-disease-c	0.1263	0.1265	0.1326	0.1118	0.1092	0.1160	0.1040
Soybean-large	0.1422	0.1151	0.1137	0.0648	0.0655	0.0811	0.0676
Primary-tumor	0.4249	0.4224	0.4413	0.4281	0.4247	0.4188	0.4112
House-votes-84	0.0410	0.0393	0.0304	0.0430	0.0406	0.0493	0.0421
Cylinder-bands	0.3117	0.2193	0.1942	0.1472	0.1501	0.1711	0.1488
Chess	0.1437	0.1398	0.1229	0.1244	0.1286	0.1397	0.1155
Balance-scale	0.1843	0.1843	0.1924	0.1905	0.1827	0.1905	0.1902
Crx	0.1180	0.1148	0.1234	0.0980	0.0953	0.0904	0.0939
Breast-cancer-w	0.0384	0.0349	0.0302	0.0338	0.0327	0.0234	0.0209
Pima-ind-diabetes	0.1946	0.1946	0.1963	0.1935	0.1941	0.1952	0.1952
Vehicle	0.2382	0.2376	0.2568	0.2401	0.2398	0.2435	0.2446
Tic-tac-toe	0.1746	0.1742	0.1266	0.2005	0.2104	0.1994	0.0967
Led	0.2251	0.2242	0.2295	0.2308	0.2331	0.2327	0.2298
Diabetic-RD	0.3206	0.3206	0.3214	0.3261	0.3416	0.3325	0.3217
Contraceptive-mc	0.3425	0.3426	0.3552	0.3811	0.3766	0.3781	0.3552
Yeast	0.3481	0.3479	0.3459	0.3455	0.3453	0.3458	0.3472
Volcanoes	0.2973	0.2973	0.2973	0.2973	0.2973	0.2973	0.2973
Car	0.0478	0.0478	0.0494	0.0556	0.0633	0.0599	0.0405
Mfeat-mor	0.2077	0.2078	0.2061	0.2475	0.2464	0.2492	0.2148
Seismic-bumps	0.0664	0.0664	0.0675	0.0639	0.0662	0.0646	0.0677
Hypo	0.0124	0.0119	0.0089	0.0069	0.0078	0.0080	0.0056
Abalone	0.3126	0.3123	0.3016	0.3199	0.3212	0.3199	0.3119
Electrical-Grid	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Firm-Teacher	0.1519	0.1520	0.1247	0.1897	0.1864	0.1897	0.1600
Nursery	0.0521	0.0522	0.0397	0.0651	0.0616	0.0658	0.0490
Magic	0.1252	0.1252	0.1244	0.1534	0.1541	0.1595	0.1294
Adult	0.1312	0.1312	0.1193	0.1378	0.1387	0.1437	0.1332
Activity-recognition-with	0.0179	0.0178	0.0131	0.0195	0.0192	0.0197	0.0142
Waveform	0.0152	0.0153	0.0182	0.0157	0.0158	0.0157	0.0157
Localization	0.3106	0.3105	0.2004	0.3129	0.3068	0.3126	0.2740
Skin-Segmentation	0.0023	0.0023	0.0020	0.0029	0.0031	0.0029	0.0023
Donation	0.0000	0.0000	0.0000	0.0002	0.0002	0.0002	0.0000

The cells in dark gray and in light gray indicate the performance of the corresponding model is better or equally well compared with other models on the corresponding dataset

Table 9 The experiment results of variance

Datasets \ Learners	TAN	WATAN	SKDB	AODE-SR	WAODE-MI	IWAODE	O^+BC
Labor	0.0211	0.0142	0.0316	0.0179	0.0221	0.0268	0.0389
Zoo	0.0303	0.0270	0.0585	0.0424	0.0424	0.0445	0.0461
Promoters	0.1329	0.1329	0.1251	0.0994	0.0654	0.1389	0.1657
Iris	0.0638	0.0618	0.0560	0.0374	0.0364	0.0436	0.0430
Teaching-ae	0.4566	0.4990	0.4598	0.1650	0.1776	0.1564	0.1780
Hepatitis	0.1712	0.1684	0.1727	0.0527	0.0541	0.0486	0.0529
Wine	0.0507	0.0531	0.0569	0.0231	0.0246	0.0141	0.0371
Autos	0.2356	0.2269	0.2265	0.1541	0.1503	0.1363	0.1549
Sonar	0.1646	0.1646	0.1675	0.0942	0.1003	0.0929	0.1075
Glass-id	0.2756	0.2748	0.2706	0.1004	0.1051	0.0999	0.1023
Audio	0.3617	0.3228	0.3095	0.1407	0.1401	0.0993	0.1245
Hungarian	0.1424	0.1491	0.1592	0.0255	0.0317	0.0270	0.0473
Heart-disease-c	0.1263	0.1265	0.1326	0.0357	0.0383	0.0305	0.0485
Soybean-large	0.1422	0.1151	0.1137	0.0842	0.0855	0.0738	0.0922
Primary-tumor	0.4249	0.4224	0.4413	0.1826	0.1859	0.1785	0.2189
House-votes-84	0.0410	0.0393	0.0304	0.0094	0.0083	0.0079	0.0186
Cylinder-bands	0.3117	0.2193	0.1942	0.1067	0.1010	0.0828	0.1090
Chess	0.1437	0.1398	0.1229	0.0422	0.0364	0.0379	0.0534
Balance-scale	0.1843	0.1843	0.1924	0.0854	0.0913	0.0854	0.0872
Crx	0.1180	0.1148	0.1234	0.0268	0.0264	0.0240	0.0409
Breast-cancer-w	0.0384	0.0349	0.0302	0.0134	0.0128	0.0122	0.0233
Pima-ind-diabetes	0.1946	0.1946	0.1963	0.0698	0.0700	0.0697	0.0622
Vehicle	0.2382	0.2376	0.2568	0.1287	0.1276	0.1245	0.1420
Tic-tac-toe	0.1746	0.1742	0.1266	0.0513	0.0604	0.0529	0.1105
Led	0.2251	0.2242	0.2295	0.0410	0.0398	0.0372	0.0530
Diabetic-RD	0.3206	0.3206	0.3214	0.0572	0.0522	0.0560	0.0576
Contraceptive-mc	0.3425	0.3426	0.3552	0.1077	0.1106	0.1086	0.1544
Yeast	0.3481	0.3479	0.3459	0.1001	0.0970	0.0967	0.1019
Volcanoes	0.2973	0.2973	0.2973	0.0052	0.0052	0.0052	0.0052
Car	0.0478	0.0478	0.0494	0.0438	0.0427	0.0430	0.0456
Mfeat-mor	0.2077	0.2078	0.2061	0.0679	0.0686	0.0676	0.0973
Seismic-bumps	0.0664	0.0664	0.0675	0.0112	0.0075	0.0128	0.0025
Hypo	0.0124	0.0119	0.0089	0.0047	0.0056	0.0068	0.0044
Abalone	0.3126	0.3123	0.3016	0.1543	0.1543	0.1539	0.1701
Electrical-Grid	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0001
Firm-Teacher	0.1519	0.1520	0.1247	0.0179	0.0188	0.0179	0.0433
Nursery	0.0521	0.0522	0.0397	0.0105	0.0111	0.0104	0.0140
Magic	0.1252	0.1252	0.1244	0.0291	0.0289	0.0291	0.0486
Adult	0.1312	0.1312	0.1193	0.0111	0.0113	0.0109	0.0153
Activity-recognition-with	0.0179	0.0178	0.0131	0.0051	0.0055	0.0050	0.0073
Waveform	0.0152	0.0153	0.0182	0.0025	0.0023	0.0024	0.0031
Localization	0.3106	0.3105	0.2004	0.0580	0.0632	0.0577	0.0770
Skin-Segmentation	0.0023	0.0023	0.0020	0.0014	0.0014	0.0015	0.0020
Donation	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000

The cells in dark gray and in light gray indicate the performance of the corresponding model is better or equally well compared with other models on the corresponding dataset

Acknowledgements This work is supported by the National Key Research and Development Program of China (No.2019YFC1804804), Open Research Project of The Hubei Key Laboratory of Intelligent Geo-Information Processing (No.KLIGIP-2021A04), and the Scientific and Technological Developing Scheme of Jilin Province (No.20200201281JC).

Author Contributions **Lanni Wang:** Conceptualization, Validation, Visualization, Writing - original draft. **Limin Wang:** Methodology, Supervision, Writing - review & editing, Funding acquisition. **Lu Guo:** Formal analysis, Project administration. **Qilong Li:** Software, Investigation. **Xiongfei Li:** Writing - review & editing, Validation.

Data Availability The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Declarations

Competing Interest The authors declare that they have no conflict of interest.

Ethical and Informed Consent for Data Used This study does not contain any studies with human participants or animals performed by any of the authors.

References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. Artif Intell 17(3):37–37
2. Yukselturk E, Ozekes S, Türel YK (2014) Predicting dropout student: an application of data mining methods in an online education program. Eur J Open, Dist E-learning 17(1):118–133
3. Wang L, Xie Y, Pang M, Wei J (2022) Alleviating the attribute conditional independence and I.I.D. assumptions of averaged one-dependence estimator by double weighting. Knowl-Based Syst 250:109078
4. Wu H, Yan G, Xu D (2014) Developing vehicular data cloud services in the IoTenvironment. IEEE Trans Ind Inform 10(2):1587–1595
5. Peña-Ayala A (2014) Educational data mining: a survey and a data mining-based analysis of recent works. Expert Syst Appl 41(4):1432–1462
6. Jiang L, Zhang L, Yu L, Wang D (2019) Class-specific attribute weighted naive Bayes. Pattern Recognit 88:321–330
7. Ren Y, Wang L, Li X, Peng M, Wei J (2022) Stochastic optimization for bayesian network classifiers. Appl Intell 52(13):15496–15516
8. Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of Plausible Inference. Morgan Kaufmann
9. Wang L, Zhang S, Mammadov M, Li K, Zhang X (2021) Semi-supervised weighting for averaged one-dependence estimators. Appl Intell 52(4):4057–4073
10. Zhang H, Petitjean F, Buntine W (2020) Bayesian network classifiers using ensembles and smoothing. Knowl Inf Syst 62(9):3457–3480
11. Jiang L, Li C, Wang S, Zhang L (2016) Deep feature weighting for naive Bayes and its application to text classification. Eng Appl Artif Intell 52:26–39
12. Kong H, Shi X, Wang L (2021) Averaged tree-augmented onedependence estimators. Appl Intell 51(7):4270–4286
13. Zhang H, Jiang L, Li C (2022) Attribute augmented and weighted naive Bayes. Sci China Inf Sci 65(12):222101
14. Chickering DM (1996) Learning Bayesian networks is NP-complete. Learn Data: Artif Intell Stat V:121-130
15. Wang L, Zhou J, Wei J, Pang M, Sun M (2022) Learning causal Bayesian networks based on causality analysis for classification. Eng Appl Artif Intell 114:105212
16. Jiang L, Zhang H, Cai Z (2008) A novel Bayes model: hidden naive bayes. IEEE Trans Knowl Data Eng 21(10):1361–1371
17. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. Mach Learn 29(2):131–163
18. Cover TM, Thomas JA (2006) Elements of Information Theory. Wiley-Interscience
19. Martínez AM, Webb GI, Chen S, Zaidi NA (2012) Scalable learning of Bayesian network classifiers. J Mach Learn Res 17(1):1515–1549
20. Zhao X, Yan H, Hu Z, Du D (2022) Deep spatio-temporal sparse decomposition for trend prediction and anomaly detection in cardiac electrical conduction. IISE Trans Healthc Syst Eng 12(2):150–164
21. Jiang L, Zhang L, Li C, Wu J (2019) A correlation-based feature weighting filter for naive Bayes. IEEE Trans Knowl Data Eng 31:201–213
22. Langley P, Iba W, Thompson K (1992) An analysis of Bayesian classifiers. In: Proceedings of AAAI conference on artificial intelligence, pp 223–228
23. Pang Y, Zhao X, Hu J, Yan H, Liu Y (2022) Bayesian spatio-temporal graph transformer network (b-star) for multi-aircraft trajectory prediction. Knowl-Based Syst 249:108998
24. Domingos P, Pazzani M (1997) On the optimality of the simple Bayesian classifier under zero-one loss. Mach Learn 29(2):103–130
25. Kononenko I (1991) Semi-naive Bayesian classifier. In: Machine learning-EWSL-91: European working session on learning porto, pp 206–219
26. Jiang L, Cai Z, Wang D, Zhang H (2012) Improving Tree augmented Naive Bayes for class probability estimation. Knowl-Based Syst 26:239–245
27. Peng F, Schuurmans D, Wang S (2004) Augmenting naive bayes classifiers with statistical language models. Inf Retr 7(3–4):317–345
28. Sahami M (1996) Learning limited dependence Bayesian classifiers. In: Proceedings of the second international conference on knowledge discovery and data mining, pp 335–338
29. Shannon CE (2001) A mathematical theory of communication. ACM SIGMOBILE Mob Comput Commun Rev 5(1):3–55
30. Wang L, Zhang X, Li K, Zhang S (2022) Semi-supervised learning for k-dependence Bayesian classifiers. Appl Intell 52(4):3604–3622
31. Jiang L, Zhang H, Cai Z, Wang D (2012) Weighted average of one-dependence estimators. J Exp Theor Artif Intell 24(2):219–230
32. Duan Z, Wang L, Chen S, Sun M (2020) Instance-based weighting filter for superparent one-dependence estimators. Knowl-Based Syst 203:106085
33. Akaike H (1974) A New Look at the Statistical Model Identification. IEEE Trans Autom Control 19:716–723
34. Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461–465
35. Suzuki J (1999) Learning Bayesian belief networks based on the MDL principle: an efficient algorithm using the branch and bound technique. IEICE Trans Inf Syst 82(2):356–367
36. Korb KB, Nicholson AE (2004) Bayesian artificial intelligence. Chapman and Hall
37. Kong H, Wang L (2023) Flexible model weighting for one-dependence estimators based on point-wise independence analysis. Pattern Recognit 139:109473

38. Sun H (2020) Simultaneous material microstructure classification and discovery using acoustic emission signals. Arizona State University
39. Heckerman D (1998) A tutorial on learning Bayesian networks. Springer, Netherlands
40. Liu Y, Wang L, Mammadov M, Chen S, Wang G, Qi S, Sun M (2021) Hierarchical independence thresholding for learning Bayesian network classifiers. Knowl-Based Syst 212:106627
41. Zhao X, Yan H, Liu Y (2021) Hierarchical tree-based sequential event prediction with application in the aviation accident report. In: 2021 IEEE 37th international conference on data engineering (ICDE), pp 1925–1930
42. Wang L, Chen S, Mammadov M (2018) Target learning: a novel framework to mine significant dependencies for unlabeled data. In: Proceedings of the 22nd Pacific-Asia conference on knowledge discovery and data mining, pp 06–117
43. Pang Y, Zhao X, Yan H, Liu Y (2021) Data-driven trajectory prediction with weather uncertainties: a Bayesian deep learning approach. Transp Res C: Emerg Technol 130:103326
44. Bache K, Lichman M, UCI machine learning repository, Available online: https://archive.ics.uci.edu/ml/datasets.html
45. Fayyad U, Irani K (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Proceedings of the 13th international joint conference on artificial intelligence, pp 1022–1029
46. Zheng F, Webb GI, Suraweera P, Zhu L (2012) Subsumption resolution: an efficient and effective technique for semi-naive Bayesian learning. Mach Learn 87(1):93–125
47. Fisher RA (1970) Statistical methods for research workers. Breakthroughs in statistics: Methodology and distribution 66–70
48. Kohavi R, Wolpert DH (1996) Bias plus variance decomposition for zero-one loss functions. In: Proceedings of the 13th international conference on machine learning, pp 275–283
49. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32(200):675–701
50. Nemenyi PB (1963) Distribution-free multiple comparisons, Princeton University
51. Demšar J (2006) Statistical comparisons of classifiers over multiple datasets. J Mach Learn Res 7:1–30

Lanni Wang received the B.Sc. degree from Jilin University, China, in 2021 and she is currently a postgraduate student in the College of Computer Science and Technology, Jilin University, China. Her research interests include Bayesian network and data analysis.

Limin Wang received the Ph.D. degree in computer science from Jilin University, China, in 2005. He is currently a professor in the College of Computer Science and Technology, Jilin University, China. He has authored or co-authored more than 60 academic articles in reputed peer-reviewed international journals and conferences. His research interests include machine learning, data mining, decision making and Bayesian network. He has supervised many M.Sc. and Ph.D. students in the above-mentioned fields. He has also been involved with reviewing and organizing different workshops, seminars, and training sessions on different technologies.

Lu Guo received the B.E. degree from Jinan University, China, in 2021 and she is currently a postgraduate student in the College of Software, Jilin University, China. Her research interests include Bayesian network and data analysis.

Qilong Li is currently a postgraduate student in the College of Instrumentation and Electrical Engineering, Jilin University, China. His research interests include Bayesian networks and data mining.

Xiongfei Li (Member, IEEE) received the B.S. degree in computer software from Nanjing University, in 1985, the M.S. degree in computer software from the Chinese Academy of Sciences, in 1988, and the Ph.D. degree in communication and information system from Jilin University, in 2002. Since 1988, he has been a member of the faculty of the Computer Science and Technology, Jilin University, Changchun, China. He is currently a Professor of computer software and theory with Jilin University. He has authored more than 60 research articles. His research interests include data mining, intelligent network, image processing, and analysis.