PLOS ONE

RESEARCH ARTICLE

Using *k*-dependence causal forest to mine the most significant dependency relationships among clinical variables for thyroid disease diagnosis

LiMin Wang[1], FangYuan Cao[1], ShuangCheng Wang[2], MingHui Sun[1]*, LiYan Dong[1]*

1 Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, ChangChun City 130012, China, **2** Lixin Accounting Research Institute, Shanghai Lixin University of Commerce, Shanghai City 201620, China

* smh@jlu.edu.cn (MHS); dongly@jlu.edu.cn (LYD)

Citation: Wang L, Cao F, Wang S, Sun M, Dong L (2017) Using *k*-dependence causal forest to mine the most significant dependency relationships among clinical variables for thyroid disease diagnosis. PLoS ONE 12(8): e0182070. https://doi.org/10.1371/journal.pone.0182070

Editor: Vladimir B. Bajic, King Abdullah University of Science and Technology, SAUDI ARABIA

Received: November 3, 2016

Accepted: July 12, 2017

Published: August 17, 2017

Data Availability Statement: We used 12 data sets for experimental study. The URLs for these data sets are different and are listed below, Echocardiogram——————https://archive.ics.uci.edu/ml/datasets/Echocardiogram Heart——————https://archive.ics.uci.edu/ml/datasets/Statlog+%28Heart%29 Heart Disease—————— https://archive.ics.uci.edu/ml/datasets/Heart+Disease Chess—————— http://archive.ics.uci.edu/ml/datasets/Chess+%28King-Rook+vs.+King-Pawn%29 Breast —————— https://

Abstract

Numerous data mining models have been proposed to construct computer-aided medical expert systems. Bayesian network classifiers (BNCs) are more distinct and understandable than other models. To graphically describe the dependency relationships among clinical variables for thyroid disease diagnosis and ensure the rationality of the diagnosis results, the proposed *k*-dependence causal forest (KCF) model generates a series of submodels in the framework of maximum spanning tree (MST) and demonstrates stronger dependence representation. Friedman test on 12 UCI datasets shows that KCF has classification accuracy advantage over the other state-of-the-art BNCs, such as Naive Bayes, tree augmented Naive Bayes, and *k*-dependence Bayesian classifier. Our extensive experimental comparison on 4 medical datasets also proves the feasibility and effectiveness of KCF in terms of sensitivity and specificity.

Background

Data mining [1] [2] is used to extract unknown but potentially useful information by using available incomplete, noisy, fuzzy, and random practical application data. The medical domain consists of a considerable amount of data, including complete human genetic code information; clinical information on the history of patients, diagnosis, inspection, and treatment; and drug management information. Data mining can be applied in the medical field to analyze medical data, extract implicit valuable information, provide correct diagnosis and treatment, and study the genetic law of human diseases and health [3].

While dealing with a large amount of historical information of patients in the database, data mining needs to confirm the diagnosis based on age, gender, auxiliary examination results, and physiological and biochemical indicators of patients. Thus, data mining should eliminate interference of human factors and establish diagnosis rules with good universality, provided that large amounts of data are analyzed in the process. Consequently, researchers can

archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29 Pima Indians Diabetes——————— https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes Tic-Tac-Toe——————— https://archive.ics.uci.edu/ml/datasets/Tic-Tac-Toe+Endgame German——————— https://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29 Spambase——————— https://archive.ics.uci.edu/ml/datasets/Spambase Mushroom——————— https://archive.ics.uci.edu/ml/datasets/Mushroom Adult ——————— https://archive.ics.uci.edu/ml/datasets/Adult Census Income——————— https://archive.ics.uci.edu/ml/datasets/Census+Income.

Funding: This work was supported by the National Science Foundation of China (Grant No. 61272209) and the Agreement of Science & Technology Development Project, Jilin Province (No. 20150101014JC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

establish a prediction model, test it, and construct an accurate algorithmic model, which can be used for diagnosis of clinical medical conditions.

Now, about 20 million Americans have some form of thyroid disease, and people of all ages and races can have the chance to get thyroid disease [4]. Recently, a fair mount of data mining methods have been investigated to diagnose this kind of disease. To explore the value of contrast-enhanced ultrasound combined with conventional ultrasound in the diagnosis of thyroid microcarcinoma, multivariate logistic regression analysis is performed to determine independent risk factors [5]. Proper interpretation of the thyroid data besides clinical examination and complementary investigation is an important issue, a comparative study of thyroid disease diagnosis is made by using three different types of neural networks, i.e. multilayer neural network, probabilistic neural network and learning vector quantization neural network [6]. An enhanced fuzzy k-nearest neighbor (FKNN) classifier based computer aided diagnostic system is presented for thyroid disease [4]. The neighborhood size k and the fuzzy strength parameter m in FKNN classifier are adaptively specified by the particle swarm optimization approach. The application of Support Vector Machines is proposed to classify thyroid bioptic specimens [7], together with a particular wrapper feature selection algorithm (i.e., recursive feature elimination). The model is able to provide an accurate discriminatory capability using only 20 out of 144 features, resulting in an increase of the model performances, reliability, and computational efficiency. To elucidate the cytological characteristics and the diagnostic usefulness of intraoperative cytology for papillary thyroid carcinoma, decision tree analysis is used to find effective features for accurate cytological diagnosis [8].

Bayesian method is an intelligent computing method used in reasoning and managing uncertainty problems [9]. BNC is a probability network based on graphical models used to provide probabilistic inference, thus it is more distinct and understandable than other methods. A BNC consists of a structural model and a set of conditional probabilities. The structural model is a directed acyclic graph, in which nodes represent classes C and a set of random attributes $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. Arcs between nodes are used to describe the conditional dependence relationships, which are quantified using conditional probabilities for each node given to the parents. Bayesian methods have gained increasing interest in medical diagnosis. BN and graph theory are used to encode causal relations among variables for diagnosis and predictions in the medical domain [10–12].

The Markov blanket of a target attribute is the minimal attribute set for explaining the target attribute based on the conditional independence of all the attributes to be connected in a BN [13]. Koller and Sahami [14] defined the Markov blanket of a target attribute as the minimal set of conditioned attributes, in which all other attributes are independent of the target attribute in the probabilistic graphical model. Hence, the Markov blanket of a target attribute removes unnecessary attributes and represents the minimal information for explaining the target attribute. In a BN model, the Markov blanket of T, i.e., MB(T) is the union of parent, child, and parent of children nodes of T [13, 15]. For example, in Fig 1, the parent nodes of T are B and C, the child node of T is F, and the parent of the children node of T is E. Thus, the Markov blanket of T is MB(T) = {B, C, F, E}, indicating that nodes A, D, and G are independent of T conditioned on MB(T).

The performance of a classifier is evaluated using two key factors, namely, classification accuracy and space complexity of a model. A BN cannot express all relationships between the attributes and the class. Thus, a trade-off should exist between the structure complexity and classification accuracy. Some restricted Bayesian classifiers, e.g., Naive Bayes (NB), tree augmented Naive Bayes (TAN), and k-dependence BNs (KDB), exhibit satisfactory performance for classification at different levels of conditional independence assumption. When carrying out medical analysis, different doctors may consider different factor or attribute as starting

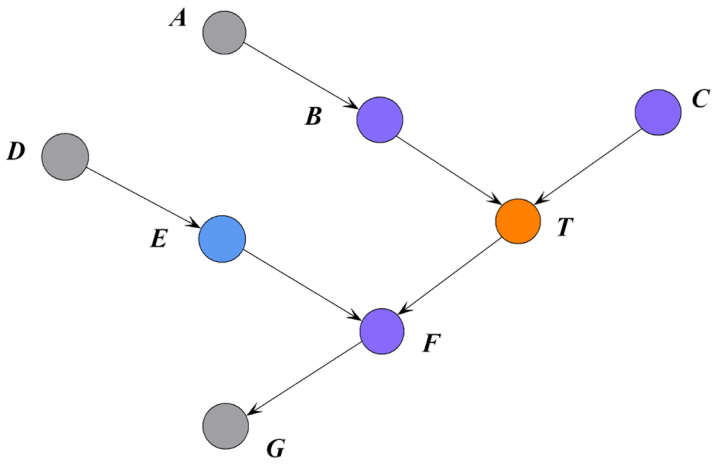

Fig 1. An example Markov blanket.

https://doi.org/10.1371/journal.pone.0182070.g001

point. One BNC is unable to express this diversity. This paper proposes a novel learning algorithm called the *k*-dependence causal forest (KCF). This algorithm generates a series of submodels, which are used to construct classifiers with different root nodes at arbitrary points (values of *k*) along the attribute dependence spectrum. The KCF algorithm aims to describe the significant dependency relationships between root node X_r and MB(X_r) while simultaneously providing accurate diagnosis to patients with thyroid diseases.

Materials and methods

Data

This research work adopts the public thyroid disease dataset from the University of California, Irvine (UCI) Machine Learning Repository [16]. The UCI database currently contains 335 datasets, and the number of sets continuously increases. The thyroid disease dataset was stored in the UCI by Ross Quinlan during his visit in 1987 for the 1987 Machine Learning Workshop; the set contains 9172 real historical instances. Each instance consists of 29 attributes, which can be classified into 20 classes. The characteristics of thyroid disease dataset are multivariate and domain theory, the characteristics of the contained attributes are categorical and real, and the associated task of the dataset is classification.

Three restricted Bayesian classifiers

BNs are often used to solve classification problems by constructing classifiers from a given set of training instances with class labels. With high classification accuracy and efficiency, BN classifiers perform outstandingly in a number of classification methods. This paper briefly introduces the three popular restricted Bayesian classifiers. In the following discussion, capital letters, such as *X*, *Y* and *Z*, denote attribute names, and lower-case letters, such as *x*, *y* and *z*, denote the specific values taken by those attributes. Sets of attributes are denoted by boldface capital letters, such as $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{Z}$, and assignments of values to the attributes in these sets are denoted by boldface lowercase letters, such as **x, y** and **z**.

The NB classifier is the simplest BN model and is very robust [17]. Given the *n* independent attributes $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ and *m* classes $c_1, c_2, \ldots, c_m$, classification will derive the

$I(X_i,X_j\|C)$	X_2	X_{17}	X_{19}	X_{21}	X_{23}
X_{17}	0.001				
X_{19}	0.001	0.142			
X_{21}	0.006	0.023	0.022		
X_{23}	0.015	0.093	0.059	0.059	
X_{25}	0.005	0.225	0.150	0.050	0.156

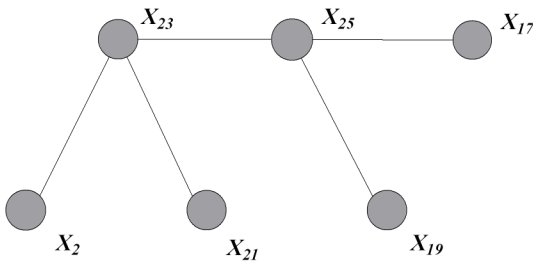

Fig 2. An example of conditional mutual information matrix (a) and corresponding undirected MST (b). Attributes {X_2, X_{17}, X_{19}, X_{21}, X_{23}, X_{25}} correspond to clinical variables *on thyroxine*, *TSH*, *T3*, *TT4*, *T4U* and *FTI*, respectively.

https://doi.org/10.1371/journal.pone.0182070.g002

maximum of $P(c_i|x)$, where $1 \leq i \leq m$. Result can be derived from the Bayesian theorem, as Eq (1) shows:

$$P(c_i|x) = \frac{P(c_i)P(x|c_i)}{P(x)} \tag{1}$$

The rigorous assumption in NB is that all attributes are conditionally independent of each other. Thus, the class assignments of the test samples are based on Eq (2).

$$\arg\max_{c_i} P(c_i)P(x|c_i) = \arg\max_{c_i} P(c_i)\prod_{j=1}^{n} P(x_j|c_i) \tag{2}$$

The basic framework of TAN [18] is the extension of the Chow-Liu tree [19], which utilizes conditional mutual information to build a maximum spanning tree (MST). TAN is a one-dependence classifier because it allows each attribute to have at most one parent in addition to the class. In practice, TAN is regarded as a good trade-off between the model complexity and classification performance. Fig 2 shows an example of the condition mutual information matrix with six attributes and corresponding undirected MST. The selected six attributes are the first few attributes with the maximum mutual information with class $I(X_i; C)$ in the thyroid disease dataset.

For a TAN model, the class assignments of the test samples are based on Eq (3).

$$\arg\max_{c_i} P(c_i)P(x|c_i) = \arg\max_{c_i} P(c_i)\prod_{j=1}^{n} P(x_j|c_i, x_{jp}) \tag{3}$$

where X_{jp} is the parent node of X_j.

After selecting each attribute as the root node and setting the outward direction of all the arcs from the attributes, six different directed MSTs are generated, as shown in Fig 3. The root node is filled in black. The directed MSTs can be regarded as different representations of the same spectrum of causal relationships under different conditions. One MST corresponds to n directed trees, and each tree uses different attributes as the root node. Although TAN can achieve a global one-dependence optimization, MST cannot be extended to arbitrary k-dependence structure when $k > 1$.

The KDB [20] is a k-dependence classifier because it allows each attribute to have a maximum number of k parents in addition to the class attribute. Starting with the highest, an attribute order is pre-determined by comparing the mutual information $I(X_i; C)$. By comparing

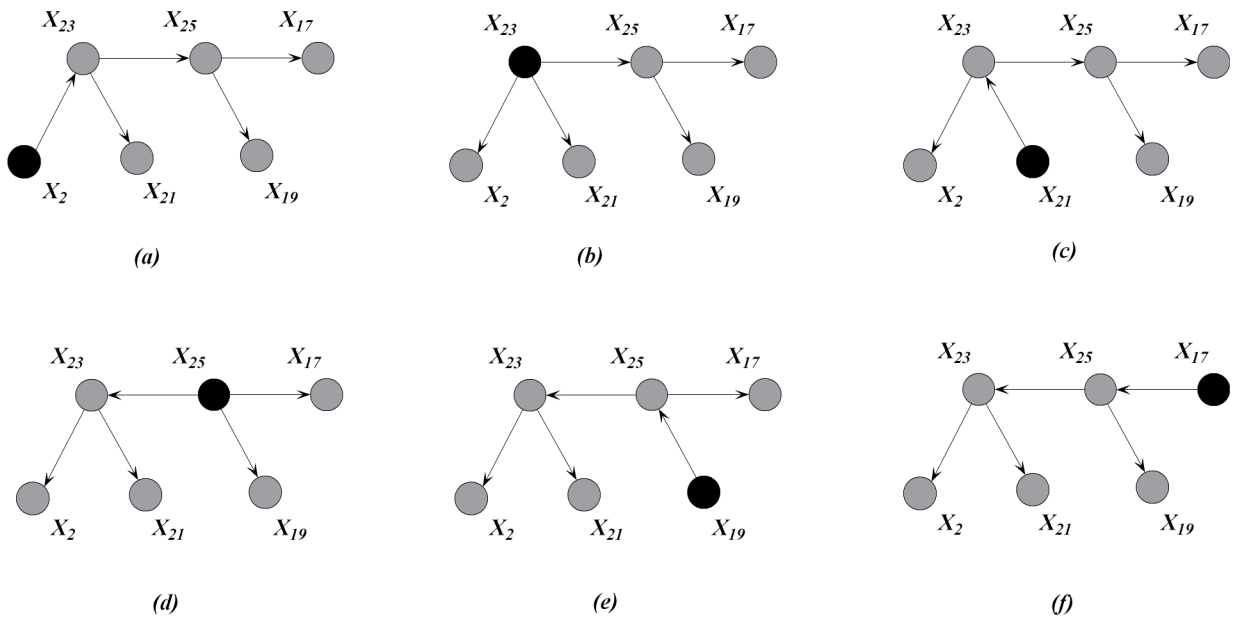

Fig 3. An example of directed MSTs with different root nodes, which are filled in black. Attributes {X_2, X_{17}, X_{19}, X_{21}, X_{23}, X_{25}} correspond to clinical variables *on thyroxine*, *TSH*, *T3*, *TT4*, *T4U* and *FTI*, respectively.

https://doi.org/10.1371/journal.pone.0182070.g003

conditional mutual information $I(X_i; X_j|C)$, each attribute can select a maximum number of k parents among the attributes ahead of itself in the pre-determined order.

For a KDB model, the class assignments of the test samples are based on Eq (4).

$$P(X|C_i) = \prod_{k=1}^{n} P(X_k|C_i, X_{i1}, \cdots, X_{ip}) \arg\max_{c_i} P(c_i)P(x|c_i) = \arg\max_{c_i} P(c_i) \prod_{j=1}^{n} P(x_j|c_i, x_{j1}, \cdots, x_{jp}) \quad (4)$$

where $\{X_{j1}, \cdots, X_{jp}\}$ are the parent attributes of X_j and $p = min(j - 1, k)$.

KCF algorithm

MST contains the most significant relationships among attributes. Thus at training time, we aim to achieve high-dependence directed trees by extending one-dependence directed trees that are inferred from MST. Each one-dependence directed tree is extended to the k-dependence conditional tree along the attribute dependence spectrum. Finally, we will obtain a series of k-dependence trees rather than one augmented tree. Leaf node X_i can be used to select other nodes as parents along the path from X_i to the root node by comparing the conditional mutual information. For example, as shown in Fig 3(a), X_2, X_{23}, X_{25} are the possible parents of X_{17}, and X_2, X_{23} are the possible parents of X_{25}. Different root nodes correspond to different spanning trees or Bayesian classifiers, the ensemble of which finally forms a forest. When $k > 1$, e.g., $k = 2$, more parents can be selected for each non-root node by comparing the conditional mutual information. Fig 4 shows the k-dependence Bayesian classifiers when $k = 2$. The newly added arcs are annotated with red color.

At the testing time, KCF estimates the class membership probabilities by using each subclassifier, and the final result is the average of the outputs of all subclassifiers. The training procedure (KCF-Training) and testing procedure (KCF-Testing) are depicted below.

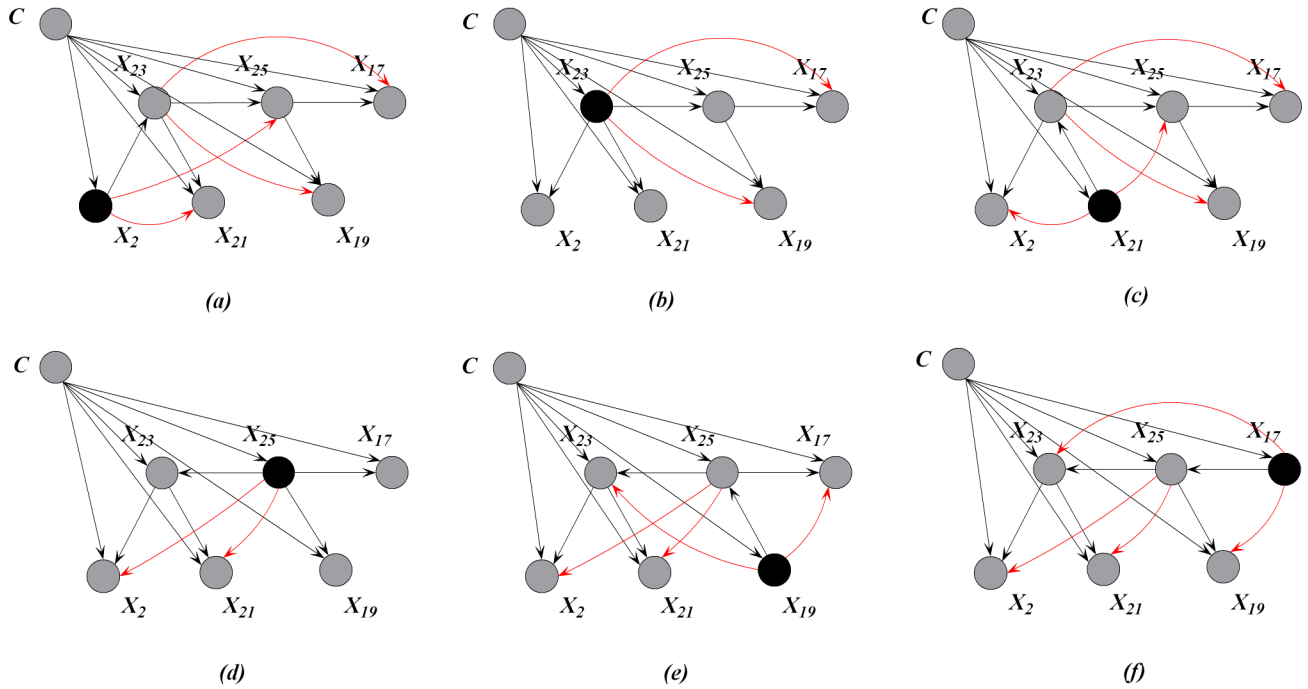

Fig 4. The KCF ($k = 2$) model corresponding to the MSTs shown in Fig 3. Attributes {X_2, X_{17}, X_{19}, X_{21}, X_{23}, X_{25}, C} correspond to clinical variables *on thyroid*, *TSH*, *T3*, *TT4*, *T4U*, *FTI* and *Class*, respectively.

https://doi.org/10.1371/journal.pone.0182070.g004

Algorithm 1 KCF-Training

Input: Pre-classified instance set DB with n predictive attributes $\{X_1, \cdots, X_n\}$.

Output: Subclassifiers $\{\mathrm{KCF}_1, \cdots, \mathrm{KCF}_n\}$.

1: Compute conditional mutual information $I(X_i; X_j|C)$ for each pair of attributes X_i and X_j, where $i \neq j$.

2: Build undirected MST by comparing conditional mutual information.

3: For each attribute X_i ($i = 1, 2, \ldots, n$)

(a) Transform the MST to be a directed one by choosing X_i as the root and setting the direction of all arcs to be outward from it.

(b) Let the Bayesian subclassifier being constructed, KCF_i, begin with the directed MST.

(c) Add a node to KCF_i representing class variable C.

(d) Add an arc from C to each node in KCF_i.

(e) For each node X_j ($j \neq i$), add $m - 1$ ($m = min(d, k)$, d is the number of nodes along the branch from root to X_j) arcs from $m - 1$ distinct attributes X_P to X_j. X_P should locate in the branch from root to X_j and correspond to the first $m - 1$ highest value for $I(X_P; X_j|C)$.

4: Compute the conditional probability tables inferred by the structure of KCF_i by using counts from DB, and output KCF_i.

Algorithm 2 KCF-Testing

Input: KCF_1, KCF_2, ..., KCF_n and a testing instance e.
Output: The conditional probabilities $\hat{P}(c|e)(p = 1, 2, \ldots, t)$, where c is the class label.
1: For each $KCF_{i\,(i=1,\,2,\,\ldots,\,n)}$, estimate the conditional probability $\hat{P}_i(c|e)$ that e belongs to class c.
2: Average all of the probabilities $\hat{P}(c|e) = \frac{1}{n}\sum_{i=1}^{n}\hat{P}_i(c|e)$.
3: Return the estimated $\hat{P}(c_1|e)$, $\hat{P}(c_2|e)$, ..., $\hat{P}(c_t|e)$.

k is related to the classification performance of a high-dependence classifier. An appropriate value of k cannot be effectively preselected to achieve the optimal trade-off between the model complexity and classification performance [21]. For each KCF_i, the space complexity increases exponentially as the value of k increases to achieve a trade-off between the classification performance and efficiency. We set $k = 2$ in the following experiments.

Results

The detailed introduction of the 29 attributes from thyroid disease dataset in UCI database is shown in Table 1. And numeric attributes in thyroid disease dataset are discretized by using 10-bin equal frequency discretization. In order to minimize the bias associated with the random sampling of the training and holdout data samples in comparing the classification accuracy of two or more methods, 10-fold cross-validation is applied to compare the general performance of KCF with three Bayesian network classifiers (i.e., NB, TAN and KDB) and five non-Bayesian network classifiers, i.e., IBK(k-Nearest Neighbours) [22], SMO(Support Vector Machine) [23], MultilayerPerception(Artificial Neural Network) [24], DecisionStump(Decision Tree) [25] and SimpleLogistic(linear logistic regression) [26]. In 10-fold cross-validation, whole data are randomly divided to 10 mutually exclusive and approximately equal size subsets. The classification algorithm trained and tested 10 times. In each case, one of the folds is taken as test data and the remaining folds are added to form training data. Thus 10 different test results exist for each training-test configuration. The average of these results gives the test accuracy of the algorithm. All the experiments have been carried out in a C++ software specially designed to deal with out-of-core classification methods. The average classification accuracy (inversely related to zero-one loss [27]) are 75.17%(NB), 80.65%(TAN), 80.43%(KDB), 81.89%(KCF), 78.15%(IBK), 79.67%(SMO), 77.34%(MultilayerPerception), 73.81%(DecisionStump) and 79.53%(SimpleLogistic). Obviously, the proposed KCF algorithm achieves the highest classification accuracy compared with other algorithms and thus performs much more effectively in thyroid disease diagnosis.

To explain the main reason of performance difference of BNCs, we will clarify from the viewpoint of Markov blanket. Compared with low-dependence BNC, high-dependence BNC can demonstrate more conditional dependencies. Thus in the following discussion, we just compare KCF with KDB, both of which are 2-dependence BNCs. KCF will generate a series of submodels, each of which corresponds to different focus for analysis. For example, if X_i is the key factor for diagnosis, then doctors can use the ith submodel for further analysis. From the definition of Markov blanket, we can get the following conclusion that X_i is directly and mutually dependent on attributes $\{Pa(X_i), Ch(X_i)\}$ while indirectly dependent on attributes $PC(X_i)$. The other attributes are useless for further consideration. The time cost for unnecessary analysis and expenditure on unnecessary physical examination will be decreased greatly. With limited time and space complexity, more Markov blanket attributes means more possible dependency relationships to be mined. The list and number of Markov blanket attributes of

Table 1. Attributes available for analysis.

Attribute	Type	Explanation	Corresponding symbol in Figs 2–8
age	Numeric	Years	X_0
sex	Binary	Female/male	X_1
on thyroxine	Binary	Yes/no	X_2
query on thyroxine	Binary	Yes/no	X_3
on antithyroid medication	Binary	Yes/no	X_4
sick	Binary	Yes/no	X_5
pregnant	Binary	Yes/no	X_6
thyroid surgery	Binary	Yes/no	X_7
I131 treatment	Binary	Yes/no	X_8
query hypothyroid	Binary	Yes/no	X_9
query hyperthyroid	Binary	Yes/no	X_{10}
lithium	Binary	Yes/no	X_{11}
goitre	Binary	Yes/no	X_{12}
tumor	Binary	Yes/no	X_{13}
hypopituitary	Binary	Yes/no	X_{14}
psych	Binary	Yes/no	X_{15}
TSH measured	Binary	Yes/no	X_{16}
TSH	Numeric	Thyroid stimulating hormone	X_{17}
T3 measured	Binary	Yes/no	X_{18}
T3	Numeric	Triiodothyronine	X_{19}
TT4 measured	Binary	Yes/no	X_{20}
TT4	Numeric	Total serum thyroxine	X_{21}
T4U measured	Binary	Yes/no	X_{22}
T4U	Numeric	thyroxine	X_{23}
FTI measured	Binary	Yes/no	X_{24}
FTI	Numeric	Free Tyroxine Index	X_{25}
TBG measured	Binary	Yes/no	X_{26}
TBG	Numeric	Thyroid binding globulin	X_{27}
referral source	Categorical	WEST, STMW, SVHC, SVI, SVHD, other	X_{28}
Category	Categorical	20 class labels are divided into 7 groups: Hyperthyroid conditions, Hypothyroid conditions, Binding protein, General health, Replacement therapy, Antithyroid treatment, Miscellaneous	C

https://doi.org/10.1371/journal.pone.0182070.t001

each attribute for KCF and KDB are shown in Fig 5 and Fig 6, respectively. From Fig 6, for 25 of all of the 29 attributes the number of corresponding Markov blanket attributes for KCF is greater than that for KDB. On average each predictive attribute has 9.1 Markov blanket attributes for KCF, whereas only 4.1 Markov blanket attributes for KDB.

Conditional mutual information $I(X_i; X_j|C)$ can be used to quantitatively evaluate the conditional dependence between X_i and X_j given C. For any given target attribute X_k, X_k is directly dependent on $Pa(X_k)$ and $Ch(X_k)$ is directly dependent on X_k. Thus the conditional dependencies are measured by $I(X_i; X_k|C)$ and $I(X_j; X_k|C)$ $(X_i \in Pa(X_k), X_j \in Ch(X_k))$, respectively. $PC(X_k)$ is conditionally dependent on X_k but directly dependent on $Ch(X_k)$. The conditional dependence is measured by $I(X_i^{'}; X_j^{'}|C)$ $(X_i^{'} \in PC(X_k), X_j^{'} \in Ch(X_k))$. All the conditional dependencies among attributes in $MB(X_k)$ can then be measured by $MB\_Info(X_k)$, which is

Attribute	X_0		X_1		X_2		X_3		X_4		X_5		X_6		X_7		X_8		X_9		X_{10}		X_{11}		X_{12}		X_{13}	
Attributes contained in the Markov blanket inKDB model and KCF model	X_{12} X_{17} X_{19} X_{28}	X_{18} X_{26} X_{27}	X_{23} X_{28}	X_4 X_5 X_7 X_8 X_9 X_{10} X_{12} X_{14} X_{21} X_{22} X_{24}	X_{21} X_{25}	X_6 X_{13} X_{15} X_{23} X_{24} X_{25}	X_{14} X_{16} X_{17} X_{19}	X_{13} X_{16} X_{21} X_{24} X_{25}	X_{19} X_{28}	X_1 X_5 X_7 X_8 X_9 X_{10} X_{12} X_{14} X_{21} X_{22} X_{24}	X_{19} X_{28}	X_1 X_4 X_7 X_8 X_9 X_{10} X_{12} X_{14} X_{21} X_{22} X_{24}	X_{23} X_{28}	X_2 X_{15} X_{23} X_{25}	X_{17} X_{28}	X_1 X_4 X_5 X_8 X_9 X_{10} X_{12} X_{14} X_{21} X_{22} X_{24}	X_{19} X_{28}	X_1 X_4 X_5 X_7 X_9 X_{10} X_{12} X_{14} X_{21} X_{22} X_{24}	X_{21} X_{28}	X_1 X_4 X_5 X_7 X_8 X_{10} X_{12} X_{14} X_{21} X_{22} X_{24}	X_{19} X_{28}	X_1 X_4 X_5 X_7 X_8 X_9 X_{12} X_{14} X_{21} X_{22} X_{24}	X_{21} X_{28}	X_{21} X_{22}	X_0 X_{17}	X_1 X_4 X_5 X_7 X_8 X_9 X_{10} X_{14} X_{21} X_{22} X_{24}	X_{19} X_{28}	X_3 X_{16} X_{21} X_{24} X_{25}

X_{14}		X_{15}		X_{16}		X_{17}		X_{18}		X_{19}		X_{20}		X_{21}		X_{22}		X_{23}		X_{24}		X_{25}		X_{26}		X_{27}		X_{28}	
X_3 X_{19}	X_1 X_4 X_5 X_7 X_8 X_9 X_{10} X_{12} X_{21} X_{22} X_{24}	X_{19} X_{28}	X_2 X_6 X_{23} X_{25}	X_3 X_{17} X_{21}	X_3 X_{13} X_{21} X_{24} X_{25}	X_0 X_3 X_7 X_{12} X_{16} X_{18} X_{19} X_{21} X_{25} X_{28}	X_{23} X_{24} X_{25} X_{27} X_{28}	X_{17} X_{19}	X_0 X_{26} X_{27}	X_0 X_3 X_4 X_5 X_8 X_{10} X_{13} X_{14} X_{15} X_{17} X_{18} X_{21} X_{25} X_{28}	X_{25} X_{26} X_{27}	X_{21} X_{23}	X_{25} X_{28}	X_2 X_9 X_{11} X_{16} X_{17} X_{19} X_{20} X_{23} X_{25} X_{26} X_{27} X_{28}	X_0 X_1 X_3 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_{11} X_{12} X_{13} X_{14} X_{16} X_{22} X_{23} X_{24} X_{25} X_{26}	X_{23} X_{25}	X_1 X_4 X_5 X_7 X_8 X_9 X_{10} X_{11} X_{12} X_{13} X_{14} X_{21} X_{24}	X_1 X_6 X_{20} X_{21} X_{22} X_{24} X_{25} X_{28}	X_2 X_3 X_6 X_{11} X_{15} X_{16} X_{17} X_{21} X_{22} X_{24} X_{25} X_{27} X_{28}	X_{23} X_{25}	X_1 X_2 X_3 X_4 X_5 X_7 X_8 X_9 X_{10} X_{11} X_{12} X_{13} X_{14} X_{16} X_{17} X_{21} X_{22} X_{23} X_{25} X_{27} X_{28}	X_2 X_{17} X_{19} X_{21} X_{22} X_{23} X_{24} X_{27}	X_0 X_1 X_2 X_3 X_5 X_6 X_8 X_{10} X_{11} X_{13} X_{15} X_{16} X_{17} X_{19} X_{20} X_{21} X_{22} X_{23} X_{24} X_{26} X_{27} X_{28}	X_{21} X_{27}	X_0 X_5 X_{11} X_{17} X_{18} X_{19} X_{21} X_{22} X_{24} X_{25} X_{27}	X_{21} X_{25} X_{26}	X_0 X_2 X_{11} X_{17} X_{18} X_{19} X_{22} X_{23} X_{24} X_{25} X_{26} X_{28}	X_0 X_1 X_4 X_5 X_6 X_7 X_8 X_9 X_{10} X_{11} X_{13} X_{15} X_{17} X_{19} X_{21} X_{23}	X_{17} X_{20} X_{23} X_{24} X_{25} X_{27}

Fig 5. The Markov blanket for KDB ($k = 2$) model is in yellow background and that for KCF ($k = 2$) model is in blue background.

https://doi.org/10.1371/journal.pone.0182070.g005

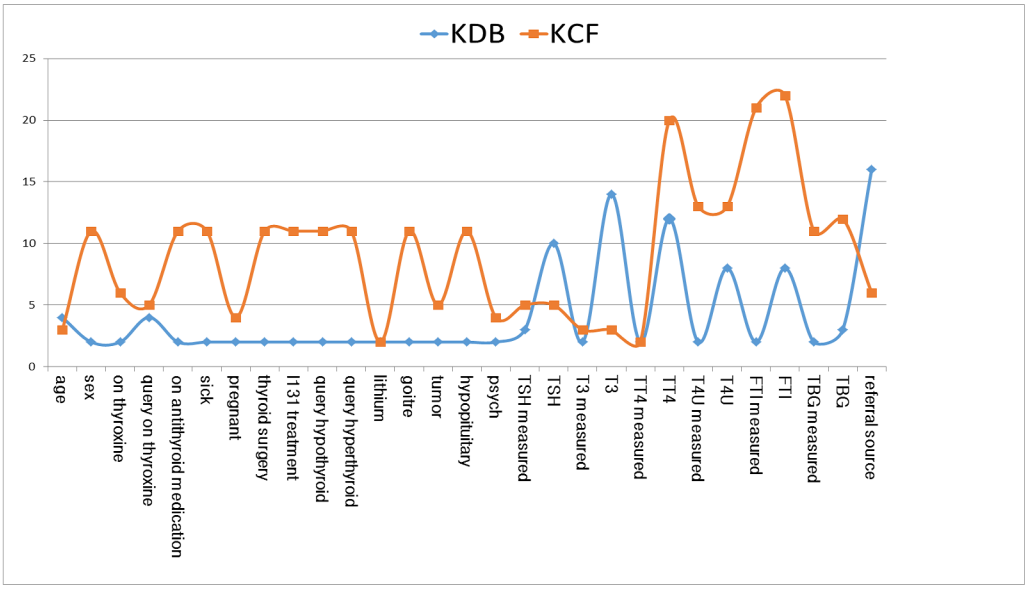

Fig 6. The number of attributes contained in the Markov blanket of each attribute in the KDB ($k = 2$) model and KCF ($k = 2$) model.

https://doi.org/10.1371/journal.pone.0182070.g006

defined by Eq (5),

$$MB\_Info(X_k) = \sum_{X_i \in Pa(X_k)} I(X_i; X_k|C) + \sum_{X_j \in Ch(X_k)} I(X_j; X_k|C) + \sum_{X_i' \in PC(X_k)} \sum_{X_j' \in Ch(X_k)} I(X_i'; X_j'|C) \quad (5)$$

We also compare the average weight of conditional dependencies implicated in $MB(X_k)$, which is defined by Eq (6),

$$Avg\_MB\_Info(X_k) = \frac{MB\_Info(X_k)}{number\ of\ attributes\ in\ MB(X_k)} \quad (6)$$

The comparison results of $MB\_Info(X_k)$ between KCF and KDB are shown in Fig 7. For the first 14 attributes, $MB\_Info(X_k) \approx 0$ $\{0 \leq k \leq 13\}$ for both KDB and KCF. Thus X_k $\{0 \leq k \leq 13\}$ is directly dependent on class variable whereas independent of any other attributes. For 13 of the other 15 attributes, the value of $MB\_Info(X_k)$ $\{14 \leq k \leq 28\}$ for KCF is greater than that for KDB. The experimental results prove that KCF can fully demonstrate dependency relationships and thus help to increase the classification accuracy.

Discussion

Thyroid cancer incidence has been rising since 1978, and its prevalence has increased dramatically over the past decade; currently, thyroid cancer is the fifth most common cancer diagnosed among women. By contrast, the incidence of other malignancies, including lung, colorectal, and breast cancer, decreases [28]. A statistical survey in 2014 showed that 10 million Chinese patients have hyperthyroidism, 90 million have hypothyroidism, more than 100

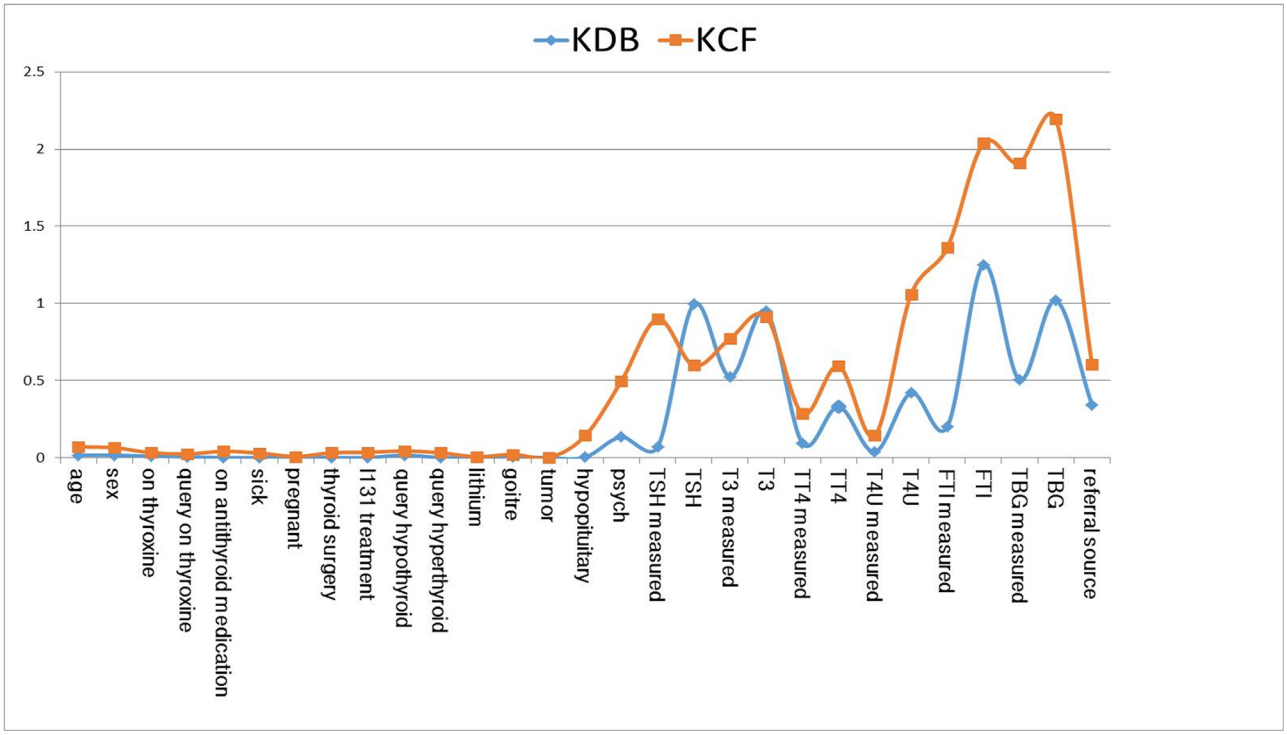

Fig 7. The sum of conditional mutual information between each attribute and the attributes contained in its Markov blanket is shown in (a). The average of conditional mutual information between each attribute and the attributes contained in its Markov blanket is shown in (b).

https://doi.org/10.1371/journal.pone.0182070.g007

million are afflicted with thyroid nodules or thyroid cancer, and conservatively; more than 200 million are estimated to have thyroid disease. As the second major disease of the endocrine system, the awareness rate and treatment rate of thyroid diseases are very low in China.

Thyroid nodule is a common clinical problem, and the prevalence of differentiated thyroid cancer increases [29]. Early detection, diagnosis, and treatment are important in curbing the development of thyroid diseases and reducing the mortality rate. Predicting the outcome of diseases and dependency among clinical variables or attributes plays pivotal roles in medical diagnosis and treatment.

For the detailed analysis, this paper calculates and compares the mutual information $I(X_i;C)$ first. The results are sorted starting from the highest. The attribute order is X_{17}, X_{25}, X_{21}, X_{19}, X_{23}, X_2, X_{28}, X_{27}, X_{16}, X_{20}, X_{26}, X_{18}, X_{22}, X_{24}, X_0, X_1, X_6, X_{10}, X_{13}, X_7, X_9, X_{15}, X_4, X_8, X_5, X_3, X_{12}, X_{11}, X_{14}. From the perspective of medical diagnosis, the attribute with the most intimate relationship with the outcome can be considered as the key attribute and should be the focus of the analysis. The attribute X_{17} represents the clinical index for thyroid stimulating hormone (TSH) and should be analyzed initially. TSH can promote the growth of thyroid secreted by adenohypophysis. In addition, TSH can completely improve the function of the thyroid, promoting early release of thyroid hormones and synthesis of T4 and T3.

To clarify the role of the TSH attribute, this paper displays the structure of the KDB and a KCF submodel in Fig 8(a) and 8(b), respectively. To make typical and fair comparison, we set X_{17} as the common root node of both models. As shown in Fig 8(a), X_{17} is the common parent of X_{25}, X_{21}, X_{28}, X_{16}, X_{18}, X_{17}, X_3, and X_{12}; X_0 and X_{19} are the parent nodes of the children of

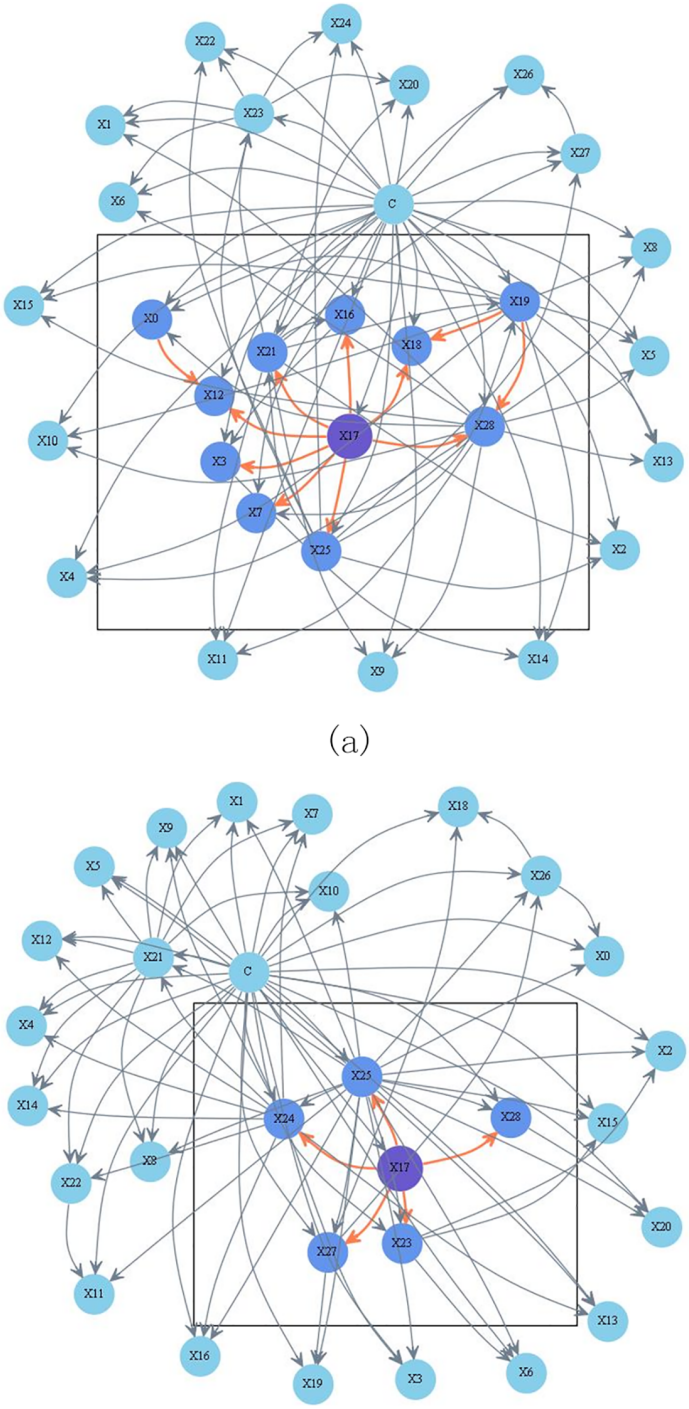

Fig 8. KDB ($k = 2$) model and a submodel of the KCF ($k = 2$) on thyroid disease date set shown respectively in (a) and (b). Attributes $\{X_0, X_1, \cdots, X_{28}, C\}$ correspond to clinical variables *age*, *sex*, *on thyroxine*, *query on thyroxine*, *on antithyroid medication*, *sick*, *pregnant*, *thyroid surgery*, *I131 treatment*, *query hypothyroid*, *query hyperthyroid*, *lithium*, *goitre*, *tumor*, *hypopituitary*, *psych*, *TSH measured*, *TSH*, *T3 measured*, *T3*, *TT4 measured*, *TT4*, *T4U measured*, *T4U*, *FTI measured*, *FTI*, *TBG measured*, *TBG*, *referral source* and *Class* respectively.

https://doi.org/10.1371/journal.pone.0182070.g008

X_{17}. X_0 is the parent node of X_{12}, and X_{19} is the common parent of X_{18} and X_{28}. $MB(X_{17})$ contains 10 attributes. $MB\_Info(X_{17})$ is 0.902 and $Avg\_MB\_Info(X_{17}) = 0.09$. In the corresponding KCF model shown in Fig 8(b), X_{17} is the common parent of X_{23}, X_{24}, X_{25}, X_{27}, and X_{28}, whereas X_{17} has no parent nodes and no parent of children nodes. Thus, $MB(X_{17})$ only contains 5 attributes. $MB\_Info(X_{17})$ and $Avg\_MB\_Info(X_{17})$ turn to be 0.597 and 0.12, respectively. Similarly, the sum of $MB\_Info(X_i)$, i.e., $\sum_{i=0}^{28} MB\_Info(X_i)$, is 14.458 for KCF, whereas it is only 6.964 for KDB. The sum of $Avg\_MB\_Info(X_i)$, i.e., $\sum_{i=0}^{28} Avg\_MB\_Info(X_i)$, is 1.576 for KDB and 1.946 for KCF. Hence, the proposed KCF model describes significant relationships among attributes.

MST contains the most significant dependency relationships, whereas the KDB model can only contain portions of the MST. Additionally, the KCF algorithm can generate a series of submodels rather than one model alone. Thus, for medical diagnosis, any clinical variable or attribute related to thyroid diseases can be regarded as the original cause, and an in-depth research can be conducted on the disease. Hence, the proposed KCF model can handle various patient conditions and is more suitable for providing appropriate treatment compared with a model with a rigid root node generated by other algorithms.

Sensitivity and specificity are statistical measures of the performance of a binary classification test, also known in statistics as classification function. In the context of medical tests sensitivity is the extent to which true positives are not missed/overlooked and specificity is the extent to which positives really represent the condition of interest and not some other condition being mistaken for it. So we select 12 datasets with binary class labels from UCI for comparison of classification accuracy. Table 2 summarizes the characteristics of each dataset, including the numbers of instances, attributes and classes. Averaged One-dependence Estimators (AODE) [30], which utilizes a restricted class of one-dependence estimators and aggregates the predictions of all qualified estimators within this class, is introduced to compare the bagging performance of KCF.

Experimental results of average classification accuracy for different BNCs are shown in Table 3. Friedman test [31], which is a non-parametric measure to compare the ranks of the algorithms for each dataset separately. The ranks of algorithms for each dataset are calculated separately (average ranks are assigned if tied values exist). The null-hypothesis is that all the algorithms performs almost equivalently and there is no significant difference in terms of

Table 2. Datasets.

No.	dataset	Instance	Attribute	Class
1	Echocardiogram	131	6	2
2	Heart*	270	13	2
3	Heart Disease*	303	13	2
4	Chess	551	39	2
5	Breast-cancer-w*	699	9	2
6	Pima-ind-diabetes*	768	8	2
7	Tic-tac-toe	958	9	2
8	German	1000	20	2
9	Spambase	4601	57	2
10	Mushroom	8124	22	2
11	Adult	48842	14	2
12	Census-income	299285	41	2

the datasets denoted with symbol "*" will be used for comparing sensitivity and specificity.

https://doi.org/10.1371/journal.pone.0182070.t002

Table 3. Experimental results of average classification accuracy for datasets with binary class labels.

Dataset	NB	TAN	KCF	KDB	AODE
Adult	84.2%	86.2%	85.1%	86.2%	86.8%
Breast-cancer-w	95.8%	96.4%	97.4%	95.3%	94.6%
Census-income	76.3%	93.6%	89.9%	94.9%	94.9%
Chess	88.7%	90.7%	90.0%	90.0%	92.4%
Echocardiogram	66.4%	67.2%	67.9%	65.6%	66.4%
German	74.7%	72.7%	75.2%	71.1%	73.0%
Heart	80.2%	80.7%	80.8%	81.9%	80.4%
Heart Disease	79.9%	79.2%	78.8%	77.6%	79.6%
Mushrooms	98.0%	100.0%	100.0%	100.0%	100.0%
Pima-ind-diabetes	75.5%	76.2%	76.2%	75.5%	76.3%
Spambase	89.8%	93.3%	93.3%	93.6%	94.1%
Tic-tac-toe	69.3%	77.1%	73.5%	79.6%	80.6%

https://doi.org/10.1371/journal.pone.0182070.t003

ranks. The Friedman statistic can be computed as Eq (7) shows,

$$F_r = \frac{12}{Nt(t+1)} \sum_{j=1}^{t} R_j^2 - 3N(t+1) \tag{7}$$

where $R_j = \sum_i r_i^j$ and r_i^j is the rank of the j-th of t algorithms on the i-th of N datasets. Thus, for any pre-determined level of significance α the null hypothesis will be rejected if $F_r > \chi_\alpha^2$, which is the upper-tail critical value having $t-1$ degrees of freedom. The critical value of χ_α^2 for $\alpha = 0.02$ is 11.668. With 5 algorithms and 12 datasets, the friedman statistic $F_r = 18.55$ and $P < 0.001$. Hence the null-hypotheses is rejected again. The average ranks of different classifiers are {NB(1.54), TAN(3.00), AODE(2.54), KDB(3.88), KCF(4.04)}. Thus KCF with the highest rank is the most effective BNC from the perspectives of classification accuracy.

When dealing with imbalanced class distribution, traditional classifiers are easily overwhelmed by instances from majority classes while the instances from minority classes are usually ignored. An useful performance measure is the balanced accuracy (BAC) [32] which avoids inflated performance estimates and defined as Eq (8) shows. It is defined as the arithmetic mean of sensitivity and specificity, which are calculated by knowing the m binary outputs of the classifiers (indicating membership to given classes). Overall performance is calculated by conducting a leave-one-out test for all training samples.

$$BAC = \frac{sensitivity + specificity}{2} \tag{8}$$

The experimental results of sensitivity, specificity and BAC for BNCs are shown in Table 4. By comparing via two-tailed binomial sign test with a 95% confidence level, Table 5 shows corresponding win/draw/loss (W/D/L) records summarizing the relative BAC of the different BNCs. The W/D/L record in cell $[i, j]$ of each table contains the number of datasets in which BNC on row i has lower, equal or higher outcome relative to the BNC on column j. We could see from Table 5 that the bagging mechanism helps AODE increase BAC significantly often relative to TAN and NB. KDB can achieve not only higher classification accuracy but also higher BAC than TAN. KCF utilizes the bagging mechanism of AODE and can represent

Table 4. Experimental results of sensitivity, specificity and BAC for medical datasets with binary class labels.

	Dataset	NB	TAN	AODE	KDB	KCF
sensitivity	Breast-cancer-w	0.969	0.973	0.965	0.958	0.971
	Heart	0.840	0.853	0.860	0.853	0.806
	Heart-disease-c	0.829	0.856	0.842	0.816	0.786
	Pima-ind-diabetes	0.820	0.842	0.824	0.838	0.816
specificity	Breast-cancer-w	0.917	0.929	0.945	0.975	0.975
	Heart	0.742	0.750	0.756	0.792	0.854
	Heart-disease-c	0.748	0.741	0.813	0.776	0.846
	Pima-ind-diabetes	0.634	0.612	0.631	0.619	0.642
BAC	Breast-cancer-w	0.943	0.952	0.955	0.966	0.973
	Heart	0.798	0.802	0.808	0.826	0.830
	Heart-disease-c	0.788	0.798	0.802	0.797	0.816
	Pima-ind-diabetes	0.727	0.726	0.727	0.728	0.729

https://doi.org/10.1371/journal.pone.0182070.t004

Table 5. Win-draw-loss records for different BNCs in terms of BAC.

	Dataset	NB	TAN	AODE	KDB
BAC	TAN	1/3/0		-	-
	AODE	2/2/0	1/3/0	-	-
	KDB	2/2/0	2/2/0	2/1/1	-
	KCF	3/1/0	3/1/0	3/1/0	2/2/0

https://doi.org/10.1371/journal.pone.0182070.t005

high-dependence relationships. This may be the main reason why KCF achieves higher BAC more often than the other four BNCs.

Conclusion

Bayesian network can graphically describe the conditional dependencies implicit in training data and Bayesian network classifiers have been previously demonstrated to perform efficiently in medical diagnosis and treatment. One single data mining model cannot deal with all difficult and complicated cases. KCF, which uses the same learning strategy as that of KDB, simultaneously provides n submodels rather than one. This improvement helps KCF to describe more significant conditional dependencies. The experimental study on UCI datasets shows that KCF enjoys obvious advantage in classification over other BNCs.

Acknowledgments

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

Author Contributions

Conceptualization: LiMin Wang, FangYuan Cao.

Data curation: FangYuan Cao.

Formal analysis: LiMin Wang, MingHui Sun.

Funding acquisition: LiMin Wang.

Investigation: LiMin Wang, ShuangCheng Wang.

Methodology: LiMin Wang.

Project administration: LiMin Wang.

Resources: LiYan Dong.

Software: LiMin Wang.

Supervision: LiMin Wang.

Validation: ShuangCheng Wang.

Visualization: FangYuan Cao.

Writing – original draft: LiMin Wang, FangYuan Cao.

Writing – review & editing: LiMin Wang, FangYuan Cao.

References

1. Yoo I, Alafaireet P, Marinov M. Data mining in healthcare and biomedicine: a survey of the literature. Journal of Medical Systems. 2012; 36(4): 2431–2448. https://doi.org/10.1007/s10916-011-9710-5 PMID: 21537851
2. Wu X, Zhu X, Wu GQ. Data Mining with Big Data. IEEE Transactions on Knowledge & Data Engineering. 2014; 26(1): 97–107. https://doi.org/10.1109/TKDE.2013.109
3. Sumalatha G, Muniraj NJR. Survey on medical diagnosis using data mining techniques. International Conference on Optical Imaging Sensor and Security. 2013; 1–8. https://doi.org/10.1109/ICOISS.2013.6678433
4. Liu DY, Chen HL, Yang B, Lv XE, Li LN. Design of an Enhanced Fuzzy *k*-nearest Neighbor Classifier Based Computer Aided Diagnostic System for Thyroid Disease. Journal of Medical Systems, 2012; 36 (5), 3243–3254. https://doi.org/10.1007/s10916-011-9815-x PMID: 22198094
5. Xu YB, Wang H, Zhou Q, Jiang J, Ma WQ. Logistic regression analysis of contrast-enhanced ultrasound and conventional ultrasound characteristics of sub-centimeter thyroid nodules. Ultrasound in Medicine and Biology, 2015; 41(12): 3102–3108. https://doi.org/10.1016/j.ultrasmedbio.2015.04.026
6. Feyzullah T. A comparative study on thyroid disease diagnosis using neural networks. Expert Systems with Applications, 2009; 36(1): 944–949. https://doi.org/10.1016/j.eswa.2007.10.010
7. Galli M, Zoppis I, Sio GD, Chinello C, Pagni F. A Support Vector Machine Classification of Thyroid Bioptic Specimens Using MALDI-MSI Data. Advances in Bioinformatics, 2016; 2016(1): 1–7. https://doi.org/10.1155/2016/3791214
8. Pyo JS, Sohn JH, Kang G. Diagnostic assessment of intraoperative cytology for papillary thyroid carcinoma: using a decision tree analysis. Journal of Endocrinological Investigation, 2016; 1–7.
9. Trotta R. Bayes in the sky: Bayesian inference and model selection in cosmology. Contemporary Physics. 2008; 49(2): 71–104. https://doi.org/10.1080/00107510802066753
10. Reiz B, Csató L. Tree-like bayesian network classifiers for surgery survival chance prediction. International Journal of Computers Communications & Control. 2008; 3(3): 470–474.
11. Arroyo-Figueroa G, Sucar LE. A Temporal Bayesian Network for Diagnosis and Prediction. Eprint Arxiv. 2013; 13–20.
12. Gadewadikar J, Kuljaca O, Agyepong K. Exploring Bayesian networks for automated breast cancer detection. IEEE Southeastcon. 2009; 153–157.
13. Lee J, Jun CH. Classification of High Dimensionality Data through Feature Selection Using Markov Blanket. Industrial Engineering & Management Systems. 2015; 14(2):210–219. https://doi.org/10.7232/iems.2015.14.2.210
14. Koller D, Sahami M. Toward Optimal Feature Selection. Proc. 13th International Conference on Machine Learning. Morgan Kaufmann, 1996; 284–292.

15. Fu S, Desmarais MC. Tradeoff Analysis of Different Markov Blanket Local Learning Approaches. Pacific-asia Conference on Advances in Knowledge Discovery & Data Mining. Springer-Verlag. 2008; 562–571.
16. UCI Machine Learning Repository. David Aha. 1987. http://archive.ics.uci.edu/ml/index.php
17. Heckerman D. A tutorial on learning with bayesian networks. Learning in Graphical Models. 1995; 25 (4):33–82.
18. Friedman N, Dan G, Goldszmidt M. Bayesian network classifiers. Wiley Encyclopedia of Operations Research & Management Science, 2011; 29(2-3):598–605.
19. Chow CK, Liu CN. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory. 1968; 14(3):462–467. https://doi.org/10.1109/TIT.1968.1054142
20. Sahami M. Learning Limited Dependence Bayesian Classifiers. In the 2nd International Conference. Knowledge Discovery and Data mining (KKD). 1996; 335–338.
21. Martinez AM, Webb GI, Chen SL, Nayyar AZ. Scalable learning of Bayesian network classifiers. Journal of Machine Learning Research. 2013; 1–30.
22. David WA, Dennis K, Marc KA. Instance-based learning algorithms. Machine Learning, 1991; 6(1): 37–66. https://doi.org/10.1007/BF00153759
23. Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation, 1989; 13(3): 637–649. https://doi.org/10.1162/089976601300014493
24. David R, James M. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. American Scientist, 1988.
25. Iba Wayne, Langley P. Induction of One-Level Decision Trees, in Proceedings of the Ninth International Conference on Machine Learning, San Francisco, CA: Morgan Kaufmann, 1992; 233–240.
26. Marc S, Eibe F, Mark H. Speeding up logistic model tree induction. in Proceeding of the 9th European conference on Principles and Practice of Knowledge Discovery in Databases, 2005; 675–683.
27. Burduk R. Comparison of cost for zero-one and stage-dependent fuzzy loss function. Intelligent Information & Database Systems. 2012; 7196:385–392.
28. O'Grady TJ, Gates MA, Boscoe FP. Thyroid cancer incidence attributable to overdiagnosis in the United States 1981-2011: Thyroid Cancer Overdiagnosis. International Journal of Cancer. 2015; 2664–2673. https://doi.org/10.1002/ijc.29634 PMID: 26069163
29. Cooper DS, Doherty GM, Haugen BR, Kloos RT. Revised American Thyroid Association management guidelines for patients with thyroid nodules and differentiated thyroid cancer. Thyroid Official Journal of the American Thyroid Association. 2009; 19(2): 1167–1214. https://doi.org/10.1089/thy.2009.0110 PMID: 19860577
30. Zheng F, Geoffrey W. Efficient lazy elimination for averaged one-dependence estimators. in Proceedings of the Twenty-third International Conference on Machine Learning, 2006; 1113–1120.
31. Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research. 2006; 7: 1–30.
32. Redlarski G, Gradolewski D, Palkowski A. A System for Heart Sounds Classification. PLoS ONE. 2014; 9(11): e112673. https://doi.org/10.1371/journal.pone.0112673 PMID: 25393113