Enhancing Dependence Representation of Clinical Variables based on Bayesian Networks to Assist in Thyroid Disease Diagnosis and Treatment

Fangyuan Cao[1, a], Limin Wang[2,b]

[1,2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, ChangChun, China

[a]18744019469@163.com, [b]wanglim@jlu.edu.cn

Keywords: Bayesian network classifier; medical diagnosis; *k*-dependence causal forest; maximum-weighted spanning tree;

Abstract. In the past few decades, a large number of data mining algorithms have been proposed to assist in the practical medical diagnosis. This paper proposes a novel algorithm, *k*-dependence causal forest (KCF) based on Bayesian network. The experiments conducted on the thyroid disease data set suggest the KCF model always has lower 0-1 loss and contains more conditional mutual information, which means the KCF model is suitable to assist in thyroid disease diagnosis and treatment.

Introduction

Data mining[1] is a process, which can extract information and knowledge potentially useful from amount of incomplete, noisy, fuzzy and random practical application data. Data mining technology has become a popular tool for medical researchers who seek to identify patterns and relationships among a large number of attributes as it is able to predict the outcome of a disease using the historical cases data sets[2]. During the past few decades, many studies have been done to construct various prediction models for medical diagnosis, such as Bayesian networks, decision trees, artificial neural networks and so on. As utilizing graphical models to express its structure, Bayesian network models are more distinct and understandable than others. Reiz et al.[3] used Bayesian network and graph theory to encode causal relations between variables. Sucar et al.[4] constructed a temporal Bayesian network to make diagnosis and prediction in medical domain. Gadewadikar et al.[5] made use of Baysian network model to construct automated breast cancer detection support tool.

Bayesian network is a kind of probability network, which is based on graphical network to make probabilistic inference. A Baysian network cannot contain all the relationships between the attributes and the class, so there should be a trade-off between structure complexity and classification accuracy. Some restricted Baysian classifiers (e.g., Naive Bayes (NB), tree augmented Naive Bayes (TAN) and k-dependence BNs (KDB)) are popular for the good performance in classification with different levels of the conditional independence assumption. Next, this paper gives an outline of the three popular restricted BN classifiers (i.e., NB, TAN and KDB), and proposes a novel learning algorithm called the *k*-dependence causal forest (KCF). The goal of this research is to construct a stable and efficient Bayesian network model for assisting medical specialist to make accurate diagnosis and treatment of thyroid disease patients.

Three restricted Bayesian classifiers

The Naive Bayesian(NB) network classifier is the simplest Bayesian network model, it assumes that all the predictive attributes are independent from each other given class label[6]. Suppose there are n independent attributes $X=\{ X_1,..., X_n \}$ and m classes $C_1,C_2,...,C_m$, then classification is to derive the maximum $P(c_i \mid x)$, where $i \in (1, m)$, this can be derived from Bayesian theorem in Eq.1:

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

The rigorous assumption in NB is that all the attributes are conditionally independent, so 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 [7] can be seen as an extension of the Chow-Liu tree [8], which utilizes conditional mutual information to build a maximum spanning tree (MST). TAN is a 1-dependence classifier, because it allows each attribute to have at most one parent besides the class. For a TAN model, the class assignments of the test samples are based on Eq.3:

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

where X_p is the parent node of X_j.

The k-dependence structure (KDB) [9] is a k-dependence classifier, as it allows each attribute to have a maximum of k parents besides the class attribute. There is a pre-determined attribute order by comparing mutual information $I(X_i;C)$ between predictive attribute and class attribute, starting with the highest. Each attribute can select a maximum of k parents among the attributes ahead of itself in the predetermined order with the maximum conditional mutual information $I(X_i;X_j|C)$. For a KDB model, the class assignments of the test samples are based on Eq.4:

$$\arg \max_{c_i} P(c_i)P(x|c_i) = \arg \max_{c_i} P(c_i)\prod_{j=1}^{n} P(x_j|c_i,x_{j1},...,x_{jp}) \tag{4}$$

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

The k-dependence Causal Forest Algorithm

As MST contains most significant relationships between attributes, so we try to achieve high-dependence directed trees by extending MSTs, there will be one k-dependence causal forest (KCF) rather than one augmented tree. An example of the condition mutual information matrix with four attributes is shown in Fig.1(a), after selecting each attribute as the root node and setting the direction of all arcs to be outward from it, four different directed MSTs are generated, as shown from Fig.1(b) to Fig.1(e), the root node is filled in black. Leaf node X_i can select other nodes as parents along the path from X_i to the root node by comparing conditional mutual information. If k > 1, e.g., k = 2, more parents can be selected for each non-root node by comparing conditional mutual information. Fig.2 shows k-dependence Bayesian classifiers when k = 2. Compared with Fig.1 the new added arcs are annotated with red.
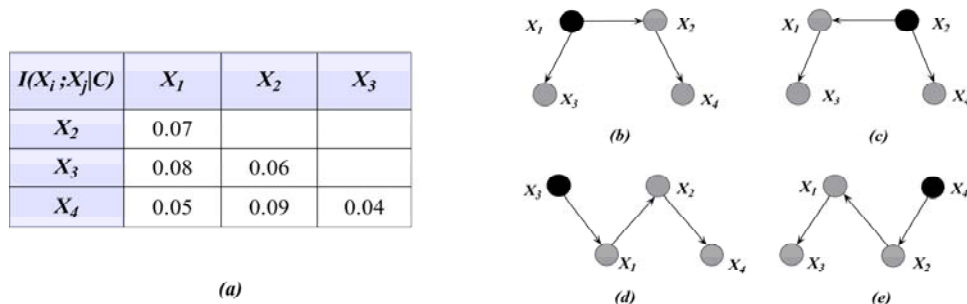

$I(X_i;X_j\vert C)$	X_1	X_2	X_3
X_2	0.07		
X_3	0.08	0.06	
X_4	0.05	0.09	0.04

(a)

(b) *(c)* *(d)* *(e)*

Figure 1 An example of directed MSTs with different root nodes, which are filled in black.
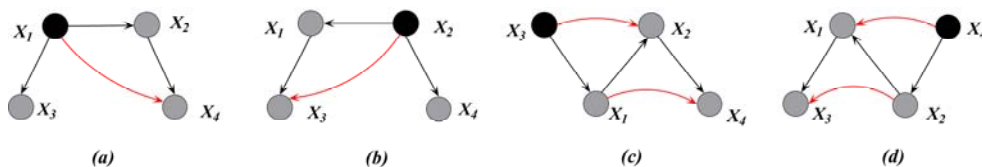

(a) *(b)* *(c)* *(d)*

Figure 2 The KCF (k=2) model corresponding to the MSTs shown in Fig.1.

0314

The detailed KCF model algorithm are depicted in Algorithm 1.

Algorithm 1 KCF

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

Output: Subclassifiers $\{ KCF_1 ,..., KCF_n \}$.

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 $(i = 1,2,...,n)$

(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.

For each KCF_i, the space complexity increases exponentially as k increases, to achieve a trade-off between classification performance and efficiency we restrict the structure complexity to be 2-dependence in the following experiments.

Experiments and results

The experiments are conducted on the thyroid disease data set in UCI database, and the comparisons are among NB, TAN, KDB(set k=2) and KCF(set k=2). For the original thyroid disease data set, MDL discretization [10] is used to discretize numeric attributes. The averaged 0-1 loss of KCF model is 0.065, which is closer to zero compared with NB(0.111), TAN(0.072) and KDB(0.071). As the results of 0-1 loss shown, the proposed KCF algorithm is computationally efficient and usually achieve lower 0-1 loss compared with other algorithms on the thyroid disease data set, which means the corresponding KCF classifier performs efficient in thyroid disease diagnosis.

Then this paper compares the capacity of dependence representation in different algorithms, as the algorithm with a higher dependence level always has better dependence representation of attributes, so we make the comparison between KCF model and KDB model. Markov blanket can be seen as a tool to estimate the capacity of dependence representation. In a Bayesian network model, the Markov blanket of T is the union of parent, child, and parent of children, nodes of T [11, 12]. This paper locates the Markov blanket of each attribute contained in thyroid disease data set in the KCF model and the KDB model respectively, the results are shown in Fig.3.
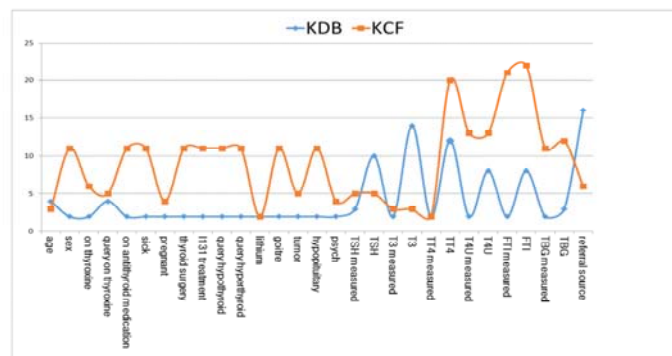

Figure 3 The comparison of Markov blanket attribute numbers between KDB ($k = 2$) and KCF ($k = 2$).

Obviously different, among all the 29 attributes, only 4 attributes (X_0, X_{17}, X_{19} and X_{28}) have the Markov blanket with more attributes contained in the KDB model than the KCF model. After statistics, there are 120 attributes contained in the Markov blanket collection of all the 29 attributes in the KDB model but 264 in the KCF model. And the sum of the conditional mutual information that contained in the Markov blanket of each attribute is 6.964 in KDB model but 14.458 in KCF model, which can fully demonstrate that the dependence representation of attributes is expressed more sufficiently in the KCF model than the corresponding KDB model on thyroid data set.

Conclusions

Bayesian classifiers have been previously demonstrated to perform efficiently in medical diagnosis and treatment. On the basis of analyzing and summarizing the working mechanisms of the three classical Bayesian classifiers, this paper proposed a novel algorithm, KCF, which will generate a *k*-dependence causal forest and the class membership probabilities produced by different subclassifiers are directly averaged. We conduct the experiments on thyroid disease data set, the results suggest that the KCF classifier has substantially lower 0-1 loss and stronger relationship dependence representation compared with the three classical models which demonstrates the KCF model is profitable to be applied in the diagnosis and treatment of thyroid disease.

Acknowledgements

References

[1] Yoo I, Alafaireet P, Marinov M, et al.: Journal of Medical Systems,Vol. 2431-2448 (2012)

[2] Wu X, Zhu X, Wu G Q, et al. : *Data Mining with Big Data*. Knowledge & Data Engineering IEEE Transactions on. 97-107 (2014).

[3] Reiz B, Csat´o L:International Journal of Computers Communications & Control 470-474 (2008)

[4] Arroyo-Figueroa G, Sucar L E. : A Temporal Bayesian Network for Diagnosis and Prediction. Eprint Arxiv 13-20 (2013).

[5] W Gadewadikar J, Kuljaca O, Agyepong K, et al : Exploring Bayesian networks for automated breast cancer detection. N/a.225-231 (2010).

[6] Heckerman D: Learning in Graphical Models. 33-82 (1995).

[7] Friedman N, Dan G, Goldszmidt M : Bayesian network classifiers. Wiley Encyclopedia of Operations Research & Management Science. 598-605 (2011).

[8] Chow C K, Liu C N : Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory. 462-467 (1968).

[9] M. Sahami : Learning Limited Dependence Bayesian Classifiers. In the 2nd International Conference, Knowledge Discovery and Data mining (KDD). 335-338 (1996).

[10] Clarke E J, Barton B A : International Journal of Intelligent Systems. 61-92 (2000).

[11] Lee J, Jun C H : Classification of High Dimensionality Data through Feature Selection Using Markov Blanket. Industrial Engineering & Management Systems. 210-219 (2015).

[12] Fu S, Desmarais M C : Tradeoff Analysis of Different Markov Blanket Local Learning Approaches. Pacific-asia Conference on Advances in Knowledge Discovery &Data Mining. 562-571 (2008).