基于条件对数似然函数导数的贝叶斯网络分类器优化算法

王中锋 王志海
(北京交通大学计算机与信息技术学院 北京 100044)

摘 要 通常基于鉴别式学习策略训练的贝叶斯网络分类器有较高的精度,但在具有冗余边的网络结构之上鉴别式参数学习算法的性能受到一定的限制.为了在实际应用中进一步提高贝叶斯网络分类器的分类精度,该文定量描述了网络结构与真实数据变量分布之间的关系,提出了一种不存在冗余边的森林型贝叶斯网络分类器及其相应的FAN学习算法(Forest-Augmented Naïve Bayes Algorithm),FAN算法能够利用对数条件似然函数的偏导数来优化网络结构学习.实验结果表明常用的限制性贝叶斯网络分类器通常存在一些冗余边,其往往会降低鉴别式参数学习算法的性能;森林型贝叶斯网络分类器减少了结构中的冗余边,更加适合于采用鉴别式学习策略训练参数;应用条件对数似然函数偏导数的FAN算法在大多数实验数据集合上提高了分类精度.

关键词 机器学习;数据挖掘;分类器;贝叶斯网络;鉴别式训练策略
中图法分类号 TP18 **DOI号**: 10.3724/SP.J.1016.2012.00364

An Optimization Algorithm of Bayesian Network Classifiers by Derivatives of Conditional Log Likelihood

WANG Zhong-Feng WANG Zhi-Hai
(*School of Computer and Information Technology, Beijing Jiaotong University, Beijing* 100044)

Abstract In general, Bayesian network classifiers trained by discriminative strategy have higher classification accuracy than others. However, the performance of discriminative parameter learning algorithms is limited in dealing with redundant edges. In order to improve the classification accuracy in a real situation, in this paper we describe the quantitative relations between Bayesian network structures and joint probability distributions, propose a Forest-Augmented Naïve Bayes classifier and its learning algorithm. An FAN classifier is a kind of Bayesian network whose structure has few redundant edges, and FAN algorithm is optimized by properties of partial derivatives of conditional log likelihood. Experimental results have shown that redundant edges in the structure of a Bayesian network classifier could degrade classification performance in common situation, and most of restricted Bayeisian network classifiers have redundant edges. Therefore, FAN classifier without redundant edges is suitable for discriminative parameter learning strategy. On most of datasets, classification accuracies of classifiers trained by FAN algorithm are enhanced.

Keywords machine learning; data mining; classifier; Bayesian network; discriminative training strategy

收稿日期:2009-02-23;最终修改稿收到日期:2011-06-12.本课题得到国家自然科学基金(60673089,60973011)资助.**王中锋**,男,1977年生,博士研究生,主要研究方向为数据挖掘和模式识别.E-mail: iewzf@163.com. **王志海**(通信作者),男,1963年生,教授,博士生导师,主要研究领域为数据挖掘和机器学习.E-mail: zhhwang@bjtu.edu.cn.

1 引　言

分类器的设计是数据挖掘和机器学习领域的主要研究内容之一.该领域提出了大量的分类器模型和学习算法.在这些模型中,贝叶斯网络分类器不但理论基础坚实,而且在实际应用中有较强的抗噪声性能和健壮性能,得到了多年的持续研究,提出了许多的学习算法[1].但是,在这些学习算法中,大多数仅是将贝叶斯网络分类器看作贝叶斯网络来处理,而没有考虑其作为分类器的特殊需求[2-4].

一般而言,以贝叶斯网络的表达能力为评价标准指导设计的贝叶斯网络分类器学习算法归为生成式学习策略,以分类器的精度为评价标准设计的归为鉴别式学习策略[4].在理想状态下,与数据变量分布一致的贝叶斯网络同时也是分类精度最高的贝叶斯网络分类器.但在实际应用中,两者难以保持一致.这主要是因为:(1)在现实中,迫于计算机的性能,从计算理论的角度考虑,搜索到精确的结构问题是一个 NP 难题[5];(2)在应用中,采样工具的选择,实验人员的熟练程度,统计方法的局限性,都会导致训练数据不能准确地代表整体数据分布,且这种差异是不可预期的[6];(3)在训练数据的数目一定的情况下,贝叶斯网络分类器的结构越复杂,条件概率表就越庞大,对网络中各个参数的估计就会越不可靠,其分类性能就有可能下降.考虑到这些实际情况,通常在设计贝叶斯网络分类器模型和算法时,限制网络结构的形式,来降低与样本的拟合程度,提高其泛化性能.这类模型和算法有朴素贝叶斯网络分类器(Naïve Bayes,NB)、树型限制性贝叶斯网络(Tree-Augmented Naïve Bayes Classifier,TAN)分类器[2]、双层贝叶斯分类器(Double-Level Bayesian Netwrok Augmented Naïve Bayes,DLBAN)[7]、消极贝叶斯网络分类器学习算法(Lazy Bayesian Rules,LBR)[8]、超父算法(Super-Parent TAN,SP)[9]、平均 1-依赖估计算法(Aggregating One-Dependence Estimators,AODE)[10]、子结构学习算法(Substructure Learning Algorithm)[11]等.对于这些限制了网络结构形式的贝叶斯网络分类器模型,最接近数据变量分布的网络结构不一定就是分类精度最高的,因此需要用鉴别式学习策略来设计算法.由于鉴别式学习策略是直接以提高分类器的精度为目标设计学习算法的,在通常情况下采用鉴别式学习策略得到的分类器精度比采用生成式学习策略得到的高[3-4,12-16].近年来,Pernkopf 等人[13]提出了一些贝叶斯网络分类器鉴别式结构学习算法,Greiner 等人[3-4]提出了 ELR 参数学习算法,Jing 等人[14-15]提出了 BNB 和 BAN 参数学习算法,Su 等人[16]提出了 DFE 参数学习算法.这些算法都是基于鉴别式学习策略设计的.

但是,本文研究表明:限制性贝叶斯网络分类器鉴别式参数学习策略仅适用于比实际数据变量分布简单的网络结构.贝叶斯网络结构与实际数据变量分布之间存在 3 种的关系,分别是简单、等于和复杂.当贝叶斯网络结构缺少表示变量之间实际存在的依赖关系的边时,网络结构简单于变量分布;当贝叶斯网络结构恰好能够表示变量之间的依赖关系时,网络结构等于变量分布;当贝叶斯网络结构存在表示变量之间依赖关系的边,而实际变量之间不存在依赖关系时,网络结构复杂于变量分布,并称这些边为冗余边.由于以往基于鉴别式学习策略设计的参数学习算法假设能够在各种网络结构上提高分类精度,所以它们仅对网络结构与实际变量分布之间的关系进行了定性的描述.本文首次定量描述了它们之间这 3 种关系,提出了一种森林型贝叶斯网络分类器(Forest-Augmented Naïve Bayes,FAN),来提高鉴别式参数学习策略的适应性,并提出了一种基于条件对数似然函数偏导数的 FAN 分类器学习算法,来识别贝叶斯网络结构中的冗余边.

本文的主要贡献包括:(1)定量描述贝叶斯网络结构与所描述的数据变量实际分布之间的关系;(2)提出一种森林型贝叶斯网络分类器 FAN,FAN 模型能够为鉴别式参数学习策略性能的发挥提供必要的保证;(3)提出一种 FAN 分类器学习算法,该算法首次应用条件对数似然函数偏导数指导贝叶斯网络分类器算法优化;(4)实验验证本文论点的正确性及所提算法的有效性.

本文第 2 节介绍相关工作;第 3 节定量描述贝叶斯网络结构与所描述的实际数据变量分布之间的关系;第 4 节提出森林型贝叶斯网络分类器 FAN;第 5 节提出一种 FAN 分类器学习算法;第 6 节实验验证森林型贝叶斯网络分类器和 FAN 算法的有效性;第 7 节总结全文.

2 贝叶斯网络分类器

贝叶斯网络分类器是一种采用贝叶斯网络进行表示的分类器函数.其学习算法的设计可以应用对

数似然函数为评价标准，也可以应用条件对数似然函数为评价标准.通常，采用前者的称为生成式学习策略，采用后者的称为鉴别式学习策略.

在正式讨论本文的内容前，为叙述的方便，先约定文中符号的基本含义.变量用大写字母表示(例如：A,B_i,Y)，其状态或取值用相应的小写字母表示(例如：a,b_i,y).变量的集合用大写黑体表示(例如：$\boldsymbol{A},\boldsymbol{\Pi}$)，变量集合的值用相应的小写黑体表示(例如：$\boldsymbol{a},\boldsymbol{\pi}$).并用花体字母(例如：$\mathbb{B},\mathbb{G}$)表示统计模型或图模型.

2.1 分类器的定义

不失一般性，首先给出分类器的定义：若问题域为类对象(或者事物的状态)，所有可能的类标签为 $c_1,c_2,\cdots,c_k$，且令 $C=\{c_1,c_2,\cdots,c_k\}$，则属性(或特征、变量)对象信息可表达为一个向量，每一个对象都是由一个相应的属性值向量 $\boldsymbol{a}\in\boldsymbol{A}$ 来描述的.给定一个属性值向量，分类任务就是判定这个向量所表达的对象属于类标签 $c_1,c_2,\cdots,c_k$ 中的哪一个.于是，一个分类器可视为一个映射 $f:\boldsymbol{A}\rightarrow C$.

对于数据集合 D，条件变量 $\boldsymbol{A}=(A_1,A_2,\cdots,A_n)$ 和类变量 C 的联合概率分布可以表示为贝叶斯网络 $\mathbb{B}=\langle\mathbb{G},\boldsymbol{\Theta}\rangle$，其中，$\mathbb{G}$ 是一个有向无环图，其中的结点都对应于 D 中的随机变量，有向边表示其终点对应的随机变量依赖于起点对应的随机变量；$\boldsymbol{\Theta}=\{\theta_1,\theta_2,\cdots,\theta_n\}$ 是网络中的参数，由一组条件概率表构成，表 θ_i 对应着变量 A_i，定义了图中结点间依赖强度的大小.从而，一个贝叶斯网络 $\mathbb{B}$ 对应于一种变量的联合概率分布，即

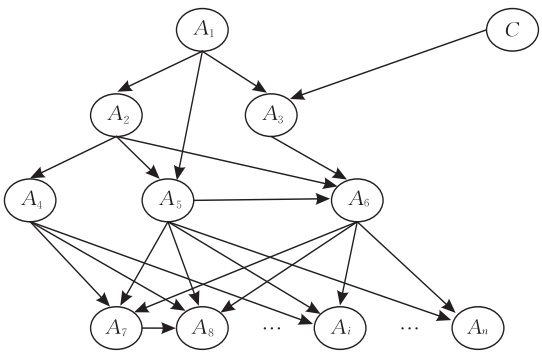
$$P_{\mathbb{B}}(A_1,A_2,\cdots,A_n)=\prod_{i=1}^{n}P_{\mathbb{B}}(A_i\mid\boldsymbol{\Pi}_i)\quad(1)$$

其中：n 表示 $\mathbb{B}$ 中随机变量的数目；$\boldsymbol{\Pi}_i$ 表示随机变量 A_i 在 $\mathbb{B}$ 中的直接父结点对应的变量的合取.进而，基于贝叶斯网络 $\mathbb{B}$ 的分类器可以定义为函数 $f_{\mathbb{B}}(\cdot)$.

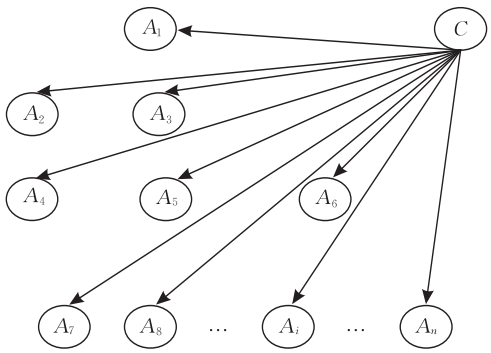
$$f_{\mathbb{B}}(\boldsymbol{a})=\arg\max_{c\in C}\{P(c)\prod_{i=1}^{n}P_{\mathbb{B}}(a_i\mid\boldsymbol{\pi}_i)\}\quad(2)$$

2.2 学习策略

贝叶斯网络分类器学习策略有两种：分别是生成式学习策略和鉴别式学习策略.生成式学习策略是先训练一个性能较好的贝叶斯网络，然后再将此网络应用于分类问题.鉴别式学习策略是直接生成一个性能较好的分类器.一般来讲，在分类器的分类精度方面，鉴别式学习策略比生成式学习策略得到的分类器更有竞争力，这是因为生成式学习策略过多地考虑了与类变量无关的属性变量之间的依赖关系.例如，图1给出了一个应用两种策略有可能得到的不同网络结构，图1(a)是应用生成式学习策略得到的，图1(b)是应用鉴别式式学习策略得到的.图1(a)是一个好的贝叶斯网络，却不一定是一个好的贝叶斯网络分类器；图1(b)是一个好的贝叶斯网络分类器，却也不一定是一个好的贝叶斯网络.



(a) 采用生成式学习策略得到的



(b) 采用鉴别式学习策略得到的

图1 贝叶斯网络分类器结构图

生成式学习策略通常是以最大化似然函数标准指导学习算法的设计，鉴别式学习策略通常以最大化条件似然函数标准指导学习算法的设计.两者的区别在于采用了不同的评估函数.

为了应用上的方便，似然函数常转换为等价的对数似然函数(Log Likelihood，LL)的形式.给定 m 个实例的数据集合 $D=\{\boldsymbol{a}_1,\cdots,\boldsymbol{a}_i,\cdots,\boldsymbol{a}_m\}$，则有对数似然函数表示为

$$LL(f_{\mathbb{B}}\mid D)=\sum_{i=1}^{m}\log(P_{\mathbb{B}}(\boldsymbol{a}_i))=$$

$$m\sum_{i=1}^{m}\sum_{\substack{a_i\in A_i\\ \boldsymbol{\pi}_i\in\boldsymbol{\Pi}_i}}\hat{P}_D(a_i,\boldsymbol{\pi}_i)\log(\theta(a_i,\boldsymbol{\pi}_i))\quad(3)$$

容易看出，当 $\theta(a_i,\boldsymbol{\pi}_i)=\hat{P}_D(a_i,\boldsymbol{\pi}_i)$ 时，对数似然函数取得极大值.因此可简单地用观察到的样本出现频率估计贝叶斯网络的参数(Observed-Frequency Estimation，OFE)[16].

鉴别式学习策略[15]直接以最大化条件似然函

数为标准优化分类器性能. 同样，条件似然函数也可以表示为对数条件似然函数(Log Conditional Likelihood，LCL)的形式，

$$LCL(f_{\mathbb{B}}\mid D)=\sum_{i=1}^{m}\log(P_{\mathbb{B}}(c_i\mid \boldsymbol{a}_i))=$$

$$m\sum_{i=1}^{m}\sum_{\substack{a_i\in A_i\\ \boldsymbol{\pi}_i\in \boldsymbol{\Pi}_i}}P_D(a_i,\boldsymbol{\pi}_i)\log(\theta(a_i\mid \boldsymbol{\pi}_i)) \tag{4}$$

虽然式(4)和式(3)表现形式类似，但当 $\theta(a_i,\boldsymbol{\pi}_i)=\hat{P}_D(a_i,\boldsymbol{\pi}_i)$ 成立时，只能保证对数似然函数取得极大值，却不能保证对数条件似然函数取得极大值，因此式(4)不能分解为线性形式或任何相近的形式直接计算.

3　网络结构与变量分布的关系

在应用中，由于很多实际原因，往往只能训练得到限制了网络结构形式的贝叶斯网络. 这种网络结构与实际数据变量的分布并不一致. 为了进一步提高分类器的性能，本文定量地定义了贝叶斯网络结构与实际变量分布之间的关系，方便具体的操作.

定义 1.　假设数据集合 D 符合某一联合概率分布 P，称两个贝叶斯网络$\mathbb{B}$和$\mathbb{B}'$相对于概率分布 P 是等价的，当且仅当

(1) $\mathbb{B}$和$\mathbb{B}'$中结点的数目与 P 中随机变量的数目相同，且一一对应；

(2) $\mathbb{B}$和$\mathbb{B}'$能准确地表示联合概率分布 P，记为$\mathbb{B}=|_P\mathbb{B}'$(简记为$\mathbb{B}=\mathbb{B}'$).

当$\mathbb{B}=\mathbb{B}'$时，说明$\mathbb{B}$和$\mathbb{B}'$关于概率分布 P 是等价的，但它们有可能表现出不同的结构形式. 这是因为在 P 中条件独立的两个变量在$\mathbb{B}$和$\mathbb{B}'$中对应的两个结点间可能存在边也可能不存在边.

定义 2.　对于一个关于联合概率分布 P 的贝叶斯网络，假设其中的每一对变量都存在着一条有向边，则称这个贝叶斯网络是联合概率分布 P 的一个完全贝叶斯网络，记为$\mathbb{B}_c$.

即便在联合概率分布 P 中条件独立的两个变量，在$\mathbb{B}_c$中对应的两个结点间也必须存在边，且所有这样的结点间都是相互弱连通的.

定义 3.　对于一个联合概率分布 P，所有描述它的贝叶斯网络集合记为 $\boldsymbol{BS}(P)=\{\mathbb{B}\mid \forall\mathbb{B}=|_P\mathbb{B}_c\}$.

定义 4.　在 $\boldsymbol{BS}(P)$中边数目最少的$\mathbb{B}$称为 P 的最简贝叶斯网络，记为$\mathbb{B}_t$，即$\mathbb{B}_t$中所有的边对描述数据集合 D 的联合概率分布 P 都是必不可少的.

一般地，在以往的研究中，用$\mathbb{G}$表示训练得到的网络结构，用$\mathbb{T}$表示符合数据集合 D 内变量实际联合概率分布 P 的网络结构. 将两者的关系描述为

$$\mathbb{G}\begin{cases}<\mathbb{T}, & \mathbb{G}\text{简单于}\mathbb{T}\\ \approx\mathbb{T}, & \mathbb{G}\text{近似于}\mathbb{T}\end{cases} \tag{5}$$

这里的$\mathbb{T}$可以视为$\mathbb{B}_t$. 当$\mathbb{G}$缺少足够的边来描述 D 中变量间的依赖关系时，记为$\mathbb{G}<\mathbb{T}$；当$\mathbb{G}$能够近似描述 D 时，记为$\mathbb{G}\approx\mathbb{T}$. 定义中所表达的“$\approx$”近似这一概念很难把握. 现在精确定义为，当$\mathbb{G}=\mathbb{B}_t$时，$\mathbb{G}=\mathbb{T}$，并追加当$\mathbb{G}\in\boldsymbol{BS}(P)$且$\mathbb{G}\neq\mathbb{B}_t$时，$\mathbb{G}>\mathbb{T}$，对于这种结构，又称其为存在冗余边的结构，简称冗余结构. 本文将两者的关系重新描述为

$$\mathbb{G}\begin{cases}<\mathbb{T}, & \mathbb{G}\notin\boldsymbol{BS}(P)\\ =\mathbb{T}, & \mathbb{G}=\mathbb{B}_t\\ >\mathbb{T}, & \mathbb{G}\in\boldsymbol{BS}(P)\text{ 且}\mathbb{G}\neq\mathbb{B}_t\end{cases} \tag{6}$$

图 2 是式(6)中 3 类关系的示例图，若图 2(a)表示与实际变量分布完全一致的贝叶斯网络分类器结构，则缺少边$\langle A_1,A_2\rangle$的图 2(b)表示比实际变量分布简单的贝叶斯网络分类器结构，而添加边$\langle A_1,A_3\rangle$的图 2(c)表示存在冗余结构的贝叶斯网络分类器结构.

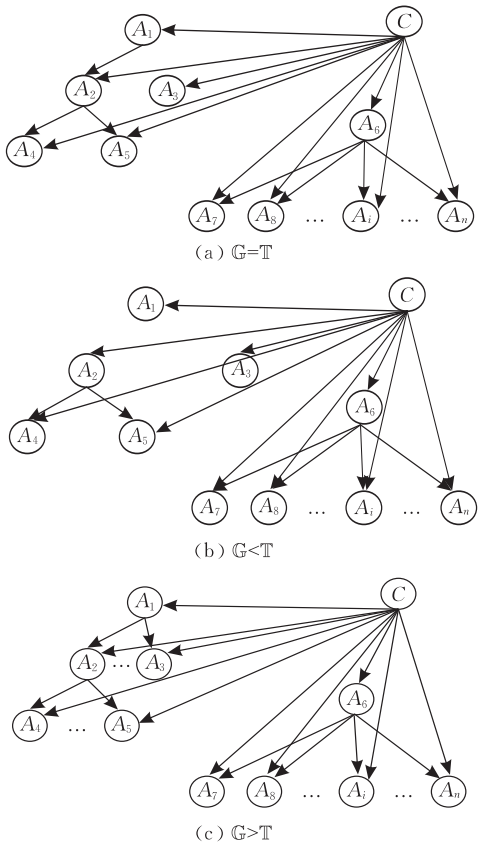


图 2　贝叶斯网络分类器结构与变量分布之间的 3 种关系示例图

因为在大多数应用问题中，$\mathbb{B}_c$是无意义的，$\mathbb{B}_t$是难以保证的，所以在分类任务中，通常采用固定父结点数目的贝叶斯网络结构. Greiner 等人[4]利用 ELR 算法进行了两类实验：第 1 类采用的网络结构简单于数据集实际分布的网络结构；第 2 类采用的网络结构接近于数据集实际分布的网络结构. 这些实验具有一定的局限性. 因为，即便对于结构简单的 NB 网络或 TAN 网络[2]，也可能由于存在多余的结点与冗余的边而导致其比实际结构复杂. 假设$\mathbb{G}=\{\mathbb{G}_1,\mathbb{G}_2,\mathbb{G}_3\}$，$\mathbb{T}=\{\mathbb{T}_1,\mathbb{T}_2,\mathbb{T}_3\}$，即$\mathbb{G}$由$\mathbb{G}_1$、$\mathbb{G}_2$和$\mathbb{G}_3$构成，$\mathbb{T}$由$\mathbb{T}_1$、$\mathbb{T}_2$和$\mathbb{T}_3$构成，设$\mathbb{G}_i$和$\mathbb{T}_i$分别包含相同的结点($i=1,2,3$)，且假设这 3 组子结构所含结点数目基本相同. 当$\mathbb{G}<\mathbb{T}$时，有可能是因为$\{\mathbb{G}_1<\mathbb{T}_1,\mathbb{G}_2<\mathbb{T}_2,\mathbb{G}_3<\mathbb{T}_3\}$，也有可能是因为$\{\mathbb{G}_1<\mathbb{T}_1,\mathbb{G}_2<\mathbb{T}_2,\mathbb{G}_3>\mathbb{T}_3\}$，$\{\mathbb{G}_1<\mathbb{T}_1,\mathbb{G}_2>\mathbb{T}_2,\mathbb{G}_3<\mathbb{T}_3\}$或$\{\mathbb{G}_1>\mathbb{T}_1,\mathbb{G}_2<\mathbb{T}_2,\mathbb{G}_3<\mathbb{T}_3\}$情况引起的. 很明显，后 3 种情况出现的概率更大. 通过以上讨论得知，在所采用的网络结构之中通常存在$\mathbb{G}<\mathbb{T}$，$\mathbb{G}=\mathbb{T}$或$\mathbb{G}>\mathbb{T}$情况.

应用贝叶斯网络结构与实际变量分布之间的关系，可以分析出，在一般的情况下，限制了父结点数目的贝叶斯网络结构中不能避免存在冗余的边. 下面提出一种不含有冗余边的贝叶斯网络分类器.

4 森林型贝叶斯网络分类器

森林型贝叶斯网络分类器 FAN 是一种限制网络形式的贝叶斯网络分类器应用模型，其限制策略包括两方面：首先，其结构的属性子网拓扑形式与 TAN 结构的属性子网近似，但比 TAN 结构更为灵活，内部结点可以有一个父结点或没有父结点，即整体上可以是一棵树，也可以是由多棵树组成的森林. 其次，类变量是属性子网中所有结点的父结点. 这样可以保证分类结果能够考虑到每一个特征变量的影响. FAN 分类器定义为函数 $f_{\mathrm{FAN}}(\cdot)$，

$$f_{\mathrm{FAN}}(\boldsymbol{a})=\arg\max_{c\in C}\{P(c)\prod_{i=1}^{n}P_{\mathrm{FAN}}(a_i\mid\boldsymbol{\pi}_i)\} \quad (7)$$

图 3 是一个 FAN 的结构图示意图. 从图中可以看出，FAN 是一种介于 NB 和 TAN 网络结构之间的限制性贝叶斯网络结构. 因为对于任一特征变量 $A_{NB}\in\mathbb{G}_{NB}$，其$|\boldsymbol{\Pi}_{\mathrm{NB}}|=1$，对于任一特征变量 $A_{\mathrm{TAN}}\in\mathbb{G}_{\mathrm{TAN}}$，$|\boldsymbol{\Pi}_{\mathrm{TAN}}|=2$，且对于任一特征变量 $A_{\mathrm{FAN}}\in\mathbb{G}_{\mathrm{FAN}}$，$1\leq|\boldsymbol{\Pi}_{\mathrm{FAN}}|\leq 2$，所以$|\boldsymbol{\Pi}_{\mathrm{NB}}|\leq|\boldsymbol{\Pi}_{\mathrm{FAN}}|\leq|\boldsymbol{\Pi}_{\mathrm{TAN}}|$，因此，可以将 NB 看做是结构最简单的 FAN，将 TAN 看做是结构最复杂的 FAN.

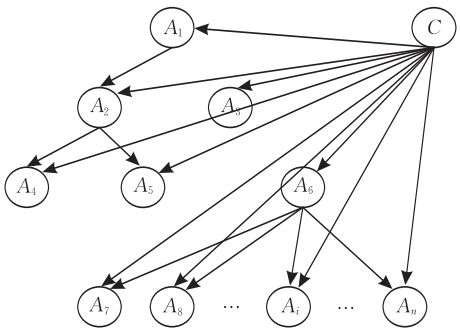


图 3 森林型贝叶斯网络分类器结构示意图

5 森林型贝叶斯网络分类器学习算法

这一节提出一种在现有贝叶斯网络分类器结构训练基础上学习森林型贝叶斯网络分类器的算法，该算法的关键是识别网络结构中的冗余边.

当贝叶斯网络结构中有冗余的边的时候，不仅仅是增加计算量的问题，甚至会误导参数学习，降低分类器的性能. 因而，常常需要识别网络结构中是否有冗余的边.

极大化对数条件似然函数的过程与提高贝叶斯网络分类器分类精度的过程是一致的. 梯度上升法以对数条件似然函数的导数为方向极大化函数值，所以对数条件似然函数的偏导数的计算与网络中结点的父结点集合关系紧密.

因此可以应用对数条件似然函数的偏导数与网络中结点的父结点集合之间的关系，消除网络结构中冗余的部分，提高鉴别式参数学习方法所得到的分类器精度. 条件概率表内的每一个参数都表示一个随机变量 A_i 和属性变量集合 $\boldsymbol{\Pi}_i$ 间的随机关系. 随机变量 A_i 是固定不变的，属性变量是一个集合，在贝叶斯网络中也就是 A_i 的父结点集合，当其中有结点的存在与否对对数条件似然函数的导数没有影响时，认为它们是冗余的信息，并记 A_i 的这些父结点为 $\mathrm{sub}(\boldsymbol{\Pi}_i)$. 即当

$$\left\langle\frac{\partial\widehat{LCL}^{(\mathbb{G})}(\boldsymbol{\Theta})}{\partial\theta(a_i\mid\boldsymbol{\pi}_i)}=\frac{\partial\widehat{LCL}^{(\mathbb{G})}(\boldsymbol{\Theta})}{\partial\theta(a_i\mid\boldsymbol{\pi}_i\backslash\mathrm{sub}(\boldsymbol{\pi}_i))}\right\rangle_{a_i\in A_i,\boldsymbol{\pi}_i\in\boldsymbol{\Pi}_i} \quad (8)$$

成立时，$\mathrm{sub}(\boldsymbol{\Pi}_i)$集合内所包含的结点不应成为变量 A_i 的父结点.

接下来，有两个技术方面的问题需要解决：一是如何计算对数条件似然函数的偏导数，二是以什么地方的导数值为判断的依据.

关于第 1 个问题，可以应用梯度方法鉴别式学习贝叶斯网络参数时的研究成果. 梯度方法研究中

提出的计算对数条件似然函数的偏导数的方法如下[4,17].

用一个贝叶斯网络分类器的网络表示这个数据集合，设条件概率中的各个参数为变量，并计算出这些变量的偏导数. 对于其中的一个参数 $\beta_{a|\pi}$，为了保证 $\beta_{a|\pi}\geq 0$ 且 $\sum_a \beta_{a|\pi}=1$，将其变形为参数“$\theta_{a|\pi}$”，使满足

$$\beta_{a|\pi}=\frac{e^{\theta_{a|\pi}}}{\sum_{a'} e^{\theta_{a'|\pi}}},$$

可得出，当 θ 的定义域是实数时，β 能够满足要求.

在梯度算法运行中的任一步，参数 θ 的变化方向可以表示为对数条件似然函数的偏导数，因而

$$\nabla \widehat{LCL}=\left\langle \frac{\partial \widehat{LCL}^{(\mathbb{G})}(\boldsymbol{\Theta})}{\partial \theta(a_i\mid \boldsymbol{\pi}_i)}\right\rangle_{a_i,\boldsymbol{\pi}_i},$$

结点 A_i 在取值为 a_i 且父结点集合取值为 $\boldsymbol{\pi}_i$ 时，对应于数据集合 D 的导数为

$$\frac{\partial \widehat{LCL}^{(\mathbb{G})}(\boldsymbol{\Theta})}{\partial \theta(a_i\mid \boldsymbol{\pi}_i)}=\sum_{\langle a_1,a_2,\cdots,a_n,c\rangle\in D}\frac{\partial \widehat{LCL}^{(\langle a_1,a_2,\cdots,a_n,c\rangle)}(\boldsymbol{\Theta})}{\partial \theta(a_i\mid \boldsymbol{\pi}_i)} \tag{9}$$

对于任一训练实例 $\langle \boldsymbol{a}',c'\rangle$，关于参数 $\theta(a_i|\boldsymbol{\pi}_i)$ 的偏导数即为

$$\frac{\partial \widehat{LCL}^{(\langle \boldsymbol{a}',c'\rangle)}(\boldsymbol{\Theta})}{\partial \theta(a_i\mid \boldsymbol{\pi}_i)}=[P_{\boldsymbol{\Theta}}(a_i,\boldsymbol{\pi}_i\mid a',c')-P_{\boldsymbol{\Theta}}(a_i,\boldsymbol{\pi}_i\mid a')]-\theta(a_i\mid \boldsymbol{\pi}_i)[P_{\boldsymbol{\Theta}}(\boldsymbol{\pi}_i\mid a',c')-P_{\boldsymbol{\Theta}}(\boldsymbol{\pi}_i\mid a')] \tag{10}$$

将式(10)代入式(9)可以计算出函数各个参数的偏导数.

关于第 2 个问题，可以在最大似然函数估计参数之处求偏导数，来判断当时的网络结构与条件对数似然函数的关系.

为了发现这些冗余结构，需要寻找在对数条件似然函数最大化过程中冗余结构对函数导数产生影响的位置，计算其导数值，进而应用导数性质分析出冗余部分. 可以从对数条件似然函数公式的变形看出

$$\begin{aligned}LCL(f_{\mathbb{B}}\mid D)&=\sum_{\langle a_1,a_2,\cdots,a_n,c\rangle\in D}P_D(a_1,a_2,\cdots,a_n,c)\cdot \log P_{\mathbb{B}}(c\mid a_1,a_2,\cdots,a_n)\\&=\sum_{\langle a_1,a_2,\cdots,a_n,c\rangle\in D}P_D(a_1,a_2,\cdots,a_n,c)\cdot \log\frac{P_{\mathbb{B}}(a_1,a_2,\cdots,a_n,c)}{P_{\mathbb{B}}(a_1,a_2,\cdots,a_n)}\\&=\sum_{\langle a_1,a_2,\cdots,a_n,c\rangle\in D}P_D(a_1,a_2,\cdots,a_n,c)\cdot \log P_{\mathbb{B}}(a_1,a_2,\cdots,a_n,c)-\\&\quad\sum_{\langle a_1,a_2,\cdots,a_n\rangle\in D}P_D(a_1,a_2,\cdots,a_n)\cdot \log P_{\mathbb{B}}(a_1,a_2,\cdots,a_n).\end{aligned}$$

因为等号右边第一项为对数似然函数[2]

$$LL(\mathbb{B}\mid D)=\sum_{\langle a_1,a_2,\cdots,a_n,c\rangle\in D}P_D(a_1,a_2,\cdots,a_n,c)\cdot \log P_{\mathbb{B}}(a_1,a_2,\cdots,a_n,c),$$

所以，对数条件似然函数可表示为

$$LCL(f_{\mathbb{B}}\mid D)=LL(\mathbb{B}\mid D)-\sum_{\langle a_1,a_2,\cdots,a_n\rangle\in D}P_D(a_1,a_2,\cdots,a_n)\cdot \log P_{\mathbb{B}}(a_1,a_2,\cdots,a_n).$$

上式可以将对数条件似然函数理解为调整后的对数似然函数. 因此，最大化对数条件似然函数的过程就是在最大对数似然的基础上，最小化

$$\sum_{\langle a_1,a_2,\cdots,a_n\rangle\in D}P(a_1,a_2,\cdots,a_n)\log P_{\mathbb{B}}(a_1,a_2,\cdots,a_n).$$

所以在对数条件似然函数最大化过程中，可以应用最大似然函数估计参数之处做判断网络结构对对数条件似然函数导数有影响的点求偏导数.

根据以上分析，本文首次基于对数条件似然函数的偏导数提出了一种森林型贝叶斯网络分类器学习算法 FAN，该算法能够识别网络中冗余的边，并删除其表示的本不存在的依赖关系. 算法的详细描述见算法 1 所示.

算法 1.　FAN 算法.

输入：训练数据集合 D

输出：一个贝叶斯网络分类器 f_{FAN}

1. 应用 TAN 算法训练一个贝叶斯网络结构$\mathbb{G}$.
2. 应用 OFE 算法估计贝叶斯网络结构$\mathbb{G}$的网络参数，生成贝叶斯网络$\mathbb{B}$.
3. 分别以条件概率表中各项为变量，计算对数条件似然函数的偏导数.
4. 对于贝叶斯网络$\mathbb{B}$中的每一个结点对应的属性变量 A_i.
5. 对于属性变量 A_i 的在贝叶斯网络$\mathbb{B}$中的父结点集合 $\boldsymbol{\Pi}_i$ 内的每一个结点对应的变量 A_j.
6. 如果将 A_j 放入集合 $\mathrm{sub}(\boldsymbol{\Pi}_i)$ 后，对于新的贝叶斯网络$\mathbb{B}'$仍然满足式(8)，则$\mathbb{B}:=\mathbb{B}'$.
7. 应用鉴别式参数学习算法重新估计$\mathbb{B}$的网络参数，生成一个贝叶斯网络分类器 f_{FAN}.
8. 返回贝叶斯网络分类器 f_{FAN}.

FAN 算法步 1 根据训练数据集合生成一个贝叶斯网络结构；步 2 采用生成式方法估计条件概率表的初始值；步 3 计算对数条件似然函数对应于各个参数项的偏导数；步 4～步 6，对贝叶斯网络参数中每一项都进行考察，若其父结点集合存在对对数

条件似然函数相对于这一项的偏导数无影响的结点,则删除;步 7 重新应用鉴别式参数学习算法估计网络参数,生成贝叶斯网络分类器;步 8 返回训练好的贝叶斯网络分类器.

FAN 算法是一种应用条件对数似然函数偏导数设计的 FAN 分类器学习算法,该算法能够使贝叶斯网络结构更适合于鉴别式参数学习算法性能的发挥.虽然鉴别式参数学习方法比生成式参数学习方法适应性更强,但要求采用的网络结构不能比实际网络结构复杂,这个问题看似容易解决,但目前的结构学习算法都很难保证,因而需要剔除传统方法生成的贝叶斯网络结构中冗余的边.

6 实 验

许多研究通常假设贝叶斯网络结构中不存在冗余的边.然而,这种假设在大多数现实问题中具有不同程度的局限性.可以验证当网络结构近似于或简单于实际数据变量分布时,鉴别式参数学习策略比生成式参数学习策略得到的分类器分类精度高.因此,只需验证在更广泛的情况下,鉴别式参数学习策略比生成式参数学习策略得到的分类器分类精度低,就可以得出通常情况下贝叶斯网络结构中是存在冗余的边的结论.由于基于梯度方法的鉴别式训练算法的运算量大,其实验结果只是在小数据集合上得到了验证.

对于自动训练算法,较难学习到恰好符合实际分布的贝叶斯网络结构,所以不易确定网络结构是简单于还是复杂于实际的分布.因此,实验中需要以某一实际联合概率分布的近似描述为基准,删除一些边,生成较简单的网络结构.这里采用 TAN 结构做为基准网络,没有采用 NB 结构,是因为在 TAN 结构中每个属性变量的父结点集合除了类变量之外还能有一个属性变量,但 NB 网络结构中的属性变量只有类变量一个父结点,若采用 NB 网络结构为基础进行分析则类同于属性选择.这样设计的实验与以往的实验不同,以往多是直接假设 TAN 网络结构简单于或近似于实际的分布进行分析的,比如 Greiner 等人[4]在验证 ELR 算法理论时的第一组实验同时采用了 NB 结构和 TAN 结构,认为这两种结构一般都比实际数据集合的分布简单.比较而言,在基准结构上删除边后作为简单的网络结构来应用更为准确.

在实验中,使用 DFE 算法进行鉴别式参数学习,这是因为 ELR 算法训练速度太慢,一般只适用于小型数据集.为了扩大本文的实验范围以及说明属性变量数目与算法的关系,在实验中采用更为有效的 DFE 算法.DFE 算法的分类精度一般与 ELR 算法基本相当[16].

表 1 实验数据集合

序号	数据集合	实例	类值	属性	缺损值
1	Audiology	226	24	69	Yes
2	Chess	551	2	39	No
3	Connect-4	67557	3	42	No
4	Lung Cancer	32	3	56	Yes
5	Lymphography	148	4	18	No
6	Primary Tumor	339	22	17	Yes
7	Soybean	683	19	35	Yes
8	Solar Flare	1389	2	9	No
9	Solar Flare-x	1389	3	10	No
10	Splice Junction Gene	3177	3	60	No
11	Zoology	101	7	16	No
12	Adult	48842	2	14	Yes
13	Annealing Processes	898	6	38	Yes
14	Balance Scale	625	3	4	No
15	Breast Cancer(Wisconsin)	699	2	9	Yes
16	Echocardiogram	131	2	6	Yes
17	Glass IdentificationA	214	7	10	No
18	Glass Identification	214	7	9	No
19	Heart	270	2	13	No
20	Hepatitis Prognosis	155	2	19	Yes
21	Hypothyroid Diagnosis	3163	2	25	Yes
22	Ionosphere	351	2	34	Yes
23	Iris Classification	150	3	4	No
24	Labor Negotiations	57	2	16	Yes
25	New-thyroid	215	3	5	No
26	Pen Digits	10992	10	16	No
27	Pima Indians Diabetes	768	2	8	No
28	Satellite	6435	6	36	No
29	Shuttle	58000	7	9	No
30	Sign	12546	3	8	No
31	Sonar Classification	208	2	60	No
32	Syncon	600	6	60	No
33	Wine Recognition	178	3	13	No

实验是在 Weka-3-4-11[18]系统平台之上设计与实现的.数据集合来源于 UCI 的数据库站点①.考虑到在属性变量较多的数据集合上训练的网络结构可能出现冗余边的概率较大,为了强调鉴别式参数学习策略在有冗余边的结构上训练的分类器的性能,选择的数据集合在本文参考文献之中所使用过的基础上适量提高了属性变量多的集合的比重,Greiner 等人[4]以 Friedman 等人[2]的实验数据为基础做的实验,Friedman 等人[2]共采用了 25 个数据集合,数据集合的属性变量数目范围从 4 个～36 个,平均每

① Asuncion A, Newman D J. UCI Machine Learning Repository. http://www.ics.uci.edu/~mlearn/MLRepository.html.

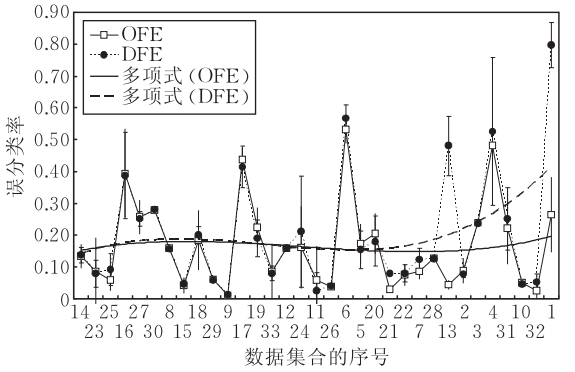
个数据集合拥有属性变量的数目为 15.68，本文采用了 33 个数据集合，数据集合的属性变量数目范围从 4 个～69 个，平均每个数据集合拥有属性变量的数目为 23.85. 很明显，Greiner 等人[4] 和 Friedman 等人[2] 实验选择的数据集合维度较低，代表性不足. 表 1 列出了这 33 个数据集的具体信息，包括实例数目、类值数目、属性变量数目以及是否有缺损值. 其中前 11 个数据集合只有名称型数据，其它数据集含有连续型属性. 实验使用 Weka 系统中无监督的离散化方法进行预处理. 另外，有 12 个数据集合具有缺损值，在实验中并没有进行特殊处理.

表 2 列出了应用 OFE 算法估计参数（简称 OFE）得到的分类器的误分类率、应用 DFE 算法估计参数（简称 DFE）得到的分类器的误分类率和 FAN 算法得到的分类器的误分类率.

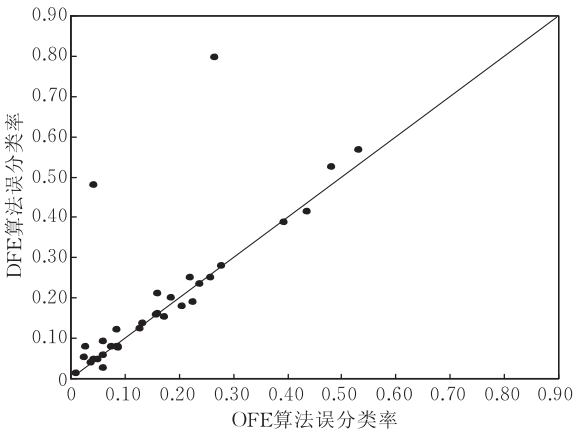
表 2　实验结果

No.	误分类率/%		
	OFE	DFE	FAN
1	26.4602±2.3705	79.7345±1.4132	24.1593±0.7404
2	8.8203±0.7973	7.6951±0.2069	8.0218±0.1519
3	23.7737±0.0143	23.5484±0.0058	23.5750±0.0097
4	48.1250±3.5630	52.5000±4.6351	46.2500±1.3975
5	17.2973±1.2274	15.4054±1.2087	13.9189±0.7704
6	53.2153±0.8751	56.7552±0.8499	56.2242±0.9916
7	8.6383±0.5279	12.2694±0.7423	9.8683±0.2450
8	15.8963±0.2061	15.7955±0.1205	15.7955±0.1205
9	0.9791±0.0394	1.3391±0.0644	1.3391±0.0644
10	5.0299±0.0784	4.7026±0.0359	4.6900±0.0385
11	5.9406±2.1003	2.5743±1.1289	2.3762±0.5423
12	15.9719±0.0634	16.0100±0.0363	16.0080±0.0383
13	4.3207±0.2882	48.0401±1.8677	25.0557±0.7130
14	13.3760±0.7640	13.7920±0.4853	13.7280±0.5355
15	4.1774±0.5098	4.6352±0.2394	3.8627±0.1431
16	39.3893±2.7839	38.7786±2.7204	33.1298±1.0241
17	43.6448±0.7819	41.4953±1.2967	41.3084±1.4251
18	18.5047±1.8866	20.0000±0.5119	9.0654±1.0238
19	22.4445±1.2450	18.9630±1.1535	17.8518±1.7646
20	20.5161±1.2410	18.0645±1.5803	14.3226±1.6699
21	2.8517±0.1330	7.9861±0.1109	7.9229±0.1918
22	7.4644±0.6802	8.0342±0.5480	8.1481±0.5908
23	8.6667±2.1082	7.8666±0.8692	7.3333±0.8165
24	16.1404±2.6022	21.0526±3.5088	14.0351±2.7739
25	5.9535±0.6899	9.2093±1.0083	7.3489±0.2080
26	3.7391±0.0863	3.8828±0.1100	3.8664±0.0826
27	25.7292±1.0682	25.1042±0.4640	24.0625±0.4159
28	12.7055±0.1687	12.4911±0.0543	12.4911±0.0543
29	6.0069±0.0024	5.8642±0.0028	5.8642±0.0028
30	27.8973±0.1684	27.9946±0.1581	27.9946±0.1581
31	22.1154±2.2292	25.0961±1.9647	24.1346±1.2900
32	2.5333±0.5452	5.3333±0.4859	1.1000±0.3028
33	8.5394±1.7497	7.9776±0.4700	3.8202±0.4700

图 4 是采用 OFE 算法和 DFE 算法得到的贝叶斯网络分类器误分类率的曲线图和散点图. 在误分类率曲线图中，横坐标是数据集合的序号，按其属性变量数目从少到多排列；纵坐标表示这两个算法生成的贝叶斯网络分类器在各个数据集上的平均误分类率的大小. DFE 算法生成的分类器的误分类率多项式趋势线用实线表示，OFE 算法生成的分类器的误分类率多项式趋势线用虚线表示.



（a）误分类率曲线图



（b）误分类率散点图

图 4　OFE 算法与 DFE 算法比较的误分类率曲线图和散点图

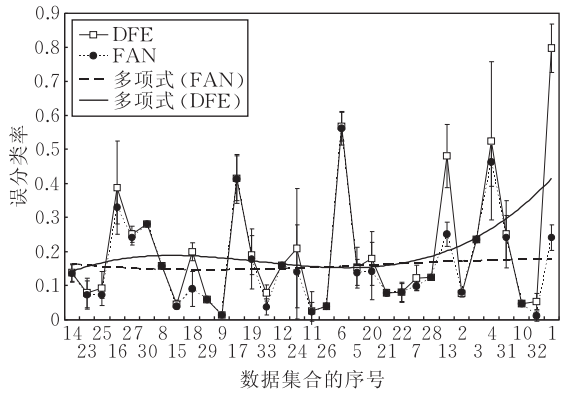
从图 4(a)可以看出，在数据集合的属性变量数目较少时，这两个算法生成的分类器的误分类率并无明显差异. 随着属性变量数目的增加，DFE 算法的误分类率有增长的趋势，而 OFE 算法的误分类率仍无显著变化. 在图 4(b)所表示的误分类率散点图内，除了在两个属性变量数目较多的数据集合之上应用 DFE 算法的效果较差之外，在其它数据集合之上的分布基本在对角线附近. 统计结果表明，在 18 个数据集上 DFE 算法生成的贝叶斯网络分类器的平均误分类率比 OFE 算法的高，进一步对之进行双尾配对 t 检验，可知有 10 个数据集合在显著性水平 0.05 下误分类率显著上升；在另外 15 个数据集上，DFE 算法的平均误分类率比 OFE 算法的低，同样进行双尾配对 t 检验，有 6 个在显著性水平 0.05 下误分类率显著下降. 这就说明在 27 个数据集上基于

鉴别式学习策略的 DFE 算法生成的分类器的误分类率没有显著下降，即两个算法的性能基本接近. Greiner 等人[4]已经验证了在$\mathbb{G}<\mathbb{T}$和$\mathbb{G}\approx\mathbb{T}$情况下，鉴别式方法生成的分类器性能优于生成式方法生成的分类器性能. 但在 27 个数据集上鉴别式方法生成的分类器性能并不优于生成式方法生成的分类器性能，所以这些网络中可能包含$\mathbb{G}>\mathbb{T}$的情况.

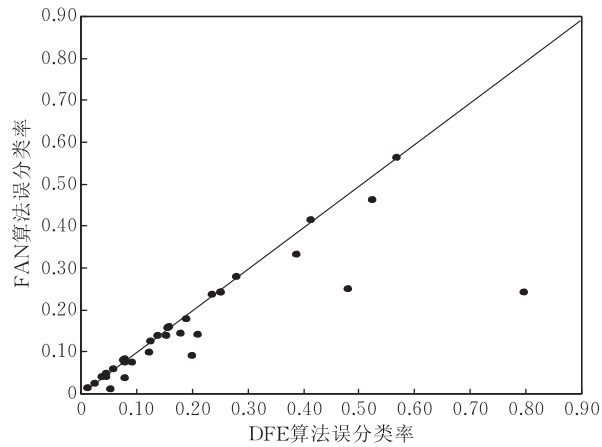
显然，本文实验中采用的大部分数据集合生成的贝叶斯网络结构存在冗余的部分，并不像通常认为的，学到的网络结构比实际的复杂只是个别现象，并说明了鉴别式参数学习策略对限制性贝叶斯网络结构的修正是有限的. 因此，FAN 分类器是一种比 TAN 分类器更适合鉴别式参数训练的分类模型.

FAN 算法运行中，式(8)是判断随机变量 A_i 的父结点集合内是否有多余结点的理论依据. 然而，现实数据集合中既可能有噪声又可能有缺损数据. 因此，在实验中设置了一个误差范围(± 0.5)，当比较结果在这个范围内时，就认为式(8)成立. 不同的随机变量可能存在不同的父结点，并且这些父结点具有多种的取值组合形式. 随着父结点数目的增多，组合方式就会迅速增加，出现差异的参数个数也就随之迅速增多. 为了控制父结点数目的增多对度量结果的影响，实验中以各个父结点产生的差异率作为比较标准，即相对于具有相应父结点的不同参数的导数个数与随机变量 A_i 的父结点取值数目乘积的比值. 结合前面的分析，当比值小于设定的阈值时，就认为式(8)成立，以此作为识别结构中是否有冗余边的依据.

图 5 是我们的 FAN 算法和 DFE 算法所得到的贝叶斯网络分类器误分类率的曲线图和散点图. 误分类率曲线图的坐标同上. DFE 算法的误分类率多项式趋势线用实线表示，而 FAN 算法的误分类率多项式趋势线用虚线表示. 从图 5(a)可以看出，FAN 算法误分类率多项式趋势线较为平稳，受数据集合的属性变量数目变化的影响较小；而除了具有中小规模属性变量数目的数据集之外，DFE 算法性能均不如 FAN 算法. 从图 5(b)也可看出，FAN 算法生成的分类器的误分类率整体上比 DFE 算法的低. 在 25 个数据集上 FAN 算法比 DFE 算法得到的贝叶斯网络分类器的平均错误率都有所降低，占总数据集合数目的 75.76%. 在显著性水平 0.05 之下的双尾配对 t 检验结果表明，FAN 算法在 12 个数据集合之上的误分类率都有显著降低，占总数据集合数目的 36.36%. 这就验证基于对数条件似然函数偏导数的变量间依赖关系度量方法，能够在贝叶斯网络结构中辨别通常的度量方法不能有效识别的冗余边，所设计的森林型贝叶斯网络分类器在多数实际应用中能取得较好的性能.



(a) 误分类率曲线图



(b) 误分类率散点图

图 5 FAN 算法与 DFE 算法比较的误分类率曲线图和散点图

7 总结与展望

本文研究了贝叶斯网络在分类问题中的应用. 定量描述了网络结构与实际数据变量分布之间的关系，指出当网络结构比实际分布复杂时，即网络中存在冗余边时，将会限制鉴别式参数学习策略性能的发挥，进而提出一种不含有冗余边的森林型贝叶斯网络分类器和学习算法 FAN. FAN 算法首次应用了对数条件似然函数的偏导数来优化算法的设计. 最后，用实验验证了已有的限制性贝叶斯网络分类器中普遍存在冗余边；在有冗余边的网络结构上，鉴别式学习策略的性能弱于生成式学习策略的，这就说明了本文所提 FAN 分类器的应用价值；同时，验证了 FAN 算法的有效性，说明了基于对数条件似然函数设计贝叶斯网络分类器是一种有效提高分类

器性能的途径.

本文下一步将研究鉴别式参数学习策略不适合在有冗余边的网络结构上训练的原因，分别从两方面进行探索：第 1 个原因可能是因为，在贝叶斯网络分类器结构中，随着冗余边增多，进行参数估计时需要的训练数据数目也将增多，若训练数据固定，分类器参数估计的可靠性将下降；第 2 个原因可能是因为条件对数似然函数是不可分解的，在利用爬山算法逐步逼近最优解时，若网络结构中冗余边增多，有可能增加爬山算法陷入局部最优的可能性.

致　谢　本文研究得到了田盛丰教授的热情指导. 田教授在本文研究进程中，提出了很多宝贵的建议，开拓了我们的研究思路. 在此表示衷心的感谢. 同时也感谢匿名审稿人对本文提出的宝贵意见!

参　考　文　献

[1] Han J, Kamber M. Data Mining: Concepts and Techniques. 2nd Edition. San Francisco, CA: Morgan Kaufmann, 2005

[2] Friedman N, Geiger D, Goldszmidt M. Bayesian network classifiers. Machine Learning, 1997, 29(2/3): 131-163

[3] Greiner R, Zhou W. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers//Proceedings of the 18th Annual National Conference on Artificial Intelligence (AAAI 2002). Edmonton, Canada, 2002: 167-173

[4] Greiner R, Su X, Shen B et al. Structural extension to logistic regression: Discriminative parameter learning of belief net classifiers. Machine Learning, 2005, 59(3): 297-322

[5] Chickering D M, Heckerman D, Meek C. Large-sample learning of Bayesian networks is NP-hard. The Journal of Machine Learning Research, 2004, 5: 1287-1330

[6] Tillman R E. Structure learning with independent non-identically distributed data//Proceedings of the 26th Annual International Conference on Machine Learning. New York, 2009: 1041-1048

[7] Shi Hong-Bo, Wang Zhi-Hai, Huang Hou-Kuan et al. A restricted double-level Bayesian classification model. Journal of Software, 2004, 15(2): 193-199(in Chinese)
(石洪波，王志海，黄厚宽等. 一种限定性的双层贝叶斯分类模型. 软件学报，2004，15(2)：193-199)

[8] Zheng Z, Webb G I. Lazy learning of Bayesian rules. Machine Learning, 2000, 41(1): 53-84

[9] Keogh E J, Pazzani M J. Learning the structure of augmented Bayesian classifiers. International Journal on Artificial Intelligence Tools, 2002, 11(4): 587-601

[10] Webb G I, Boughton J R, Wang Z. Not so naive Bayes: Aggregating one-dependence estimators. Machine Learning, 2005, 58(1): 5-24

[11] Näele A, Dejori M, Stetter M. Bayesian substructure learning — Approximate learning of very large network structures//Proceedings of the 18th European Conference on Machine Learning (ECML, 2007). Warsaw, Poland, 2007: 238-249

[12] Pernkopf F, Bilmes J. Discriminative versus generative parameter and structure learning of Bayesian network classifiers//Proceedings of the 22nd International Conference on Machine Learning (ICML 2005). Bonn, Germany, 2005: 657-664

[13] Pernkopf F. Discriminative learning of Bayesian network classifiers//Proceedings of the 25th IASTED International Multi-Conference. Innsbruck, Austria, 2007: 422-427

[14] Jing Y, Pavlovic V, Rehg J M. Efficient discriminative learning of Bayesian network classifier via boosted augmented naive Bayes//Proceedings of the 22rd International Conference on Machine Learning(ICML 2005). Bonn, Germany, 2005: 369-376

[15] Jing Y, Pavlovic V, Rehg J M. Boosted Bayesian network classifiers. Machine Learning, 2008, 73(2): 155-184

[16] Su J, Zhang H, Ling C X et al. Discriminative parameter learning for Bayesian networks//Proceedings of the 25th International Conference on Machine Learning (ICML 2008). Helsinki, Finland, 2008: 1016-1023

[17] Darwiche A. A differential approach to inference in Bayesian networks. Journal of the ACM, 2003, 50(3): 280-305

[18] Witten I H, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann, 2005

WANG Zhong-Feng, born in 1977, Ph. D. candidate. His research interests include data mining and pattern recognition.

WANG Zhi-Hai, born in 1963, professor, Ph. D. supervisor. His research interests include data mining and machine learning.

Background

Supervised learning is an important task in machine learning and data mining, assigning predefined class labels to data items described by means of a set of features or attributes. A classifier is a function that maps an instance into a

class label. The problem of the automatic induction of classifiers from data sets of pre-classified instances has long received considerable attention within the machine learning community. Bayesian network is an effective approach to this problem and has also been successfully applied in many ways by inducing classifiers using different types of Bayesian network learning algorithms. However, learning (the most probable a posteriori under certain conditions) Bayesian network from data is an NP-hard problem.

There are many learning algorithms for automatically building restricted Bayesian networks from data. Some of these are based on testing conditional independences, others are based on the so-called score+search paradigm. Recently, some researches show that discriminative parameter learning strategies are good at restricted Bayesian network structures. This paper is interested in optimizing Bayesian classifier on this field. We indicate that there are always maintaining a portion which is more complex than truth, even in the simplest Bayesian network structures. And discriminative parameter learning strategy is not feasible to the case above. In order to enhance the performance of classifier, we propose an algorithm for simplifying structure based on the relationship between redundant edges in network structure and gradient orientation of conditional log likelihood function.

The work is supported by National Natural Science Foundation of China (60673089, 60973011). From the viewpoint of real applications, this project focuses on more efficient and accurate learning techniques based on restricted Bayesian network classifiers, and tastes to actually apply them to large-scale databases. The mainly details are: (1) Research on how to specify directions of dependent relations efficiently and effectively in Bayesian networks; (2) In the space of discrete random variables, focus on the algorithms for solving a near-maximum spanning forest based on directed graph with asymmetric dependent relations, and the mechanizes for multi-level dependences transformation; (3) Develop on semi-lazy learning techniques and their applications; and (4) Research on the applications of emerging pattern techniques in building a restricted Bayesian network. These results will be meaningful not only for learning techniques of restricted Bayesian network, but also for improvements on other learning models. At the same time, they would be more important for the problems of general network optimization both on theoretical aspects and real applications. Until now, the research team has published more than 20 papers about the restricted Bayesian networks learning strategies.