压缩候选的贝叶斯信念网络构造算法

Algorithm of Bayesian Belief Network Structure of Compressed Candidature

杨本良

Yang Benliang

(**广西梧州市商务局,广西梧州** 543000)

(Wuzhou Business Bureau of Guangxi, Wuzhou, Guangxi, 543000, China)

摘要:针对传统算法分类速度较慢的不足,改进传统算法中候选变量的搜索方式,提出用依赖度量函数测量变量之间的依赖程度,得出压缩候选的贝叶斯信念网络构造算法.该算法在不影响原有算法可靠性的前提下,提高了学习速度.

关键词:贝叶斯信念网络 压缩候选 算法 数据挖掘

中图法分类号:TP311 **文献标识码**:A **文章编号**:1002-7378(2005)04-0207-02

Abstract: A learning algorithm of compressed candidates based on Bayesia belief network is developed to solve slow running problem of traditional Bayesian belief network constructing algorithm. The improved method for searching candidates with a modified dependence measure is used in the presented algorithm which can speed up the study process without sacrificing the reliability of the traditional method.

Key words: Bayesian belief network, compressed candidature, algorithm, data mining

随着数据库技术的广泛应用,各行各业都积累了海量的数据.这些海量数据所隐含的内在联系有可能是有价值的知识,如何发现、提取这些知识和规则并加以利用就成了当务之急.数据挖掘就是从大量的数据中提取隐含的、未知的、对决策有潜在价值的知识和规则的过程,贝叶斯方法则是分类规则挖掘的常用方法[1].贝叶斯信念网络说明了联合条件分布,它允许在变量的子集之间定义类条件独立性,并提供一种因果关系图形.它克服了朴素贝叶斯分类方法无法定义变量之间的依赖关系的弱点.但传统的贝叶斯信念构造算法在对海量数据分类时速度较慢,针对传统算法的不足,本文改进了传统算法中候选变量的搜索方式,提出用依赖度量函数测量变量之间的依赖程度,得到压缩候选的贝叶斯信念网络构造算法,该算法在不影响传统算法性能的基础上,极大地提高了运算速度.

1 贝叶斯信念网络

定义1 给定一个随机变量集 $\chi=\{X_1,X_2,\cdots,X_n\}$,贝叶斯信念网络说明 χ 上的一条联合条件概率分布,其中 X_i 是一个 m 维向量.贝叶斯信念网络定义为:

$$B=\langle G,\theta\rangle, \tag{1}$$

其中,G 是一个有向无环图,其顶点对应于有限集 χ 中的随机变量 $X_1,X_2,\cdots,X_n$;其弧代表一个函数依赖关系.如果有一条弧由变量 Y 到 X,则 Y 是 X 的双亲或者直接前驱,而 X 则是 Y 的后继.一旦给定其双亲,无环图中的每个变量独立于图中该节点的非后继.在图 G 中 X_i 的所有双亲变量用集合 $Pa(X_i)$.

(1) 式中,θ 代表用于量化网络的一组参数.对于每一个 X_i,$Pa(X_i)$ 的取值 x_i,存在

$$\theta_{x_i|pa(x_i)}=P(x_i|pa(X_i)),$$

它指明了在给定 $Pa(X_i)$ 发生的情况下 x_i 事件发生的条件概率.因此实际上一个贝叶斯信念网络给定了变量集合 χ 上的联合条件概率分布:

$$P_B(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P_B(X_i|Pa(X_i)).$$

因此,贝叶斯网络构造算法可以表示如下:给定一组训练样本 $D=\{x_1,x_2,\cdots,x_n\}$,x_i 是 X_i 的实例,寻找一个最匹配该样本的贝叶斯信念网络.常用的学习算法通常是引入一个评估函数 $S(B|D)$,用

收稿日期:2005-05-19
修回日期:2005-08-05
作者简介:杨本良(1971-),男,助理工程师,主要从事数据库、数据仓库与数据挖掘研究。

该函数来评估每一个可能的网络结构与样本之间的契合度，并从所有这些可能的网络结构中寻找一个最优解.

2 压缩候选的贝叶斯网络结构算法

传统的贝叶斯网络构造算法寻找网络结构时，要从 $n-1$ 个候选节点中逐一搜索 X_i 的父亲变量，由于没有考虑元素之间的相互联系，花费了大量时间搜索那些极不合理的候选变量. 如对蕴涵式：

$X \longrightarrow Y \longrightarrow Z$，

我们不难发现 X 和 Y、Y 和 Z、X 和 Z 之间存在依赖联系. 但是一旦考虑 X 和 Y 都是 Z 的父节点，如果将 Y 视为 Z 的父节点，X 对于 Z 的发生没有任何帮助. 基于上述思想，本提出了压缩候选的方法. 可以通过一个依赖度量函数 $I(X,Y)$ 来测量两个变量之间的依赖程度，$I(X,Y)$ 越大，说明变量 X 和 Y 之间联系越强，X 和 Y 也就越有可能存在父子关系；$I(X,Y)$ 很小，就说明 X 和 Y 互为父子关系的可能性越小. 为了计算变量之间的依赖关系，可以在选择父节点时不是选择所有的变量，而是集中扫描那些最可能是 X_i 的父亲的变量集 $Y_{i1},Y_{i2},\cdots,Y_{ik}, k \ll n$.

根据上述思想，本文提出了压缩候选的贝叶斯信念网络构造算法：

输入：训练样本集 $D=\{x^1,x^2,\cdots,x^N\}$，

初始化网络 B_0，

评价函数 $S(B|D)=\sum_i S(X_i|Pa(X_i),D)$，

参数 k.

输出：最优网络，$(1,2,\ldots,n)$

（Ⅰ）压缩. 根据 D 和 B_{n-1}，使用候选压缩，从 $X_1,\cdots,X_n$ 中为 X_i 选择一个候选父集 $C_i^n(|C_i^n|\leqslant k)$，定义一个有向图 $H_n=(\chi,E)$，其中 $E=\{X_j \longrightarrow X_i|\forall\ i,j,X_j\in C_i^n\}$.

（Ⅱ）最大化. 从中寻找一个能够最大化评价函数 $S(B_n|D)$ 的贝叶斯网络 $B_n=\langle G_n,\Theta_n\rangle$，其中，$G_n\subset H_n$，$\forall\ X_i, Pa^{G_n}(X_i)\subseteq C_i^n$.

（Ⅲ）返回 B_n.

压缩候选的贝叶斯信念网络构造算法最关键的一步就是：在计算评价函数中利用候选压缩算法对 X_i 可能的父亲集 $Pa(X_i)$ 进行筛选，从中选出 k 最有可能成为 X_i 的父亲的变量. 为了计算变量之间的联系紧密度，引入了依赖度量函数 $I(X,Y)$：

$$I(X,Y)=D_{KL}(P(X,Y)|P(X)P(Y)),$$

其中 $D_{KL}(P(X)|Q(X)=\sum_X P(X)\log\frac{P(X)}{P(Y)}$.

使用依赖度量函数后的候选压缩算法如下：

输入：数据集 $D=\{x^1,x^2,\cdots,x^N\}$，

一个贝叶斯网络 B_n，

权值计算函数 $S(B|D)$，

参数 k.

输出：对每个变量 X_i，返回一个 k 候选父集 C_i，其中 $X_i, i=1,2,\cdots,n$.

（Ⅰ）对每个 X_i，计算 $I(X_i,X_j), X_j\neq X_i$ 而且 $X_j\notin Pa(X_i)$；

（Ⅱ）挑选具有最高权值的 $k-l$ 的元素，$l=|Pa(X_i)|$；

候选集合 $C_i=Pa(X_i)Y\{x_1,\cdots,x_{k-l}\}$；

返回 $\{C_i\}$.

3 算法测试及分析

我们对有20000条的病人病况记录分别进行了传统的贝叶斯信念网络学习算法和压缩候选的贝叶斯信念网络学习算法的测试和比较. 在参数 k 取20～25时，压缩候选算法构造网络所需要的时间仅为传统学习算法的1/3至1/5，同时学习得到网络结构的评估函数的值仍然相当高. 这说明了压缩候选的贝叶斯信念网络学习算法所构造的网络仍然与训练样本中包含的默认网络结构有较高的契合度.

4 结束语

本文针对传统算法在对海量数据分类时速度较慢的缺点，提出了压缩候选的贝叶斯信念网络构造算法. 该算法对传统的贝叶斯信念网络构造算法做了许多改进，在不影响原有算法的可靠性的前提下，大大提高了学习速度.

参考文献：

[1] 范 明，孟小峰. 数据挖掘概念和技术[M]. 北京：机械工业出版社，2001.

（责任编辑：黎贞崇）