基于贝叶斯信念网络的数据分类挖掘算法*)

李 芸
(广东纺织技术学院 佛山 528041)

摘 要 贝叶斯方法是概率统计学中一种很重要的方法。分类知识发现是数据挖掘的一项重要内容,研究各种高性能、高速度的分类算法是数据挖掘面临的主要问题之一。本文介绍了贝叶斯信念网络,并针对传统算法在对海量数据进行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。它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,并通过在实际工作的执行情况来证明该算法的有效性。

关键词 贝叶斯网络,分类,数据挖掘

The Algolrithm of Data Classification Unearth Based on Bayesian Belief Network

LI Yun
(Guangdong Textikle Polytechnic, Foshan 528041)

Abstract Bayesian approach is an important method in statistics. Data classification is an important task of data miming. To discover a high-performance, high-speed classification is one of key problems for data mining. In this paper, we introduce the Bayesian belief nerwork. Because these algorithms are very slow, we introduce a method based on compressive candidates, which greatly speed up the study process. At last we prove that this method is reasonable for its application on live data.

Keywords Bayesian belief network, Classification, Data mining

1 前言

生活在18世纪的托马斯-贝叶斯(Thomas Bayes)生前是位受人尊敬的英格兰长老会牧师,为了证明上帝的存在,他发明了概率统计学原理,遗憾的是,他的这一美好愿望至死也未能实现。不过,200多年后的今天,他的这一理论却成了21世纪计算机软件的理论基础,尤其是在数据管理软件领域。随着数据库技术的广泛应用,各行各业都积累了大量有用数据[1]。这些数据所隐含的内在联系可能就是有价值的知识,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。它包括关联分析、分类、预测、聚类分析和孤立点分析等几个方面。分类作为数据挖掘的主要内容之一,主要是通过分析训练数据样本,产生关于类别的精确描述[2]。这种类别通常由分类规则组成,可以用来对未来的数据进行分类预测,有着广泛的应用前景。本文根据样本数据建立的贝叶斯网络的传统算法,提出"压缩侯选的贝叶斯信念网络算法",它对传统的贝叶斯网络算法做了许多改进,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,并在不影响其性能的基础上,极大地提高了运算速度,对在大型数据库上运用贝叶斯信念网络进行数据挖掘有极大的现实意义。

2 贝叶斯网络结构

2.1 贝叶斯网络理论基础

贝叶斯网络推理能够处理不完备数据集。如:规则挖掘,决策树,人工神经网络,密度估计,分类,回归和聚类等方法,传统推理是无法解决的,对于传统的推理必须知道所有可能的数据输入,如果缺少其中的某一输入就会对建立的模型产生偏差。贝叶斯方法则可以解决这个问题,因为贝叶斯网络反映的是整个数据域中数据间的概率关系,即使缺少某一数据变量仍然可以建立精确的模型。贝叶斯网络是根据因果关系进行推理的。在数据分析处理中获得变量域的理解是十分重要的,而且贝叶斯网络可以在缺少插入值的情况下进行决策。它说明了联合条件分布,允许在变量的子集之间定义类条件独立性提供了一种因果关系图形,可以在其上学习并根据学习结果进行分类。贝叶斯网络和贝叶斯统计概率是紧密相关的,这促进了知识和数据域之间的关联关系。它不需要知道处理数据域的先验知识就可以建立正确的预测模型,由于贝叶斯网络具有语义的因果关系因而可以直接地进行因果先验知识的分析,因此在贝叶斯网络中可以获得较全面的先验知识[3]。

2.2 贝叶斯用于分类的方法

为阐述单个变量的分布函数的求法,首先讲一个掷图钉的例子:设掷为头的可能性是 t,那么 t 的可能性概率分布函数 $P(t|\xi)$。那么下一次掷为头的概率是 $P(x=\text{heads}|\xi)=\int p(x=\text{heads}|t,\xi)p(t|\xi)dt=\int t*p(t|\xi)dt=E(t|\xi)$。而且,进一步如果掷为头后的 t 的分布概率就为 $p(t|x=\text{heads},\xi)=c*p(x=\text{heads}|t,\xi)*p(t|\xi)=c*t*p(t|\xi)$。这样的话 $p(t|\text{mheads},\text{ntails},\xi)=c*t(m)*(1-t)(n)*p(t|\xi)$[其中 $t(m)$ 表示 t 的 m 次方],也就求得 m 次掷为"头",n 次掷为尾后的 t 的概率分布情况[4]。上面的是对于两个结果的情况的分析,那么对于离散的多种结果的情况,我们可以用同样的方法进行分析。

*)本课题得到全国教育科学十五规划重点课题(No:AYA010034)基金资助。李 芸 讲师,研究方向:计算机软件工程。

2.3 贝叶斯信念网络

定义 1 给定一个随机变量集 $\chi=\{X_1,X_2,\cdots,X_n\}$，若 χ 上的一条联合条件概率分布。则用贝叶斯信念网络定义如下：

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

其中，X_i 是一个 m 维向量；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 $X_1,X_1,\cdots,X_n$；其弧代表一个函数依赖关系；θ 代表用于量化网络的一组参数。

定义 2 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 Y 是的后继。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，则 G 中 X_i 的所有双亲变量为集合 $Pa(X_i)$。

定义 3 如果对于每一个 X_i，$pa(X_i)$ 的取值 x_i 存在如下一个参数：$\theta_{x_i|pa(x_i)}=P(x_i|pa(X_i))$，且给定 $pa(X_i)$ 发生情况下事件发生的条件概率为 x_i，则贝叶斯信念网络给定的变量集合 χ 上的联合条件概率分布：

$$P_B(X_1,X_2,\cdots,X_n)=\prod_{i=1}^{n}P_B(X_i|Pa(X_i)) \tag{2}$$

4 贝叶斯信念网络结构算法

贝叶斯信念网络算法可以表示如下：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 $S(B|D)$，使用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。常用的评价函数有贝叶斯权矩阵（Bayesian Score Metric）及最小描述长度函数（Minimal Description Length）[5]。由于基于这两种评价函数的学习算法非常复杂，限于篇幅，本文不再详细阐述解释，读者可以查看文[1]和[2]。

在传统的贝叶斯网络算法中，当进行寻找网络结构时，要从 $n-1$ 个侯选节点中逐一搜索 X_i 的父亲变量[6]。这个算法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侯选变量[7]。比如对如下的蕴涵式：$X\rightarrow Y\rightarrow Z$。我们可以发现 X 和 Y；Y 和 Z；X 和 Z 之间存在依赖联系。如果考虑 X 和 Y 都是 Z 的父节点，就可以发现一旦将 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，本文提出了“压缩侯选”的贝叶斯信念网络算法，即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以来关系；$I(X,Y)$ 越小，就说明 X 和 Y 互为父子的可能性越弱。因此通过计算变量之间的依赖关系，可以在选择父节点时选择所有的变量，而且是集中扫描那些最可能是 X_i 的父亲的变量集，$Y_{i1},Y_{i2},\cdots,Y_{ik}$，$k<<n$。贝叶斯信念网络算法描述如下：

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
　从 1,2,…，到 n
(1)压缩：根据 D 和 B_{n-1}，使用侯选压缩，从 $X_1,X_2,\cdots,X_n$ 中，为 X_i 选择一个侯选父集 $C_i^n(|C_i^n|\leqslant k)$，这里定义了一个有向图 $H_n=(\chi,E)$，其中，$E=\{X_j\rightarrow X_i\mid\forall i,j,X_j\in C_i^n\}$。
(2)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，$G_n\subset H_n$，$\forall X_i,Pa^{G_n}(X_i)\subseteq C_i^n$
返回 B_n。

这个算法最关键的一步就是在计算评价函数之间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，我们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，$D_{KL}(P(X)|Q(X)=\sum_X P(X)\log\frac{P(X)}{P(Y)}$

使用上面定义的依赖度量函数的侯选压缩算法如下：

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 侯选父集 C_i
　对于每个 X_i，$i=1,2,\cdots,n$
(1)对每个 X_j，计算 $I(X_i,X_j)$，$X_j\neq X_i$ 而且 $X_j\notin Pa(X_i)$
(2)挑选具有最高权值 $k-1$ 的的元素，$l=|Pa(X_i)|$
　候选集合 $C_i=Pa(X_i)Y\{x_1,\cdots,x_{k-1}\}$
返回 $\{C_i\}$。

实验及分析 在广东省纺织学院学费征收管理系统平台上，分别测试了传统的贝叶斯网络算法和压缩候选的贝叶斯信念网络算法。在参数 k（规费的征收）取 20～25 时，压缩侯选算法构造的网络所需要时间减少为传统算法的 1/3 到 1/5，而同时所获网络结构的评估函数值仍然是一个相当高的值。这说明了该算法所构造的网络仍然和训练样本中包含的默认网络结构有较高的契合度。

参 考 文 献

1 Cooper G F, Herskovtis E. A Bayesian method for the induction of probabilistic network from data. Machine Learning, 1992(10)
2 范明，孟小峰，等. 数据挖掘概念与技术. 机械工业出版社，2001
3 张剑. 自治 Agent 的分布式入侵检测系统研究. 计算机工程与应用，2003，24(8)
4 Curtisdalton Getting Personal Witting Personal with Fire walls [J]. Network Magazine, 2001, 16(1): 102～106
5 http://tech. sina. com. cn 2003/10/20 14:55
6 Anomaly E E. Detection over noisy data using learned probability data using learned probability distributions [A]. In: Proceedings of the International Conference on Machine Learning, 2000
7 Lam W, Bacchus F. Learning Bayesian belief networks: An approach based on the MDL principle. Com. Int., 1994(10)