基于数据挖掘技术的算法研究*

张德丰[1], 马子龙[2], 梁忠宏[2]
(1. 佛山教育学院计算机系, 广东 佛山 528000;
2. 哈尔滨工业大学航空航天学院, 黑龙江 哈尔滨 150001)

摘 要: 分析了贝叶斯信念网络和数理统计方法在数据挖掘中的作用, 提出了一种贝叶斯信念网络和基于数理统计的数据挖掘模型, 并用实例证明该数据挖掘模型有效性。对传统的贝叶斯信念网络构造算法做了许多改进, 在不影响其性能的基础上, 极大地提高了运算速度。

关键词: 数据挖掘; 乔里斯基法; 雅可比法; 贝叶斯信念网络
中图分类号: TN919.2 文献标识码: A 文章编号: 0529-6579 (2003) 04-0036-04

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。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。

1 数理统计算法模型的建立

数理统计学是一门关于数据资料的收集、整理、分析和推理的科学, 在时下的数据挖掘热潮中, 数理统计方法仍是一种不可或缺的方法。模式化是数据挖掘的核心, 用的最广泛又最为经典的模式化方法当数数理统计分析, 一般情况下, 在数据库或数据仓库字段之间存在两种关系: 函数关系(能用函数公式表示的确定关系) 和相关关系 (不能用函数公式表示, 但仍是相关确定关系), 对它们可进行回归分析、相关分析、主成分分析。

为了分析数据仓库中的数据关联性, 需进行多元线性回归分析和利用相关系数表进行特征值与特征向量分析, 以确定主成分。

1.1 利用乔里斯基分解法解出回归系数

从数据仓库中抽取随机变量 y 及 m 个自变量 $x_0, x_1, \cdots, x_{m-1}$, 给定 n 但观测数据 $(x_{0i}, x_{1i}, \cdots, x_{m-1,i}, y_i)$ $(i = 0, 1, \cdots, n-1)$, 用线性表达式:

$$y = a_0 x_0 + a_1 x_1 + \cdots + a_{m-1} x_{m-1} + a_m$$

对观测值进行回归分析, 其中 $a_0, a_1, \cdots, a_{m-1}, a_m$ 为回归系数, 根据最小二乘原理, 为使 $q = \sum_{i=0}^{n-1} [y_i - (a_0 x_0 + a_1 x_1 + \cdots + a_{m-1} x_{m-1} + a_m)]^2$ 达最小, 回归系数 $a_0, a_1, \cdots, a_{m-1}, a_m$ 应满足下列方程组:

$$(\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}})[a_0, a_1, a_2, \cdots, a_{m-1}, a_m]^{\mathrm{T}} = [y_0, y_1, y_2, \cdots, y_{n-1}, y_n]^{\mathrm{T}}$$

其中

$$\boldsymbol{C} = \begin{bmatrix} x_{00} & x_{01} & x_{02} & \cdots & x_{0,n-1} \\ x_{10} & x_{11} & x_{12} & \cdots & x_{1,n-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{m-1,0} & x_{m-1,1} & x_{m-1,2} & \cdots & x_{m-1,n-1} \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}$$

乔里斯基 (cholesky) 分解法是假定 $\boldsymbol{AX} = \boldsymbol{D}$ 中的 $\boldsymbol{A}$ 为对称正定时, 可以惟一的分解为 $\boldsymbol{A} = \boldsymbol{UTU}$, 其中 $\boldsymbol{U}$ 为上三角矩阵, 即:

$$\begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0,n-1} \\ a_{10} & a_{11} & \cdots & a_{1,n-1} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n-1,0} & a_{n-1,1} & \cdots & a_{n-1,n-1} \end{bmatrix} = \begin{bmatrix} u_{00} & 0 & \cdots & 0 \\ u_{10} & u_{11} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ u_{n-1,0} & u_{n-1,1} & \cdots & u_{n-1,n-1} \end{bmatrix}。$$

* 收稿日期: 2003-12-28
基金项目: 国家自然科学基金资助项目 (60133010, 60073053); 国家“八六三”高技术研究发展计划资助项目 (863-317-03-05-98)
作者简介: 张德丰 (1963 年生), 男, 硕士; E-mail: zhangdf@foshan.net

$$\begin{bmatrix} u_{00} & u_{01}0 & \cdots & u_{0,n-1} \\ 0 & u_{11} & \cdots & u_{1,n-1} \\ 0 & 0 & \cdots & \cdots \\ 0 & 0 & \cdots & u_{n-1,n-1} \end{bmatrix}$$

$\boldsymbol{U}$ 矩阵中的各元素由以下计算公式确定：

$$u_{00} = \sqrt{a_{00}}$$

$$u_{ii} = \left(a_{ii} - \sum_{k=0}^{i=1} u_{ki}^2 \right)^{1/2}, \quad i = 1, 2, \cdots n-1$$

$$u_{ij} = \left(a_{ij} - \sum_{k=0}^{i=1} u_{ki} u_{kj} \right) / u_{ii}, \quad j > i$$

于是，方程组 $\boldsymbol{AX}=\boldsymbol{B}$ 的解可由下述公式计算：

$$y_i = \left(b_i - \sum u_{ki} y_k \right) \backslash u_{ii}$$

$$x_i = \left(y_i - \sum_{k=i+1}^{n-1} u_{ik} x_k \right) \backslash u_{ii}$$

1.2　求 x_i，x_j 的相关系数 ρ

公式如下：

$$\rho_{x_i x_j} = \frac{\sum_{k=1}^{n} x_{ik} x_{jk} - n\overline{xy}}{\sqrt{\sum_{k=1}^{n} x_{ik}^2 - n\overline{x^2}} \sqrt{\sum_{k=1}^{n} x_{jk}^2 - n\overline{y^2}}}$$

构造相关系数矩阵，它有如下特征：

由于 $\rho_{x_i x_j} = \rho_{x_j x_i}$，故该矩阵为对称矩阵对角线系数为 1，因为自相关系数为 1。

1.3　用雅可比方法求特征值和特征向量

由于上述相关系数矩阵为实对称矩阵，故可用 Jacobi 法求该矩阵的全部特征值及相应的特征向量，雅可比法的关键思想如下。

设 n 阶矩阵 $\boldsymbol{A}$ 为对称矩阵，在 N 阶对称矩阵 $\boldsymbol{A}$ 的非对称矩阵的非对角线元素中选取一个绝对值最大的元素，设为 a_{pq}，利用平面旋转变换矩阵 $\boldsymbol{R}_0(p, q, \theta)$ 对 $\boldsymbol{A}$ 进行正交相似变换：

$$\boldsymbol{A}_1 = \boldsymbol{R}_0(p, q, \theta)^{\mathrm{T}} \boldsymbol{A} \boldsymbol{R}_0 \boldsymbol{p}(p, q, \theta)$$

其中，$\boldsymbol{R}_0(p, q, \theta)$ 的元素为 $r_{pp} = \cos\theta$，$r_{qq} = \cos\theta$，$r_{pq} = -\sin\theta$，$r_{qp} = \sin\theta$，$r_{ij} = 0$，$i, j \neq p, q$

如果按下式确定角度 θ

$$\tan 2\theta = \frac{2a_{pq}}{a_{pp} - a_{qq}}$$

则对称矩阵 $\boldsymbol{A}$ 经上述变换后，其非对角线元素的平方和将减少 $2a_{pq}^2$，对角线元素的平方和增加 $2a_{pq}^2$，而矩阵中所有元素的平方和保持不变。由此可知，对称矩阵 $\boldsymbol{A}$ 每经过一次变换，其非对角线元素的平方和向零接近一步，因此只要反复进行上叙变换，就可逐步将矩阵 $\boldsymbol{A}$ 变为对角矩阵，对角矩阵中对角线上的元素 λ_0，λ_1，$\cdots\lambda_{n-1}$ 即为特征值，而每一步的平面旋转矩阵的乘积的第 i 列（$i=0, 1, 2, \cdots, n-1$）即为与 λ_i 对应的特征向量。

雅可比方法可分为下面 3 种：

(1) 经典的雅可比法：就是在每一步选绝对值为最大的非对角线元素作为化零（或说扫除）的对象，不难证明，这样的非对角线全体元素平方和 τ_k 确实趋于零，但每一步都要就所有非对角元进行比较，才能挑出绝对值为最大的元素。这很费时间。

(2) 循环的雅可比法：对非对角元位置规定一个顺序，然后按照这个顺序进行扫除，碰到非零元就用上述办法把它化为零，否则就跳过去，反复按规定顺序进行扫除，这种办法也是收敛的。

(3) 限值的循环法：除了和循环法一样要对非对角元规定一个顺序外，再取下降于零的数列 a_1，a_2，…作为限值，首先就 a_1 来进行扫掠，即按规定顺序扫掠，碰到绝对值小于 a_1 的非对角元就跳过去；否则才做化零工作，反复循序进行，直到所有的非对角元的绝对值都小于 a_1 为止，然后取 a_2 作为新的限量防止进行，这样继续就 a_3，a_4，…扫掠下去，只至达到我们的要求为止。一般常用如下方法取限制：对于原给的 $\boldsymbol{A}$，记 $r = \sqrt{\tau_k}$，任取一个 $m \geqslant n$，并用 $a_k = r / m^k$ 作为限值。这种方法也是收敛的。

雅可比方法步骤如下：令 $\boldsymbol{S} = I_n$（I_n 为单位矩阵），在 $\boldsymbol{A}$ 中选取非对角线元素中绝对值最大者，设为 a_{pq}，若 $|a_{pq}| < \varepsilon$，则跌代过程结束。此时对角线元素 a_{ii}（$i=0, 1, \cdots, n-1$）即为特征值 λ，矩阵 $\boldsymbol{S}$ 的第 i 列为与 λ_i 对应的特征向量。否则，继续下一步。

计算平面旋转矩阵的元素及其变换后的矩阵 $\boldsymbol{A}_1$ 的元素。其计算公式如下：

$$y = \frac{a_{qq} - a_{pp}}{2}$$

$$\tilde{\omega} = \mathrm{sign}(y) \frac{x}{\sqrt{x^2 + y^2}}$$

$$\sin 2\theta = \tilde{\omega}$$

$$\sin\theta = \frac{\tilde{\omega}}{\sqrt{2\left(1 + \sqrt{1 - \tilde{\omega}^2}\right)}}$$

$$\cos\theta = \sqrt{1 - \sin^2\theta}$$

$$a_{pp}^{(1)} = a_{pp}\cos^2\theta + a_{qq}\sin^2\theta + a_{pq}\sin 2\theta$$

$$a_{qq}^{(1)} = a_{pp}\cos^2\theta + a_{qq}\sin^2\theta - a_{pq}\sin 2\theta$$

$$a_{pq}^{(1)} = a_{qp}^{(1)} = 0$$

$$a_{pj}^{(1)} = a_{pj}\cos\theta + a_{qj}\sin\theta$$

$$a_{qj}^{(1)} = -a_{pj}\sin\theta + a_{qj}\cos\theta \qquad j \neq p, q$$

$$a_{ip}^{(1)} = a_{ip}\cos\theta + a_{iq}\sin\theta$$
$$a_{iq}^{(1)} = -a_{ip}\sin\theta + a_{iq}\cos\theta \quad i \neq p, q$$
$$a_{ij}^{(1)} = a_{ij} \quad i, j \neq p, q$$

$\boldsymbol{S} = \boldsymbol{S}^{*}\boldsymbol{R}(p, q, \theta)$，转第二步。

1.4 求出主成分

把特征值从大到小排列，求出累计贡献在70%左右的前几个特征值，这几个特征值就是主成分。

2 贝叶斯信念网络算法模型

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。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。

首先介绍贝叶斯信念网络，然后介绍了根据样本数据建立贝叶斯信念网络的传统构造算法并提出了“压缩候选的贝叶斯信念网络构造算法”，它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。

2.1 贝叶斯信念网络

定义 1 给定一个随机变量集 $\mathrm{X} = \{X_1, X_2, \cdots, X_n\}$，其中 X_i 是一个 m 维向量。贝叶斯信念网络了说明 X 上的一条联合条件概率分布。贝叶斯信念网络定义如下：

$$B = \langle G, \theta \rangle$$

第一部分 G 是一个有向无环图，其顶点对应于有限集 $\boldsymbol{X}$ 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，图中的每个变量独立于图中该节点的非后继。在图 G 中的 X_i 所有双亲变量用集合 $Pa(X_i)$。

第二部分 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，存在如下一个参数：$\theta_{x_i|pa(x_i)} = P(x_i \mid pa(X_i))$，它指明了在给定在 $pa(X_i)$ 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 X 上的联合条件概率分布：

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

贝页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 $S(B \mid D)$，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）。

2.2 压缩候选的贝叶斯网络结构算法

在传统的贝叶斯网络构造算法中，当进行寻找网络结构时，要从 $n-1$ 个候选节点中逐一搜索 X_i 的父亲变量。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候选变量。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。但是一旦我们考虑 X 和 Y 都是 Z 的父节点，我们就可以发现一旦将 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，我们提出了“压缩候选”的贝叶斯信念网络构造算法。我们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；$I(X, Y)$ 很小，就说明 X 和 Y 互为父子的可能性越小。因此通过计算变量之间的依赖关系，我们可以在选择父节点时是选择所有的变量，而是集中扫描那些最可能是 X_i 的父亲的变量集 $Y_{i1}, Y_{i1}, \cdots, Y_{ik}$，$k \ll n$。

根据上述思想，我们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 $S(B \mid D) = \sum_i S(X_i \mid Pa(X_i), D)$。参数 k，输出：最优网络，从 1, 2, 到 n。

(1) 压缩。

根据 D 和 B_{n-1}，使用候选压缩，从 $X_1, X_2, \cdots, X_n$ 中为 X_i 选择一个侯选父集 $C_i^n(|C_i^n| \leqslant k)$ 在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 $S(B_n \mid D)$ 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 $I(X, Y)$：

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

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

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 $S(B \mid D)$，参数 k，输出：对每个变量 X_i，返回一个 k 候选父集 C_j。对于每个 X_i，$i = 1, 2, \cdots, n$。

(1) 对每个 X_j，计算 $I(X_i, Y_j)$，$X_j \neq X_i$，而且 $X_j \notin Pa(X_i)$。

(2) 挑选具有最高权值的 $k - l$ 的元素，$l = \mid Pa(X_i) \mid$。

候选集合 $C_i = Pa(X_i) \cup \{x_1, \cdots, x_{k-l}\}$，返回 $\{C_j\}$。

3　结　论

某企业为了挖掘出哪几种因素影响产品的质量，从产品数据集市中抽取 19 种主要影响产品质量的因素，每个因素取 36 个测试数据（要有代表性）。利用乔里斯基（cholesky）分解法解出回归系数，确定函数关系。利用相关系数矩阵和雅可比方法求出主成分，确定影响产品质量的主要因素。然后进行结论评价，如果结论不合理，再进行数据抽取，利用以前的方法进行再分析，直至结论合理有效。

在一个有 20 000 条记录上分别测试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。在参数 k 取 20～25 时，压缩候选算法构造网络所需要的时间减少为传统学习算法的 1/3～1/5 倍，而同时学习得到网络结构的评估函数的值仍然是一个相当高的值。这说明了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。

数据挖掘涉及的学科领域和方法很多，有多种分类法。统计方法中，可细分为：回归分析（多元回归、自回归等）、判别分析（贝叶斯判别、费歇尔判别、非参数判别等）、聚类分析（系统聚类、动态聚类等）、探索性分析（主元分析法、相关分析法等）等。数据挖掘是一个崭新的计算机应用领域，它将极大地促进信息对于人类社会进步所起的作用。

参考文献：

[1] 范明，孟小峰，赵军. 数据挖掘概念与技术[M]. 北京：机械工业出版社，2001.

[2] COOPER G F, HERSKOVTIS E. A Bayesian method for the induction of probabilistic network from data[M]. Machine Learning. 1992.

[3] 高洪深. 决策支持系统(DSS)理论·方法·案例[M]. 第 2 版. 北京：清华大学出版社，2000.

[4] 王林书，吴渝. 概率论与数理统计[M]. 北京：科学出版社，1999.

[5] 陈代春. 数据仓库技术及其应用研究[D]. 中南大学计算机系，2001.

[6] LAM W, BACCHUS F. Learning Bayesian belief networks: An approach based on the MDL principle[J]. IEEE Trans on Image Processing, 1996, 5(1): 4—15.

[7] 邢乃宁，孙自辉. 基于增量式遗传算法的分类规则挖掘算法[J]. 计算机应用，2001，21(11)：22—24.

The Algorithm of Data Mining Technique

ZHANG De-feng[1], *MA Zi-long*[2], *LIANG Zhong-hong*[2]

(1. Computer Department, Foshan College of Education, Guangdong Foshan, 528000, China;
2. Harbin Institute of Technology, Harbin 150001, China)

Abstract: The function of Bayesian belief network and Mathematical Statistics method is analysed in the field of data mining, then a kind of Bayesian belief network and model of data mine is provided for Mathematical Statistics. The effectiveness of the data mine model is proved with an example. The traditional Bayesian belief network-learning algorithm, is improved to speed up the operation process.

Key words: data mining; Cholesky method; Jacobi method; Bayesian belief network