文章编号：1671－4598(2003)12－0914－04 中图分类号：TP912.31 文献标识码：A

数据挖掘中的统计方法概述

赵广社，张希仁
(西安交通大学，陕西 西安 710049)

摘要：统计方法有成熟的数学基础，可以很好的对数据进行解释，在数据挖掘中有着大量的运用。文章回顾了数据挖掘中常用的统计方法，包括传统的统计方法（回归分析、主成分分析、判别分析和聚类分析）和其他一些非机器学习的方法（模糊集、粗糙集和统计学习理论），分析了各种统计方法的优缺点。

关键词：数据挖掘；回归分析；主成分分析；判别分析；聚类分析；模糊集；粗糙集；支持向量机

Introduction Statistical Techniques in Data Mining

ZHAO Guang-she, ZHANG Xi-ren
(Xi'an Jiaotong University, Xi'an 710049, China)

Abstract: Statistical techniques have mature bases of maths, enable to interpret characteristics of data and have much application in data mining. In this paper, we review statistical techniques used usually, including traditional statistical techniques and techniques which are not machine learning technique. The advantages and disadvantages of every statistical techniques are analyzed.

Key words: data mining; regression analysis; principal component analysis; discriminant analysis; clustering analysis; fuzzy set; rough set; support vector machine

1 引言

近年来，数据挖掘引起了信息产业界的极大关注，其主要原因是存在大量数据，可以广泛使用，并且迫切需要将这些数据转换成有用的信息和知识。获取的信息和知识可以广泛用于各种应用，包括商务管理、生产控制、市场分析和科学探索等。

知识挖掘最新的描述性定义是由 Usama M. Fayyyad 等给出的[1]：数据挖掘是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程。数据挖掘的基本过程和主要步骤如图 1 所示[10]。

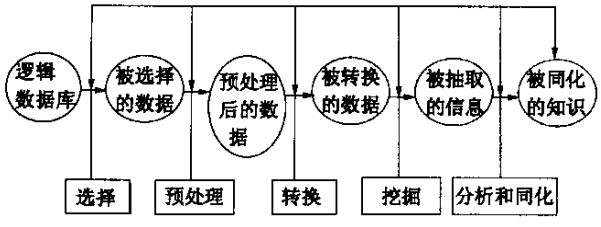

图 1 数据挖掘过程的步骤

数据挖掘是一个多学科领域[2]，包括数据库技术、人工智能、机器学习、神经网络、统计学、模式识别、知识库系统、知识获取、信息检索、高性能计算和可视化。数据挖掘主要有统计方法（因为基本的数据分析来自于这个领域，许多数据分析问题存在统计解决方法）和机器学习（它从另外的方式去处理数据，可以自动地产生和证实假设，分别描述假设）两种方法。此文中主要是回顾数据挖掘中的统计方法。

收稿日期：2003－07－21。
基金项目：国家重点基础研究发展规划项目(2001CB309405)
作者简介：赵广社(1966－)，男，陕西省乾县人，副教授，主要从事图像测量与信息融合理论、数据挖掘与知识发现以及嵌入式系统的研究。

2 数据挖掘的统计方法

2.1 回归分析[4,5]

设有 k 个自变量（预报因子）x_1，…，x_k，p 个因变量 y_1，…，y_p，相应的 n 组观测资料是

$$x_{i1},\cdots,x_{ip},y_{i1},\cdots,y_{ip},i=1,2,\cdots,n.$$

用矩阵来表示，可得资料矩阵

$$\underset{n\times p}{Y}=(y_{i\alpha}),\underset{n\times k}{X}=(x_{i\alpha}),$$

如果因变量 y_α 与自变量 x_1，…，x_k 之间有线性关系式，且 y_α 的值又带有误差，于是有

$$y_\alpha=\beta_{0\alpha}+\beta_{1\alpha}x_1+\beta_{2\alpha}x_2+\cdots+\beta_{k\alpha}x_k+\varepsilon_\alpha,\alpha=1,\cdots.p.$$

写成矩阵形式，就是

$$\begin{bmatrix}y_1\\ \vdots\\ y_p\end{bmatrix}=\begin{bmatrix}\beta_{01}&\beta_{11}&\cdots&\beta_{k1}\\ \beta_{02}&\beta_{12}&\cdots&\beta_{k2}\\ \vdots&\vdots&&\vdots\\ \beta_{0p}&\beta_{1p}&\cdots&\beta_{kp}\end{bmatrix}\begin{bmatrix}1\\ x_1\\ \vdots\\ x_k\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\ \vdots\\ \varepsilon_p\end{bmatrix}$$

统计问题就是从已知的资料矩阵 Y 和 X 出发，如何求得上式中的这些未知常数 β_{ij}，并且对误差 ε_a 作出估计和推断。β_{ij} 称为回归系数，而略去误差后的关系式

$$y_\alpha=\beta_{0\alpha}+\beta_{1\alpha}x_1+\beta_{2\alpha}x_2+\cdots+\beta_{k\alpha}x_k,\alpha=1,\cdots.P.$$

称为回归方程。β_{ij} 用最小二乘法求解。

回归是学习一个函数，这个函数是将数据项映射为

实值的预测变量。在数据挖掘中回归分析应用的例子很多，例如，回归分析可用来预测由微波遥感探测的森林里的生物的数量；可以估计在给出诊断结果的情况下病人的能存活的概率；可以作为一个广告花费函数来预测消费者对新产品的需求；可以预测输入变量可能是预测变量的时间滞后形式的时间序列。

2.2 主成分分析（PCA）

主成分分析是设法将原来众多具有一定相关性的指标（比如 p 个指标），重新组合成一组新的相互无关的综合指标来代替原来指标。通常数学上的处理就是将原来 p 个指标作线性组合，作为新的综合指标。如果将选取的第一个线性组合即第一个综合指标记为 F_1，我们希望 F_1 尽可能多的反映原来指标的信息。最经典的方法就是用 F_1 的方差来表达，即 Var（F_1）越大，表示 F_1 包含的信息越多。因此在所有的线性组合中所选取的 F_1 应该是方差最大的，故称 F_1 为第一主成分。如果第一主成分不足以代表原来 p 个指标的信息，再考虑选取 F_1 即选取第二个线性组合，为了有效地反映原来信息，F_1 已有的信息就不需要再出现在 F_2 中，用数学语言表达就是要求 Cov（F_1，F_2）$=0$，称 F_2 为第二主成分，依次类推可以造出第三，四，……，第 p 个主成分。在实际工作中，就挑选前几个最大主成分。将主成分分析推广，就是因子分析[4]。

在数据挖掘中，PCA 常常用于数据的预处理，消除噪声，消除数据属性间的相关性，降低数据的维数，以减少计算量，提高运算速度。但通过 PCA 我们也可以呈现数据变量的响应趋势，在文献［3］就给出了这样的例子：利用 PCA 对数据进行预处理，就可以抓住过程的特征，将特征空间划分成许多小区域，在每个区域里，数据点都有相似的响应趋势；而不同区域的数据点有不同的响应趋势。这样对新的数据我们只要进行相应的变换，将其映射到特征空间里相应的区域，我们就可以预测它的响应趋势。在这里作者还将变量的动态响应描述为概念，这样可以进行逻辑推理，判断设备在哪个区域运行。

将 PCA 用于状态识别、监视和诊断[3]：操作员对工厂安全和运作的监视；工程师对工厂长期性能分析的监视。结合 PLS（部分最小二乘法）和 SPE（平方预测误差），可以推断实例偏离正常状态的原因：分析哪些因素是使实例偏离的主要原因；给操作员提出该如何操作的建议。

2.3 聚类分析

为了将样品进行分类，就需要研究样品之间的关系，一种方法是用相似系数，性质越接近的样品，他们的相似系数越接近于 1（或 -1），而彼此无关的样品他们的相似系数则越接近于 0，比较相似的样品归为一类，不怎么相似的样品属于不同的类。另一种方法是将每一个样品看作 m 维空间的一点，并在空间定义距离，距离较近的点归为一类，距离较远的点应属于不同的类，样品之间的相似系数和距离有各种各样定义，而这些定义与指标（变量）的类型关系极大，通常指标按照测量它们的尺度来进行分类[5]。

聚类方法很多，但其核心只有两个，一个是样本的相似度量问题，另一个是聚类的准则问题。所谓样本相似性度量有距离（距离度量还有许多种）、相关系数、夹角方向余弦 3 种。聚类的准则可以分为两类，一类是启发式方法，根据经验和直观确定一些准则，比如 A 和 B 一类，B 和 C 一类，那么 A 和 C 也是一类。另一类是最优化的技术，根据聚类问题的实际背景确定一个目标函数，比如类内样本间差别最小，而类间样本的差别最大，这样，聚类问题就转化成一个最优化问题。

大体上，主要的聚类算法可以划分为如下几类[5]：划分方法、层次的方法、基于密度的方法、基于网格的方法和基于模型的方法。

聚类是一种常用的描述方法，寻求可以识别有限的类类别或聚类集合来描述数据。聚类在数据挖掘中应用的例子很多[1,8,9]，如发现超市数据库中有类似购买行为的消费者群体，识别从天空测距红外线光谱的子类。

2.4 判别分析[3]

判别分析：有 k 个总体 G_1，…，G_k，它们的分布函数分别是 F_1（y），…，F_k（y），每个 F_i（y）均是 m 维分布函数。对给定的一个样品 y，我们要判断它来自哪个母体。

判别分析是根据样品的类别属性计算 cdfs 和找到能最好的划分这些类别的因子。判别分析给出了判别函数集，每个判别函数如下：

$$cdf_{ci} = \omega_0 + \omega_1 X_{1ci} + \omega_2 X_{2ci} + \cdots + \omega_a X_{aci}$$

其中 cdf_{ci} 是类别 c 中样品 i 的规范判别函数，ω_a 是每个变量 X_{aci} 的系数，X_{aci} 是 c 类别中样品 i 变量 X_a 的值。

为了估计 cdf 的数量得到 cdf 的系数，要计算整个空间数据点的数量和在此空间中每个类别的好坏。我们通过离差矩阵 DISP 来实现这点。其定义如下：

$$disp_{ab} = \sum_{l=1}^{c} \sum_{m=1}^{n_c} (X_{alm} - X_{a..})(X_{blm} - X_{b..})$$

这里，$disp_{ab}$ 是规范矩阵 DISP 中的元素（a 和 b 是独立的属性），c 是类别数，n_c 是 c 类别中对样品变量 X_a 的值，$X_{a..}$ 是对所有样品（不依赖于类）变量 X_a 的平均值。

同理，用类平均 X_{la} 代替总体平均 $X_{a..}$ 就可以计算群内离差矩阵 $DISP_{\text{within-groups}}$。一个类里的离差就是将数据点归纳为同一类的测量方法。如果类不能代表群（即类在整个数据空间是随机分布的），则 $DISP_{\text{total}}$ 等于 $DISP_{\text{within-groups}}$。$DISP_{\text{total}}$ 是数据的平方叉积的总和。在这样的例子中好的可能的分类（Wilks Λ（值的范围是［0－1］）将临近单元的值描述为群的重心，这种描述显示了大的相似性）。

通过 $DISP_{\text{between-groups}} = DISP_{\text{total}} - DISP_{\text{within-groups}}$ 来计算群间和的平方叉积矩阵。只要在条件 $DISP_{\text{within-groups}} < DISP_{\text{total}}$ 下类的重心没有重合，则矩阵 $DISP_{\text{total}}$ 就是变量总的偏离的测量。残差偏离被解释为有多少个类是关联的（群间）。因此我们用如下方法可以计算特征值（λ）和它们的属性 v ' s：

$$\lambda = \frac{\sum b_{1a}v_a}{\sum \omega_{1a}v_a} = \frac{\sum b_{2a}v_a}{\sum \omega_{2a}v_a} = \cdots = \frac{\sum b_{pa}v_a}{\sum \omega_{pa}v_a}$$

其中 v ' s 是 p 个系数的集合，它用于产生规范判别函数 cdf_r（指的是 $cdfs$ 中第 r 个 cdf）中的 w。当 v 的平方和设定为 1.0 时，对于这个等式就可以找到有限个唯一的解答。对应于 λ 和 v ' s 的唯一的解答指定一个 cdf。对于每一个初始化的 cdf，v ' s 都要被变换是为了得到标准的系数方法。这样，所有实例的判别式的和为 0 而标准偏差为 1.0。变换 v ' s 是为了得到 cdf_r 的 w ' s：

$$\omega_p = v_a\sqrt{n-c} \qquad \omega_0 = -\sum_{t=1}^{p}\omega_t X_t$$

对于 v ' s 和对应的 λ 的每个解答都指定了一个判别函数。现在通过特征值的诠释就有可能决定选取多少个判别函数。这些特征值通常是有序的，特征值越接近 0，则对应的判别函数就越不重要。因此可以仅仅选取那些对应的特征值远大于 0 的判别函数。

通过计算相关系数（CCC）来判别判别函数和类间的关联，相关系数定义如下：

$$CCC_r = \sqrt{\frac{\lambda_r}{1+\lambda_r}}$$

当 CCC_r 接近于 0 时，则判别函数 r 和类之间无关联；而当 CCC_r 接近于 1.0（它的最大值）时，则判别函数 r 和类之间有大的关联。常用的判别分析有 Bayes 判别、Fisher 判别和非参数判别等。

2.5 模糊集[2]

模糊集是表示和处理不确定性数据的重要方法。模糊集不仅可以处理不完全数据、噪音或不精确数据，而且在开发数据的不不确定性模型方面是有用的，与传统方法相比可提供更灵巧、更平滑的性能。

模糊逻辑将传统的二值逻辑推广到无穷多值逻辑，使得命题的真值可取［0，1］间的任何实数，它用数值的大小来表示命题的真度，使得在描述前提或结论成立的程度时，不至于只要真假两种状态，从而使得对逻辑规律的描述更符合现实。此外，人们经常运用的逻辑中前提与结论间的关系本身往往不总是一清二楚的，而包含着各种模糊性，这也是在模糊逻辑中要考虑的问题。模糊逻辑推理公式一般由前提部分和结论部分组成，而每一部分又可由若干简单命题组成。对简单命题作进一步分析，人们又可以分解出其中的个体词、谓词和量词等成分，并通过研究它们之间的形式结构和逻辑关系，总结其正确的推理形式和规则。因此，按照逻辑的组成层次，模糊逻辑分为模糊命题逻辑和模糊谓词逻辑。

2.6 粗糙集[2]

粗糙集（rough set）理论由 Zdziskew Pawlak 在 1982 年提出。他是一种新的实现工具，用于处理含糊性和不确定性，在数据挖掘中发挥了重要作用。粗糙集是由集合的下近似、上近似来定义的。下近似中的每个成员都是集合的确定成员，而不是上近似中的成员肯定不是该集合的成员。粗糙集的上近似是下近似和边界区的合并。边界区的成员可能是该集合的成员，但不是确定的成员。可以认为粗糙集是具有三值隶属函数的模糊集，即是、不是、也许。与模糊集一样，它是一种处理数据不确定性的数学工具，常与规则归纳、分类和聚类方法结合起来使用，很少单独使用。

2.7 支持向量机[7]

1992 年，Vapnik 提出支持向量机（support vector machine，SVM）的概念，解决了对非线性函数来求解超平面的问题。其基本考虑可以分为 3 个部分：

（1）通过内积函数定义的非线性映射将样本空间映射到一个高维线性空间，称为特征空间，然后，在这个线性空间中求出分类超平面。即，在特征空间中，求出使方程 $\omega x - \theta = 0$ 成立的超平面，它可以将给定样本中的样本正确划分。

（2）强调分类超平面是最优的，并由此给出支持向量的概念。最优分类超平面的含义是：存在一个超平面满足对样本正确划分的条件，而且，在这个超平面两侧的与这个超平面距离最近的两个点之间的距离最大。这样，通过超平面两侧的这两个点作与超平面平行的向量，所以这样的向量称为支持向量。在 Vapnik 的统计机器学习理论中，泛化是与学习同时考虑的，这就是在定义最优超平面时，需要考虑被划分为两个区域的点之间需要满足距离最大的原因。

（3）由于问题的目标是最优的分类超平面，它可以被变换为一个不等式约束下的二次优化问题。这个优化问题有一个重要的特点，即，其优化方程涉及样本之间的内积，根据 Mercer 定理可知，这就意味着，对于最优分类超平面问题，只要采用适当的内积函数，就可以实现非线性变换的线性分类。

Vapnik 的统计机器学习理论最重要的思想是将非线性函数映射到一个高维线性空间。一般地，对于非线性问题，通过非线性映射变换为一个高维空间线性问题的困难在于这个变换可能非常复杂，Vapnik 根据 Mercer 定理证明了这种变换所需要满足的条件与可以采用的方法（内积），并使用这种方法描述了 BP 算法、径向基算法等。

而强调最优分类超平面有两个原因：其一，Vapnik 试图将统计机器学习理论与泛化问题联系在一起，其二，求解最优分类超平面问题可以变换为在一定约束条件下，求下述函数的二次规划问题：

$$Q(a_i)\sum a_i - \sum a_i a_j y_i y_j K(x_i \cdot x_j)$$

式中 K 是一个满足 Mercer 条件的核函数。对上述

函数的二次规划问题，可以正，对 a_i 存在唯一解，其中 $a_i \neq 0$，所对应的样本就是支持向量。

3　结论

从统计学发展而来的方法与从机器学习中发展而来的方法同样重要。可以把统计方法看作是数据分析的参数方法，这也就暗含了可以获得或收集到正确的函数类型，参数数目以及参数可能的值。在统计学里，需要对数据彻底地了解来获得正确（参数化）的模型。而机器学习能自动的产生和证实拟合数据的模型，不用预先定义模型。但是在实际运用中，统计方法和机器学习方法可以互相补充，也可以融合在一起使用。统计学为数据分析提供了大量的技术方法和结论，它有许多优点：

（1）实际应用和合理的条件下可以证明数据挖掘中所使用的估计和搜索过程是一致性的；

（2）用和挖掘不确定性而不是隐藏它；

（3）证实搜索误差即诚信度和模型均值的好处；

（4）不会混淆带干涉的条件，即不要将假设检验的误差概率误认为搜索过程的误差概率。在理解搜索结构中很少用到统计，但在搜索过程中的评估，在搜索结论的评估，在结论的合理运用中用到大量的统计。

参考文献：

[1] HAN J W. KAMBER M. 数据挖掘：概念与技术 [M]. 范明，孟小峰，译. 北京：机械工业出版社，2001.

[2] 陆汝钤. 世纪之交的知识工程与知识科学 [M]. 北京：清华大学出版社，2001.

[3] WANG X Z. Data mining and knowledge discovery for process monitoring and control [M]. Springer－Verlag London，1999.

[4] 于秀林，任雪松. 多元统计分析 [M]. 北京：中国统计出版社，1999.

[5] 张尧庭，方开泰. 多元统计分析引论 [M]. 北京：科学出版社，1982.

[6] ENGELS R. Component－based user guidance in knowledge discovery and data mining [M]. Sankt Augustin：infix，1999.

[7] 边肇祺，张学工. 模式识别 [M]. 北京：清华大学出版社，1999.

[8] GLYMOUR C，MADINGAN D，PREGIBON D，et al. Statistical inference and data mining [J]. Communications of the ACM，1996，39：35－41.

[9] FAYYAD U，SHAPIRO G P，SMYTH P. From data mining to knowledge discovery in databases [M]. AAAI，1997.

[10] 罗晓沛. 数据挖掘在科学数据库中的应用探索 [C]. 中国科技大学，2002.

（上接第 913 页）

[8] COLORNI A，DORIGO M，et al. Ant system for job－shop scheduling [J]. Belgian Journal of Operations Research，Statistics and Computer Science (JORBEL)，1994，34：39－53.

[9] 孙新宇，万筱宁，孙林岩. 蚁群算法在混流装配线调度问题中的应用 [J]. 信息与控制，2002，31(6)：486－490.

[10] 侯立文，蒋馥. 一种基于蚂蚁算法的交通分配方法及其应用 [J]. 上海交通大学学报，2001，35(6)：930－933.

[11] BULLNHEIMER B，et al. An improved ant system algorithm for the vehicle routing problem [R]. Technical Report POM－10/97. Institute of Management Science，University of Vienna，1997.

[12] BULLNHEIMER B，et al. Applying the ant system to the vehicle routing problem [A]. In OSMAN I H，et al. Meta－Heuristics：Advances and Trends in Local Search Paradigms for Optimization [C]. Kluwer Academics，1998：109－120.

[13] COSTA D，HERTZ A. Ants can colour graphs [J]. Journal of the operational Research Society，1997，48：295－305.

[14] GAMBARDELLA L M，DORIGO M. An hybrid ant system for the sequential ordering problem [R]. Technical Report，IDSIA，Lugano，CH，1997：11－97.

[15] SCHOONDERWOERD R，et al. Ant－based load balancing in telecommunications networks [J]. Adaptive Behavior，1996，5(2)：169－207.

[16] WHITE T，et al. Connection management using adaptive mobile agents [A]. In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications [C]. CSREA Press，1998：802－809.

[17] BONABEAU E，et al. Routing in telecommunication networks with "Smart" ant－like agents telecommunication applications [A]. In Proceedings. of IATA'98 Second Int. Work Shop on Intelligent Agents for Telecommunication Applications [C]. Lectures Notes in AI vol. Springer Verlag，1998：1437.

[18] DICARO G，DORIGO M. Extending AntNet for best－effort Quality－of－Service routing [A]. Proc. ANTS'98－First International Workshop on Ant Colony Optimization [C]. Brussels，Belgium，1998：15－16.

[19] 张素兵，吕国英，刘泽民，周正. 基于蚂蚁算法的 QoS 路由调度方法 [J]. 电路与系统学报，2000，5(1)：1－5.

[20] 张素兵，刘泽民. 基于蚂蚁算法的分级 QoS 路由调度方法 [J]. 北京邮电大学学报，2000，23(4)：11－15.

[21] 顾军华. 基于蚂蚁算法的 QoS 组播路由问题求解 [J]. 河北工业大学学报，2002，31(4)：19－24.

[22] 张徐亮，张晋斌. 基于协同学习的蚁群电缆敷设系统 [J]. 计算机工程与应用，2000，5：181－182.

[23] 何清华，肖人彬，师汉民. 蚂蚁算法在机构同构判定中的实现 [J]. 模式识别与人工智能，2001，14(4)：406－412.

[24] 庄昌文. 基于协同工作方式的一种蚁群布线系统 [J]. 半导体学报，1999，20(5)：400－406.

[25] CASILLAS J，et al. Learning Cooperative fuzzy rules using ant colony optimization algorithms [R]. Technical Report－00119，Spain：Department of Computer Science and Artificial Intelligence，University of Granada，2000.