计算机科学2003Vol. 30№. 12

数据挖掘中的统计方法

魏 玲[1,2] 祁建军[3] 张文修[2]

(西北大学数学系 西安710069)[1] (西安交通大学理学院 西安710049)[2]

(西安交通大学计算机系统结构与网络研究所 西安710049)[3]

Statistical Methods in the Data Mining

WEI Ling[1,2] QI Jian-Jun[3] ZHANG Wen-Xiu[2]

(Dept. of Math., Northwest University, Xi'an 710069)[1] (Faculty of Science, Xi'an Jiaotong University, Xi'an 710049)[2]

(Institute of Computer Architecture & Network, Xi'an Jiaotong University, Xi'an 710049)[3]

Abstract Data mining is the process of secondary analysis of large databases aimed at finding unsuspected relationships which are of interest or value to the database owners. We analyze the statistical methods in the classification in data mining, include: preprocessing techniques, classification algorithms, and post-classification analysis. Also, we introduce the Bayesian networks for data mining.

Keywords Data mining, Preprocessing, Test, Statistical methods, Bayesian networks

1. 引言

随着计算机技术的日趋成熟与广泛应用，人们发现针对大型数据库如商业数据、银行数据和医疗数据等，如果采取适当的方法和技术，可以从中发掘出更多更有用的信息、知识，这些知识可以形成预报和分类模型，最终提供给决策支持系统。知识发现就是专门描述这些方法和技术的。它是建立在统计学、数据库技术、模式识别、机器学习等诸多领域之上的一门新学科。

数据挖掘作为知识发现这一过程的一个环节，指应用数据分析和数据发现算法，从数据库中获取潜在可用的模式或指导性规则的过程。

数据挖掘对传统的统计学提出了更高的要求，也推动了统计学的发展。本文对数据挖掘的分类过程中经常用到的统计方法给出较为全面的介绍，包括对原始数据的预处理、分类算法以及分类结果分析。

2. 数据挖掘的基础——数据

数据挖掘所获取的模式或模型的好坏在很大程度上依赖于所提供的数据。虽然，在很多实际系统中均不同程度地存在着不确定性因素，使采集到的数据常常包含着噪声，不精确。但分析人员应尽量确保这些数据蕴涵丰富的信息。

数据质量以其完整性、可分离性及分析人员所能理解的程度为特征。数据的完整性是指记录完整，无丢失数据；数据可分离性是指在特征化后的数据集内可以轻而易举地进行分类。

从数据类型分，原始数据可分为数字型与字符型两种。数字型数据包括整形数和浮点数，其数值之间的差异可以用问题域中在数字意义方面不同的函数来描述。字符型数据可用数字和字符来表现，它所取的每一个可能的值，其意义是等价的。针对不同类型的数据，有不同的数据挖掘处理方法。

从数据库容量来讲，现代生活中的数据库往往包含上百万条记录，包含上千兆(10^9)和上千京(10^{12})字节的数据库也是司空见惯的。比如，美孚石油公司需要对石油勘探中超过10万京(10^{14})字节的数据进行存储(1993)；美国电话电报公司拥有1亿顾客，其远程网络上每天就有2亿次呼叫记录(1997)[1]。如此庞大的数据库，很难用有意义的形式将其表现出来。所以在对数据进行深入分析之前，最好能对数据进行预分类。比如，把数据看作多维空间的点集，寻找数据中是否存在自然的聚类。

3. 预处理技术

数据挖掘过程中使用预处理技术或者特征提取(比如数据净化)可以获得对数据信息的更深入理解。一般来讲，利用预处理的结果进行模型挖掘，比用原始数据更方便、有效，速度、准确率都将有所提高。

常用的预处理技术有两个：数据标准化与主成分分析。

①数据标准化

两个最基本的数据标准化方法是：对输入数据的属性进行均值归零和方差标准化。

记第 j 类模式向量第 k 个样本的第 i 个属性为 a_{ikj}，第 j 类的样本数为 M_j，类别总数为 N。易得，所有类的所有样本的第 i 个属性的均值为：$E\{a_i\}=\frac{1}{N}\sum_{j=1}^{N}\frac{1}{M_j}\sum_{k=1}^{M_j}a_{ikj}$

相应的零均值属性为 $\alpha_{ikj}=a_{ikj}-E\{a_i\}$。

所有类的所有样本的第 i 个属性的方差为：$V\{a_i\}=\frac{1}{N}\sum_{j=1}^{N}\frac{1}{M_j}\sum_{k=1}^{M_j}(a_{ikj}-E\{a_i\})^2$

相应的标准化方差为 $\beta_{ikj}=\frac{a_{ikj}}{\sqrt{V\{a_i\}}}$。

②主成分分析(PCA)

主成分分析方法是通过把相关属性转变为不相关成分(如原始分量的线性组合)的方法降维，以简化问题，从而降低

魏 玲 博士生，讲师，主要研究方向为概率统计，机器学习；祁建军 博士生，主要研究方向为机器学习，网络安全；张文修 教授，博士生导师，主要研究方向为人工智能。

分类器的计算复杂性。它用较少的变量来代替原来较多的变量，而这些较少的变量尽可能多地反映原来变量的信息。但是，由于主成分分析方法是一种无监督的特征提取方法，因此，用来代替原来模式的新的不相关特征（由 PCA 所提取）对于模式分类来讲，并非是最好的。所以，使用主成分分析方法可以非常有效地降维，但并不有助于提高分类准确率。

4. 数据挖掘中的常见分类算法

对不同的数据集，最优的模型算法往往是不同的，常见的有人工神经网络、决策树、统计分类、遗传算法等。

针对统计分类，可细分为参数统计分类与非参数统计分类两种。

用于分类的参数统计方法是用类别所属的分布寻找数据集的分类平面，从而将数据进行分类。判别分析是用于判别个体所属群体的一种统计方法，通常适用于参数较少的情况。这类判别分析模型属于古典统计方法，通过寻找特征变量的线性或二次组合来区分类界限。如 Fisher 线性判别，基于极大似然的二次判别（分类面是二次曲面），以及对数判别。

非参数方法则不需要假设基本的概率分布，而是采用类分布的局部密度估计。典型的方法如 K——最近邻方法，根据 K 个到特性空间最近的点的属性来确定。近邻规则为：给定 N 个待分类的模式$\{x_1,\cdots,x_n\}$，要求按距离阈值把这些模式分到聚类中心 $z_1,\cdots,z_N$。当分类概率非常光滑地分布于特性空间时，该方法是有效的，在实际应用中有较高的分类准确性；但遗憾的是选择合适的 K 及距离是个难题，且使用此方法的计算量较大。其它的非参数统计方法还有密度估计、投影寻踪等。

近年来，基于概率统计的一些新方法被用于数据挖掘，如贝叶斯网络。贝叶斯网络最初主要在专家系统中用来表述不确定的专家知识。90年代中、后期，才逐渐被国外的科技人员用于数据挖掘[2~4]。目前，贝叶斯网络已成为数据挖掘领域的非常重要而且有效的方法之一，也是研究热点之一。

贝叶斯网络是用来表示变量集合的连接概率分布的图形模型，它提供了一种自然地表示因果信息的方法。贝叶斯网络本身并没有输入和输出的概念，各结点的计算是独立的，因此，贝叶斯网络的学习既可以由上级结点向下级结点推理，也可以是由下级结点向上级结点的推理。用于数据挖掘的贝叶斯网络方法主要有以下几个特点[5]：

(1)贝叶斯网络可以处理不完整和带有噪声的数据集。它用概率测度的权重来描述数据间的相关性，从而解决了数据间的不一致，甚至是相互对立的问题。

(2)贝叶斯网络用图形的方法描述数据间的相互关系，语义清晰，可理解性强，这将有助于利用数据间的因果关系来进行预测分析。

(3)由于贝叶斯网络具有因果和概率性语义，它有助于先验知识和概率的结合，容易与优化决策方法相结合。

用贝叶斯网络找出数据之间潜在的关系，正是数据挖掘所需要完成的功能。但是利用贝叶斯网络进行数据挖掘，主要问题是先验知识的重要性。由于不可能对所有的网络结构进行计算，特别是当变量增多时，可能的网络结构成倍增加，因此必须在现有的知识下进行网络选择，这在很大程度上依赖于专家知识。

5. 分类结果分析

分类终止后，必须充分了解分类结果的意义与有效性。

首先，评价分类器的性能优劣，即评价模型对数据的拟合程度。可以通过以下三种方法来进行：把数据分为训练集和检测集，分别用于对数据的训练和检测（检测是否发生过度训练）；样本容量大时用交叉验证方法做深层分析；样本容量小时用自举方法代替交叉验证。

其次，以错误分类代价最小为原则，用代价函数对分类错误进行加细。

第三，用假设检验的方法分析分类检验的统计意义。

分类结果的统计分析是基于混淆矩阵进行的[6]，该矩阵显示被分类器正确分类与错误分类的数目。

假定含有 N 个样本的数据集中包括 A 和 B 两类，其相应的混淆矩阵如下所示：

		实际类别 A	实际类别 B
预报类别	A	N_{11}	N_{12}
	B	N_{21}	N_{22}

其中，N_{11}为类别 A 中被正确分类的样本数，N_{22}为类别 B 中被正确分类的样本数，N_{21}为类别 A 中被错误分类的样本数，N_{12}为类别 B 中被错误分类的样本数；

且 $N_{11}+N_{12}+N_{21}+N_{22}=N$。

检验分类显著性的标准统计方法是卡方检验，采用的卡方统计量是：

$$K=\sum_{i=1}^{m}\frac{(O_i-E_i)^2}{E_i}$$

其中，m：样本的类别数，O_i：第 i 类中观察到的样本数，E_i：第 i 类的期望频数。

根据卡方检验方法，K 值大到一定程度，则认为分类结果不随机，分类具有统计意义。

结束语 数据挖掘是一个新兴的研究领域，存在着很多问题与困难。本文仅对其分类过程中可能采用的统计方法做了较为全面的介绍与分析，有关数据挖掘的更多的内容这里暂略。与传统的统计学方法相比，数据挖掘的重点多数在“学习”上，更关注模型的复杂性和所需的计算量，而较少放在大样本的渐近推论上；且某些模型技术，并不能将其归于古典的统计框架中。但是，在数据挖掘的过程中，统计模型的适用是非常重要的；反过来，统计学也不断向智能化靠近。所以，数据挖掘中采用的统计方法不应该是单一的，而必须和其它领域的知识相结合，才能高效、高速地获取数据挖掘的目标。

参 考 文 献

1 Hand D J. Data Mining: Statistics and More?. The American Statistician, 1998,52(2):112~118

2 Heckman D, Geiger D, Chickering D. Learning Bayesian Networks: the combination of knowledge and statistical data. Machine Learning,1995,20(3):197~243

3 Heckman D. A Bayesian approach for learning causal networks. In: Besnard P, Hanks S, eds. Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence. CA: Morgan Kaufmann Publishers, Inc., 1995. 285~295

4 Heckman D. Bayesian Networks for Data Mining, Data Mining and Knowledge Discovery,1997. 79~119

5 Chickering D. Learning equivalence classes of Bayesian networks structures, In: Horvitz E, Jensen F, eds. Proc. of the 12th Conference on Uncertainty in Artificial Intelligence. San Francisco, CA: Morgan Kaufmann Publishers, Inc., 1996. 54~61

6 Massy W F. On methods: discriminant analysis of audience characteristics. J of Advertising Research,1965,15:39~48