中国农业大学学报 2003,8(2):39～41
Journal of China Agricultural University

数据挖掘技术在高等学校决策支持中的应用

陶 兰[1,2] 王保迎[1] 吕建军[1]
(1.中国农业大学 信息与电气工程学院,北京 100083;2.深圳大学信息工程学院,深圳 518060)

摘 要 为有效利用高等学校教学管理工作多年来积累的大量数据,利用数据挖掘技术,对北京地区高等学校1996—2001年毕业生数据库(Beijing Graduation Database,BGD)进行了数据挖掘研究。采用多种数据预处理方法对原始BGD数据进行了处理,提出并利用FAP方法进行了属性构造;对关联规划挖掘常用的Apriori算法进行了改进,以此为基础根据实际需要设计并实现了关联规则挖掘系统;利用所实现的系统对GBD数据库进行挖掘分析,得到了有益于高等学校教学管理决策及毕业生就业指导的挖掘结果。

关键词 数据挖掘;北京毕业生数据库;关联规则;Apriori算法
中图分类号 TP 18;TP 311 **文章编号** 1007-4333(2003)02-0039-03 **文献标识码** A

Research on the application of Data Mining techniques

Tao Lan[1,2], Wang Baoying[1], L üJianjun[1]
(1. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China;
2. Faculty of Information Engineering, Shenzhen University, Shenzhen 518060, China)

Abstract To utilize the data accumulated by universities about pedagogic management effectively, the BGD(Beijing Graduation Database, BGD from 1996 to 2001) was researched by using DM(Data Mining) method. After thorough analysis of BGD,BGD was preprocessesd by many methods to standardize this method and it is made fit for knowledge discovery. In the pro-management progress, FAP method was put forward and adopted, which is used to construct attributes. Association Rules was researched and Apriori algorithm was improved. Association Rule was designed and realized according to practical requirement on the base of improved Apriori algorithm. It is used to mine BGD and some results which is helpful to decision making of universities.

Key words Data Mining; Beijing Graduation Database; association rule; apriori algorithm

高等学校多年来的教学和管理工作积累了大量的数据,目前这些数据还未能得到有效利用,只是一个待开发的"宝藏"。鉴于社会对高等学校发展的需求和目前高校数据管理现状,利用这些数据理性地分析学校各方面工作的成效以及学生培养过程中的得失变得十分重要。数据挖掘技术能从大量数据中发现有用的知识,这些知识对高等学校教学管理的决策支持将是十分有意义的,而此类研究目前国内尚不多见。笔者尝试将数据挖掘技术引入高等学校教学管理之中,主要通过1996—2001年北京市高校毕业生数据(Beijing Graduate Database,BGD)分析北京市高校以及中国农业大学毕业生就业情况,以期发现对学校教学管理、学生管理有用的信息,并对高等学校数据管理工作提出建议。

1 BGD数据库及其数据预处理分析

1.1 BGD数据库简介

由北京市毕业生就业指导中心提供的BGD数据库,是自1996年以来北京各高等学校上报给北京市教委的毕业生派遣情况数据库。每年由各学校就业指导部门将毕业生的个人基本信息、就业情况等按规定格式上报,由北京市就业指导部门汇总。原始数据库为关系数据库,由6个子数据库组成,每个子数据库由若干个关系表组成,每个学校一张dbf

收稿日期:2002-10-08
作者简介:陶 兰,博士生导师,教授,主要从事智能信息处理、语义Web、数据挖掘和智能Agent的研究。

表,每张表共有 61 个属性(字段)。

原始数据库存在如下问题:1)数据不完整,存在大量的空缺值;2)含噪声数据,存在大量冗余和噪声数据;3)数据不一致,原始数据取自各实际应用系统,而各应用系统的数据缺乏统一标准,数据结构也有较大差异;4)不同的数据挖掘算法对数据有相应的要求,因此在挖掘之前需要对原始数据进行大量的预处理工作,以减少挖掘过程中的故障,提高数据挖掘模式的质量,降低实际挖掘所需要的时间。

1.2 BGD 数据库预处理[1,2]

1)数据集成。主要将多文件或多数据库运行环境中的异构数据进行合并处理,解决语义的模糊性。该部分主要涉及数据的选择、数据的冲突问题,以及不一致数据的处理问题。原始 BGD 数据库的 6 个库可集成为一张表。数据合并后降低了处理中发生错误的可能性,同时给以后的数据变换提供了方便。在集成的过程中需考虑以下问题的一致性:字段数据格式、字段长度、字段名称等。

2)数据清理。去除源数据集中的噪声数据和无关数据,处理遗漏数据和清洗脏数据,考虑时间顺序和数据变化等。主要包括噪声数据处理和缺值数据处理,并完成一些数据类型的转换。本文数据清理工作包括空缺值、噪声数据和不一致数据的处理。

3)数据变换。主要是找到数据的特征表示,用维变换或转换方法减少有效变量的数目或找到数据的不变式,将数据转换成适合于挖掘的形式。对于 BGD 数据库,主要通过以下方式进行数据变换:数据概化,学校名称代码统一,“出生日期”字段转换,代码转换成数值型。

4)属性构造。BGD 数据库中就业方面的信息涉及到 14 项,30 个字段。这些信息都是以“毕业去向”、“单位名称”、“主管单位”、“单位所在地”、“单位行业”、“单位类型”、“单位所有制”、“重点类型”等形式描述,无法进行量化分析,因此需要采用科学方法对有关信息进行评估。本文中采用 FAP(Fill in with Average Poll result)法[2]对就业信息进行属性构造,以便评估分析。

5)数据归约。BGD 数据库有 24 万多条记录,在这样大的数据集上进行挖掘所需时间将较长。数据归约技术可以用来得到数据集的归约表示,它虽然小得多,但仍接近于保持源数据的完整性。本文中所采用的数据归约方法为:数据立方体聚集和维归约。

6)布尔转换。本文中对 BGD 数据库进行关联规则发现,须将 BGD 数据库进行布尔转换。

2 BGD 数据库关联规则挖掘

2.1 关联规则挖掘算法及改进[1,2]

2.1.1 Apriori 算法:使用候选项集寻找频繁项集

Apriori 算法是一种常用的布尔关联规则频繁项集的挖掘算法。该算法使用一种称作逐层搜索的迭代算法,k-项集用于搜索$(k+1)$-项集。首先,找出频繁 1-项集的集合,记作 L_1。L_1 用于寻找频繁 2-项集的集合 L_2,而 L_2 用于寻找 L_3,如此下去,直到不能找到频繁 k-项集。寻找每个 L_k 需要一次数据库扫描。利用 L_{k-1}寻找 L_k 的过程由连接和剪枝组成。

2.1.2 对 Apriori 算法的改进

Apriori 算法目前经常被使用,但它存在很多不足,其中主要是算法的效率问题。为了提高挖掘速度,对 Apriori 算法进行改进,其基本思想为:对该数据库中的每一条记录增加一个标志位 flag,初始值均为 1。由于不包含任何 k-项集的事务,故不可能包含任何$(k+1)$-项集,这样,在扫描数据库时,没有涉及到的事务可以将标志位 flag 置 0,在产生 j-项集$(j>k)$时不再需要它们,可略过标志位 flag 为 0 的记录。这样将大大减小比较的工作量,利于算法的实现和应用,并有助于加快运算速度。

2.1.3 由频繁项集产生关联规则[1]

一旦由数据库 D 中的事务找出频繁项集,由它们产生强关联规则就较容易了(强关联规则满足最小支持度和最小置信度)。置信度可以用式(1)获得,其中条件概率用项集支持度计数表示。

$$\text{confidence}(A\Rightarrow B)=P(A|B)=\frac{\text{support\_count}(A\cup B)}{\text{support\_count}(A)} \tag{1}$$

式中:support_count$(A\cup B)$是包含项集$(A\cup B)$的事务数,support_count(A)是包含项集 A 的事务数。根据该式,关联规则可以产生如下:

® 对于每个频繁项集 l,产生 l 的所有非空子集。

® 对于 l 的每个非空子集 s,如果

$$\frac{\text{support\_count}(l)}{\text{support\_count}(s)}\geq \text{min\_conf}$$

则输出规则“$s\Rightarrow(l-s)$”。其中 min_conf 是最小置信度阈值。

由于规则由频繁项集产生,故每个规则都自动满足最小支持度。

2.2　BGD 数据库关联规则挖掘

在对 BGD 数据库进行关联规则挖掘时,首先用 1.2 节讨论过的各种方法对 BGD 数据库进行数据预处理,然后利用改进后的关联规则挖掘算法进行挖掘。限于篇幅,在此仅将挖掘结果中的部分关联规则和结论摘录如下:

规则 1　某大学 ∧工学专业 ∧男生 ∧本科毕业 ∧学制五年 ∧学士 ⇒就业系数为 9,support = 0.69, confidence = 0.73(表明此大学工学专业五年制本科男毕业生的就业情况很好)。

规则 2　北京市属的一些小院校 ⇒就业系数为 0～3,support = 0.72,confidence = 0.78(可见北京市属的一些小院校学生多,生源质量较差。虽然这些学生大都能就业,但就业去向大多数学生不看好)。

规则 3　男生 ⇒工学,support = 0.61,confidence = 0.72(表明工学专业男生占大多数。)。

规则 4　党员 ⇒就业系数 7～9,support = 0.73, confidence = 0.75(表明党员或预备党员就业情况较好。高校学生党员综合素质一般较高,较受社会的欢迎,因此应加强学生的综合素质教育)。

规则 5　工学专业 ⇒就业系数 7～9,support = 0.58,confidence = 0.70(表明工学专业学生较受社会欢迎)。

结论 1　对整个数据库进行挖掘,与其他项目具有强关联的就业系数集中在 4～6,与其强相关的项目大多是团员、汉族、本科毕业生、学制四年、学籍无变动等项目。这符合目前毕业生的就业情况。

结论 2　若考察与就业系数相关性较高的项目,发现就业系数为 7～9 的数据集中在北京 2 所名校,其数量超过总数的 1/3。其原因与这 2 所学校毕业生大多出国留学,或读研究生有关。

结论 3　男生的就业情况和女生没有太大差别,北京市男女就业系数基本持平,中国农业大学女生平均就业系数(5.47)稍好于男生(5.35),这与人们普遍认为女生就业难是不相符的。

结论 4　生源为华东地区的学生就业系数为 7～9 的占总数的 41%,比其他地区都高,华北地区次之。建议在华东地区扩大招生,或对其他地区生源的学生加强就业指导。

结论 5　中国农业大学毕业生的就业情况 1996—1998 年呈上升趋势,1998 年达到顶峰,平均系数为 5.55;1999 年下降至 5.19,2000 与 2001 年基本持平,平均系数为 5.50。与北京市相比,该校的就业情况不容乐观(北京市平均系数为 6.50)。

结论 6　中国农业大学毕业生中,与就业系数强相关的项无有意义的特征。单考察就业信息,总系数为 7～9 的就业去向基本上为考研究生或出国,因此希望将来能与学生在校情况相结合,综合分析学生就业的影响因素,如成绩等。

3　关联规则挖掘系统的设计与实现

针对 BGD 数据库以及其他学生管理数据,笔者设计并实现了一个发现 BGD 数据库特征模式的专用数据挖掘系统。系统基于 Windows 2000 Server 平台,采用 SQL Server 2000 数据库,用 Visual C++ 6.0 作为开发工具。系统总体结构如图 1 所示。

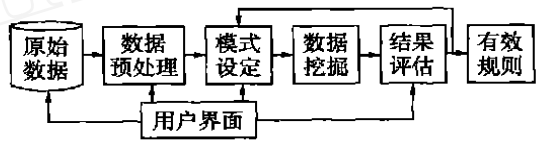

图 1　BGD 数据挖掘系统结构图

Fig. 1　The structure of BGD Data Mining system

1)用户界面:图形化的用户界面简单易用,为用户提供其在挖掘的各阶段与系统进行交互的接口。

2)数据预处理:对数据进行多种预处理工作。

3)模式设定:供用户进行支持度、置信度、年度、学校等信息设定。

4)数据挖掘:利用改进后的关联规则算法对数据挖掘库中的数据进行挖掘。

5)结果评估:将数据挖掘模块所获结果以可视化的形式表示。用户可对结果进行评价。

4　结束语

利用所建的系统和数据挖掘工具对高校数据进行挖掘,所得结论对高等学校教学管理、教育行政部门的决策,尤其对学生就业指导部门的工作提供了有益的参考。

参　考　文　献

[1] Han Jiawei, Kamber M. Data Mining: concepts and techniques[M]. San Fransisco: Morgan Kaufmann Publishers, 2001. 225～244

[2] 吕建军. 数据挖掘技术的应用研究[D]. 北京:中国农业大学, 2002