基于判定树的项目 R &D 中的数据挖掘质量测评研究 *

Study on the Quality Evaluation of the Data Mining in the Project R&D Based on Decision Tree

陆 瑶[1] 张 杰[2] 冯英俊[1]

(1. 哈尔滨工业大学管理学院 哈尔滨 150001;2. 山东科技大学经济管理学院 青岛 266510)

摘 要 项目 R &D 活动中涉及大量的信息管理和知识管理,而数据挖掘是信息管理和知识管理的基础,因此,数据挖掘的质量关系到项目 R &D 活动中信息管理的水平和知识管理的绩效,也关系到项目 R &D 活动的成败。基于此,首先论述了项目 R &D 活动中数据挖掘的涵义及其质量问题,接着提出了项目 R &D 活动中数据挖掘质量的测评指标,并构建了评价指标体系框架,最后运用判定树归纳算法对若干项目 R &D 活动中数据挖掘质量进行了实证评价和分析,以便为项目 R &D 活动提高数据挖掘质量的测评提供借鉴和方法论指导。

关键词 判定树 项目研发(R &D) 数据挖掘 质量测评

中图分类号 TP274 **文献标识码** A **文章编号** 1002 - 1965(2009)05 - 0048 - 04

项目 R &D 活动在当前的形势下具有重要的意义,它是企业生存和发展的关键,也是一个国家实现技术升级和产业更新的重要途径,它关系到一个国家的自主品牌的发展和科技实力的增强,也是一个国家掌握更多核心技术的重要渠道。项目 R &D 活动是以信息集成和知识创新为特征,信息管理和知识管理无疑成为项目 R &D 的最重要内容和活动,无论是信息和知识的获取、共享和运用,都涉及到大量的数据及其处理问题,这就决定着数据挖掘是项目 R &D 活动中的信息管理和知识管理的起点和基础,进而,数据挖掘的质量关系到项目 R &D 活动的效率和成败。因此,数据挖掘质量的测评对项目 R &D 活动和数据挖掘本身都具有重要的意义。

1 项目 R &D 中的数据挖掘的涵义及质量问题

数据挖掘(Data Mining,简称 DM)是 20 世纪 90 年代中期兴起的一项新技术,是多门学科和多种技术相结合的产物。1989 年在美国底特律召开的第 11 届国际人工智能联合会议的专题讨论会上,首次提出了知识发现(KDD)这个概念,1995 年,美国计算机学会(ACM)会议提出了数据挖掘。所谓的数据挖掘,顾名思义就是从大量的数据中挖掘出有用的信息[1~2]。一般人们认为它是从大量的、不完全的、有噪声的、模糊的、随机的数据集中识别有效的、新颖的、潜在有用的以及最终可理解模式的非平凡过程。它也可理解为是在一些事实或观察数据的集合中寻找模式的决策支持过程。数据挖掘是一门涉及面很广的交叉学科,包括机器学习、数理统计、神经网络、数据库、模式识别、粗糙集、模糊数学等相关技术,它也常被称为"知识发现"[3~4]。相对来讲,"数据挖掘"主要流行于统计界、数据分析、数据库和管理信息系统界;而"知识发现"则主要流行于人工智能和机器学习界。

数据挖掘质量的核心问题是数据挖掘的质量,数据挖掘的质量与源数据的质量、数据集成时的质量及数据分析时的质量密不可分,同时,数据挖掘算法对数据有一定的要求,如数据冗余性小、数据属性之间相关性小、数据出错率小等。而现实中,所得到的数据总是存在着一些质量问题,最常见的问题包括数据缺失、测量误差、采样失真、人为错误等。很多情况下数据挖掘对数据的采集方式没有任何控制,有时数据集可能是

收稿日期:2008 - 11 - 10 修回日期:2009 - 01 - 12

基金项目:本文系山东省统计科研重点课题"基于问卷调查的统计数据质量的测评与控制研究"(编号:KT0832)和青岛市软科学课题"青岛市企业高新技术创新项目 R &D 绩效评价研究(07R - 22)资助项目。

作者简介:陆 瑶,女,1977 年生,博士研究生,研究方向为管理有效性、技术经济与管理及虚拟经济研究;张 杰,男,1975 年生,工学博士,副教授,研究方向为数学建模、复杂大系统的应用研究。

所希望描述的总体失真样本,所以数据的质量问题更加重要。

为了有效发挥出数据挖掘技术的作用,必须对这些问题进行处理,从而使得数据挖掘的有效性和质量得到切实的保障。

2　项目 R &D 中的数据挖掘质量的测评体系

目前学术界对数据挖掘质量问题的研究基本上局限于数据预处理方面,即源数据的质量问题研究,本文从数据挖掘的全局出发,从数据挖掘的整个过程来研究数据挖掘的质量问题。数据挖掘的结果及其质量与用于挖掘的源数据的质量息息相关,也与数据挖掘过程中的处理方法与控制技术有关,同时还与数据分析的角度与要求有关。

因此,高质量的数据挖掘结果依赖于高质量的源数据、科学的数据集成和客观的数据分析,否则就会出现所谓的"垃圾进,垃圾出"的现象。根据项目 R &D 活动中的数据挖掘整个过程的不同步骤对数据挖掘质量问题进行分类,可以分为以下三类:数据准备阶段的源数据质量问题、数据集成时的质量问题和数据分析时的质量问题。

随着各种技术的不断发展,收集和积累数据的技术和渠道日益广泛,通过各种技术和渠道收集和积累的数据储存在项目 R &D 活动的数据库或数据仓库中,构成了项目 R &D 活动用于数据挖掘的源数据。但是由于各种各样的原因,如市场调查中的无回答,数据输入错误等,导致了源数据的各种质量问题。主要包括:数据缺失、数据异常和数据重复等问题;由于每一个数据源都是为了满足特定的需要而进行设计的,其结果是在数据库管理系统、数据编码、数据模式、数据格式等方面都存在很大的不同,所以在将多个数据源进行集成时数据时,这些问题表现的尤为突出,主要包括:数据编码冲突、数据模式冲突、数据本身冲突和数据冗余;运用数据挖掘技术进行数据分析时,面对数据挖掘库中的海量数据,常常会为了提高速度而选择部分数据或部分属性用于数据分析,由此导致了数据分析时的质量问题,主要包括:代表性问题、转换性问题、生成性问题及模式或模型选择性问题。

根据项目 R &D 活动中数据挖掘的内容和禀性,从项目 R &D 活动的角度出发,在文献调研和专家咨询的基础上,结合上文的论证,将项目 R &D 中的数据挖掘质量测评指标体系分为源数据质量问题、数据集成时的质量问题和数据分析时的质量问题三个维度,每个维度分为若干个指标,其具体的测评指标体系框架,如图 1 所示。

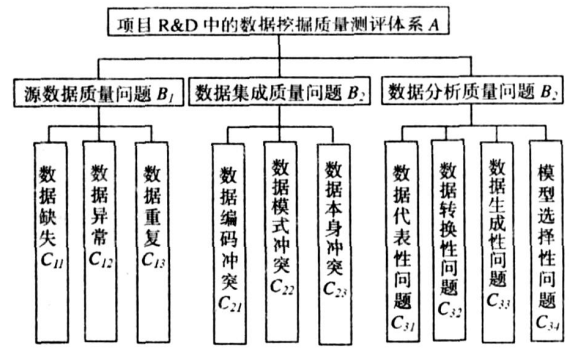

图 1　项目 R &D 中的数据挖掘质量的测评体系

3　项目 R &D 中的数据挖掘质量的测评方法及实证评价

3.1　判定树算法及其基本原理

3.1.1　判定树简介。判定树是一个类似于流程图的树结构,其中每个内部节点表示在一个属性上的测试,每个分支代表一个测试输出,而每个树叶节点代表类或类分布。判定树由决策结点、分支和叶子组成。判定树中最上面的结点为根结点,每个分支是一个新的决策结点,或者是树的叶子。每个决策结点代表一个问题或决策,通常对应于待分类对象的属性。每一个叶子结点代表一种可能的分类结果。沿判定树从上到下遍历的过程中,在每个结点都会遇到一个测试,对每个结点上问题的不同测试输出导致不同的分支,最后会到达一个叶子结点,这个过程就是利用判定树进行分类的过程[5~6]。

3.1.2　判定树算法的基本策略。判定树的基本算法是贪心算法,它以自顶向下递归的各个击破方式构造判定树。算法的基本策略如下[7]:a. 树以代表训练样本的单个节点开始。b. 如果样本都在同一个类,则该节点成为树叶,并用该类标记。c. 否则,算法使用称为信息增益的基于熵的度量作为启发信息,选择能够最好地将样本分类的属性。该属性成为该节点的"测试"或"判定"属性。所有的属性都是分类即取离散值的。连续值的属性必须离散化。d. 对测试属性的每个已知的值,创建一个分支,并据此划分样本。e. 算法使用同样的过程,递归地形成每个划分上的样本判定树。一旦一个属性出现在一个节点上,就不必考虑该节点的任何后代。f. 递归划分步骤仅当下列条件之一成立时停止:给定节点的所有样本属于同一类;没有剩余属性可以用来进一步划分样本;分支已没有样本。

3.1.3　判定树算法中的属性选择变量[8]。该树的每个节点上使用信息增益(Information Gain)度量选择测试属性。这种度量称为属性选择变量或分裂的优良性度量。选择具有最高信息增益(或最大熵压缩)的

属性作为当前节点的测试属性。该属性使得对结果划分中的样本分类所需的信息量最少,并反应划分的最小随机性或"不纯性"。这种信息理论方法使得对一个对象分类所需的期望测试数目达到最小,并确保找到一棵简单的树。

设 S 是 s 个数据样本的集合。假定类标号属性具有 m 个不同值,定义 m 个不同类 $C_i(i = 1,2,\cdots,m)$。设 s_i 是类 C_i 中的样本数。对一个给定的样本分类所需的期望信息由下式给出:

$$I(s_1,s_2,\cdots,s_m) = -\sum_{i-1}^{m} p_i \log_2(p_i) \qquad (1)$$

其中 p_i 是任意样本属于 C_i 的概率,并用 s_i/s 估计。注意,对数函数以 2 为底,因为信息用二进位编码。

设属性 A 具有 v 个不同值$\{a_1,a_2,\cdots,a_v\}$。可以用属性 A 将 S 划分为 v 个子集$\{S_1,S_2,\cdots,S_v\}$;其中,S_j 包含 S 中这样一些样本,他们在 A 上具有值 a_j。如果 A 选作测试属性(即最好的分裂属性),则这些子集对应于由包含集合 S 的节点生长出来的分枝。设 s_{ij} 是子集 S_j 中类 C_i 的样本数。根据由 A 划分成子集的熵(entropy)或期望信息由下式给出:

$$E(A) = \sum_{i=1}^{v} \frac{s_{1j}+s_{2j}+\cdots s_{mi}}{s} I(s_{1j},s_{2j},\cdots,s_{mj}) \qquad (2)$$

项 $\frac{s_{1j},s_{2j},\cdots,s_{mi}}{s}$ 充当第 j 个子集的权,并且等于子集(即 A 值为 a_j)中的样本个数除以 S 中的样本总数。熵值越小,子集划分的纯度越高。注意,对于给定的子集 S_j,有:

$$I(s_{1j},s_{2j},\cdots,s_{mj}) = -\sum_{i-1}^{m} p_{ij}\log_2(p_{ij}) \qquad (3)$$

其中,$p_{ij} = \frac{s_{ij}}{|S_j|}$ 是 S_j 中的样本属于类 C_i 的概率。

在 A 上分枝将获得的编码信息是:

$$\mathrm{Gain}(A) = I(s_1,s_2,\cdots,s_m) - E(A) \qquad (4)$$

换言之,$\mathrm{Gain}(A)$ 是由于知道属性 A 的值而导致的熵的期望压缩。

算法计算每个属性的信息增益。具有最高信息增益的属性选作给定集合的测试属性。创建一个节点,并以该属性标记,对属性的每个值创建分枝,并据此划分样本。

3.2 项目 R&D 中的数据挖掘质量的实证评价

3.2.1 数据实例及研究步骤。高质量的决策必须依赖于高质量的数据,然而现实世界中的数据极易受噪声数据、空缺数据和不一致性数据的侵扰,为了提高项目 R&D 活动中数据挖掘的质量,必须对采集到的数据有所选择,以保证其属性的完整性,并在此基础上离散化以适于后面的数据挖掘处理。

在表 1 中给出了一个数据训练集,表格数据按照属性值格式显示,第一列显示表中属性值的属性名称。从表的第二列开始,每一列都是一个项目 R&D 中的数据实例,最后一行显示每个项目 R&D 活动数据挖掘质量的等级(评判结果)。

表 1 项目 R&D 活动数据挖掘质量的训练数据

评价指标	1	2	3	4	5	6	7	8	9	10
C_{11}	是	否	是	否	否	否	是	否	是	是
C_{12}	是	是	是	否	是	是	否	否	是	否
C_{13}	是	是	是	否	是	是	否	否	否	否
C_{21}	是	否	是	是	是	否	否	是	否	是
C_{22}	是	否	是	否	是	否	否	否	否	否
C_{23}	是	否	是	否	是	否	否	否	是	否
C_{31}	是	是	否	否	是	是	否	否	否	否
C_{32}	是	否	否	是	是	否	否	是	否	否
C_{33}	是	否	是	否	是	否	否	否	是	否
C_{34}	是	是	否	否	是	是	否	否	否	否
判定结果	差	一般	差	良	差	一般	良	良	一般	良

在表 1 数据训练集中,类标号属性"判定结果"有 3 个不同的值("差","一般","良"),因此有 3 个不同的类,设类 M_1 对应于"差",类 M_2 对应于"一般",类 M_3 对应于"良"。类 M_1 有 3 个样本,类 M_2 有 3 个样本,类 M_3 有 4 个样本。使用式(1)计算对给定样本分类所需的期望信息:

$$I(s_1,s_2,s_3) = I(3,3,4)$$

$$= -\frac{3}{10}\log_2\frac{3}{10} - \frac{3}{10}\log_2\frac{3}{10} - \frac{4}{10}\log_2\frac{4}{10} = 1.571$$

接着,需要计算每个属性的熵。先计算属性为"数据缺失"的熵。观察属性"数据缺失"的每个样本值为"差"、"一般"和"良"的分布,对每个分布计算期望信息。

对于数据缺失 C_{11} = "是" $s_{11} = 2, s_{21} = 1, s_{31} = 2$,使用(3)式得:

$$I(s_{11},s_{21},s_{31}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{1}{5}\log_2\frac{1}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 1.522$$

对于数据缺失 C_{11} = "否" $s_{12} = 1, s_{22} = 2, s_{32} = 2$ 使用(3)式得:

$$I(s_{12},s_{22},s_{32}) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{2}{5}\log_2\frac{2}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 1.522$$

使用(2)式,如果样本按"C_{11}"划分,对一个给定样本分类所需的期望信息为:

$$E(C_{11}) = \frac{5}{10} I(s_{11},s_{21},s_{31}) + \frac{5}{10} I(s_{12},s_{22},s_{32})$$

$$= 15.22$$

使用(4)式,可求出这种划分的信息增益是:

$$\mathrm{Gain}(C_{11}) = I(s_1,s_2,s_3) - E(C_{11})$$

$$= 1.571 - 1.522 = 0.049$$

类似地,可以计算:

$Gain(C_{12}) = I(s_1, s_2, s_3) - E(C_{12}) = 0.726$;

$Gain(C_{31}) = I(s_1, s_2, s_3) - E(C_{31}) = 0.099$;

$Gain(C_{13}) = I(s_1, s_2, s_3) - E(C_{13}) = 0.088$;

$Gain(C_{32}) = I(s_1, s_2, s_3) - E(C_{32}) = 0.036$;

$Gain(C_{21}) = I(s_1, s_2, s_3) - E(C_{211}) = 0.154$;

$Gain(C_{33}) = I(s_1, s_2, s_3) - E(C_{33}) = 0.045$;

$Gain(C_{22}) = I(s_1, s_2, s_3) - E(C_{22}) = 0.436$;

$Gain(C_{34}) = I(s_1, s_2, s_3) - E(C_{34}) = 0.137$;

$Gain(C_{23}) = I(s_1, s_2, s_3) - E(C_{23}) = 0.515$。

4　实证结果分析与评价

由以上计算结果,可知 C12(数据异常)在属性中具有最高信息增益,被选作测试属性。创建一个节点,C12(数据异常)标记(此节点为根节点),并对于每个属性值引出一个分枝。样本按此划分,对每个分枝,再用判定树归纳分类法进行分类,再引出分枝,最后,算法返回的最终判定树如图 2 所示。

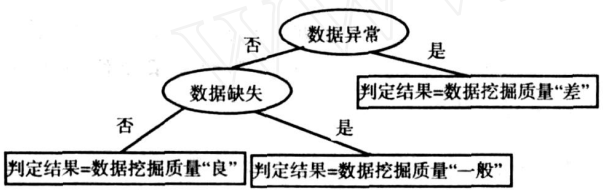

图 2　表 1 中数据判定树

从图 2 中判定树可以看出:将注意力仅仅放在数据挖掘质量是数据异常还是数据缺失上,就可准确地为项目 R &D 中的数据挖掘质量做出判定。像数据重复、数据生成性问题和数据模型选择性问题等这些属性对获得判定结果没有起到任何作用。正如我们能够看到的,判定树概化了数据,并为我们总结出:哪些属性(数据模式冲突、数据本身冲突)和属性的关系对于准确的判定是非常重要的。

另外,还可以将建立的判定树分类法模型所确定的检验集实例分类与正确分类值进行比较,检验集分类的正确性预示模型将来的性能。下面使用判定树对表 2 给出的前两个实例进行分类。

a. 因为编号为 11 的项目 R &D 活动数据挖掘质量测评指标"数据本身冲突"为"是",从判定树的根结点沿着右边的链走,而右链为终极结点,表示项目 R &D 活动数据挖掘质量测评结果为一般,说明该项目 R &D 活动数据挖掘质量不高,处在一般水平。

b. 因为编号为 12 的项目 R &D 活动数据挖掘质量测评指标"数据本身冲突"为"否",从判定树的根结点沿着左边的链走,并检查属性"数据模式冲突"的值,因为数据模式冲突为"否",表示项目 R &D 活动数据挖掘质量测评结果为良,说明该项目 R &D 活动数据挖掘质量较好。

表 2　未知分类数据的实例

编号	C_{11}	C_{12}	C_{13}	C_{21}	C_{22}	C_{23}	C_{31}	C_{32}	C_{33}	C_{34}	判定结果
11	是	否	否	否	是	是	否	是	否	是	?
12	否	否	是	否	否	否	否	否	是	否	?

5　结束语

数据挖掘是一项新兴的技术,它具有广泛的应用前景,在很多领域中都需要用到这项技术,项目 R &D 活动需要进行大量的数据挖掘工作,因此,如何保证数据挖掘的质量是项目 R &D 活动的重要内容和问题,特别的对数据挖掘的质量进行测评更是一个难点问题。本文在阐述数据挖掘质量测评体系的基础上,运用判定树算法给项目 R &D 活动中的数据挖掘质量进行了实证评价,以便为项目 R &D 活动提高数据挖掘质量提供借鉴和方法论指导。

参 考 文 献

1　Jiawei Han, Micheline Kambr. Data Mining Concepts and Techniques[J]. Morgan Kaufmann Publishers, 2000

2　Fayad U M, Simoudis E. Data Mining and Knowledge Discovery. Proceedings of 1st International conf[J]. KDD and Data Mining, 1997

3　Pawlak Z. Rough Set Theory and Its Application to Data Analysis [J]. Cybernetics and System, 1998; (29)

4　王政霞,黄大荣. 基于统计方法的数据挖掘算法研究[J]. 湖北民族学院学报(自然科学版), 2005, 3(1)

5　魏　玲. 数据挖掘中的统计方法[J]. 计算机科学, 2003; (12)

6　Quinlan J R. Induction of Decision Trees[J]. Machine Learning, 1986, 62(1)

7　邵峰晶,于忠清. 数据挖掘原理与算法[M]. 北京:中国水利水电出版社, 2003

8　刘同明. 数据挖掘技术及其应用[M]. 北京:国防工业出版社, 2001

(责编:王平军)