关联规则算法研究及其在教学系统中的应用①

Research on Algorithm of Association Rules and its application in Education System

曲守宁　董彩云　徐德军　吴 桐　(济南大学信息科学与工程学院　250022)

摘要:本文通过对关联规则挖掘算法 Apriori 算法的分析与研究,指出了其在实用中存在的主要问题。提出了与以往改进算法不同的策略,即在预处理阶段引入聚类分析,以此对关联规则算法进行改进,实现两种算法相结合的挖掘,并给出了基于聚类的关联规则改进算法描述。最后将算法应用到学生学习指导中,得到了合理的结果,实验表明了该算法的有效性。

关键词:数据挖掘　关联规则　聚类分析　K－MEANS 算法

随着基于园区网络教务管理数据仓库中学生成绩记录的急剧增长,很难直接根据学生的成绩数据分布找出规律,并根据此规律进行决策指导学生学习。数据挖掘技术可以用于从大量的数据中发现隐藏于其后的规律或数据间的关系,它通常采用机器自动识别的方式,不需要更多的人工干预。采用数据挖掘技术,可以为用户的决策分析提供智能的、自动化的辅助手段。本文在研究关联规则算法的基础上,提出了一种新的改进关联规则的方法,提高了挖掘的效率,将其应用在学习指导数据挖掘系统上,能对学生学习提供有利的指导。

1　关联规则挖掘算法

关联规则[1,2]的概念首先由 R. Agrawal 等人提出,是描述数据库中数据项(属性、变量)之间所存在的(潜在)关系的规则。目前已成为数据挖掘中非常重要的一个研究方向。

发现关联规则要经过以下四个步骤:

(1) 预处理与挖掘任务有关的数据。根据具体问题的要求对数据库进行相应的操作,从而构成规格化的数据库 D;

(2) 针对 D,求出所有满足最小支持度的项集,即大项集。由于一般情况下我们所面临的数据库都比较大,所以此步是算法的核心;

(3) 生成满足最小置信度的规则,形成规则集 R;

(4) 解释并输出 R。

经典关联规则挖掘算法 Apriori,它是一种找频繁项集的基本算法。算法的核心主要在寻找频繁项目集上。主要是基于 Apriori 性质:频繁项集的所有非空子集都必须也是频繁的。利用这个性质可以有效地压缩搜索空间。算法主要思路如下:为找 L_k,通过 L_{k-1} 与自己连接产生候选 k－项集的集合,该候选项的集合记作 C_k,依次下去直到 C_{k+1} 为空。在产生 C_k,k=1,2,…,k 时,利用剪枝策略压缩 C_k。利用任何非频繁的(k－1)－项集都不可能是频繁 k－项集这一 Apriori 性质,删去那些(k－1)－子集不在 L_{k-1} 中的 k－候选项目集。

使用 Apriori 算法进行关联规则挖掘,可以比较有效地产生关联规则,但也存在着以下两种缺陷[3]:

① 算法产生太多虚假(冗余)的规则。当数据仓库太大或支持度、信任度阈值太低时产生的规则太多,用户很难人为地对这些规则做出区分、判断,因而很难找到真正对用户有用的知识。

② 算法在效率上存在着问题。主要原因为数据库扫描的次数太多,寻找每一个 k－1 频繁项目集(k=

① 基金项目:教育部的世行贷款——21 世纪初高等教育教学改革项目(项目编号:1283B0843);
国家 863 高技术发展计划(项目编号:2002AA4Z3240)

1,2,…k) 都需要扫描数据库一次,共需要扫描 K 次。另外,当模式太长时产生的候选项目集也多得让人无法接受。因此当数据库或 K 太大时,算法的时耗太大或无法完成。故算法的可扩展性也不强,难于推广。这两种缺陷是数据挖掘的难点也是热点,并已成为约束系统性能的瓶颈。

目前已有很多文献提出了 Aprior 算法的变形,旨在提高算法效率。这些方法是通过减少扫描数据库次数或减少数据集等方式来提高算法的效率,有一些也取得了很好的效果。但都是单纯从 Apior 算法本身出发,而忽略了对待挖掘数据的分析与处理。

随着数据仓库中数据的持续增加,在数据挖掘过程中,进行一次数据挖掘的时间越来越长,规则越来越多,最终用户将面对着堆积如山的规则。许多的用户对总体数据含有的规则并不感兴趣,他们只关心某些细化区域的隐含规则。采用总体数据进行挖掘时,不仅挖掘时间相应的增长、有用的规则淹没于用户不感兴趣的规则海洋里,而且可能有的规则由于整体数据的"稀释"而无法挖掘出来。既降低了效率,又得不到有用的知识。因此,进行关联规则的挖掘需要根据用户的兴趣方向进行数据区域细化。本文即是预处理阶段利用聚类分析,分析待挖掘数据,先将数据进行聚类,将区域细化,选择聚类后的用户感兴趣的数据进行关联规则挖掘,即将两种挖掘方法有效的结合起来,在聚类的基础上进行挖掘。针对不同的用户,可以采用此算法来快速高效的得到其所需要的信息,提高了算法的效率。

2　基于聚类的关联规则改进算法

聚类分析是数据挖掘中重要的研究课题之一。聚类被广泛研究并应用于机器学习、统计分析、模式识别以及数据库数据挖掘与知识发现等不同的领域。所谓聚类[2,4]就是将物理或抽象对象的集合组成为由类似的对象组成的多个类的过程。由聚类所生成的类是一组数据对象的集合,聚类分析的原理是使属于同一类别的个体之间距离尽可能小,而不同类别的体体之间距离尽可能大。目前在文献中存在大量的聚类算法。算法的选择取决于数据的类型、聚类的目的和应用。如果聚类分析被用作描述或探查的工具,可以对同样的数据尝试多种算法,以发现数据可能揭示的结果。主要的聚类算法可以划分为如下几类: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。划分方法中的 K – MEANS 算法[5,6]是聚类分析中理论上最可行,应用最为广泛的算法之一。

通过聚类,人们能够识别密集和稀疏区域,发现全局的分布模式和数据属性之间有趣的相互关系。聚类分析作为数据挖掘中的一个模块,它既可以作为一个单独的工具以发现数据库中数据分布的一些深入的信息,并且概括出每一类的特点,或者把注意力放在某一个特定的类上以作进一步的分析;也可以作为数据挖掘算法中其他分析算法的一个预处理步骤,对数据仓库中的数据进行聚类,生成相互区分的类,其他算法再在特定类上进行处理,完成数据挖掘,产生对用户有用的知识。

基于聚类的关联规则挖掘算法的基本思想是将聚类分析作为关联规则算法的一个预处理步骤。也即是先对数据仓库中的数据按照一定的方法进行聚类,将数据按照用户感兴趣的方向进行数据区域细化,将数据集放在相应的类型中。用户根据其关心区域的选定数据类进行关联分析,使得在关联规则分析的过程中数据集的范围大大缩小,从而提高挖掘的效率。

基于园区网络的数据挖掘系统[7]使用 Apriori 算法的扩展,不仅要进行基于特定数据类的关联分析,还要进行多类数据的综合分析。进行各类数据的关联分析后,再进行多类数据的关联分析时,关联规则必然会发生变化。重新运行挖掘系统必然费时,同时也意味着以前基于单个类生成的频繁项目集和关联规则都被白白浪费了,显然不利于快速高效的发现规则。研究在原有频繁项目集的基础上,快速高效的生成新的频繁项目集,系统使用了以下算法,以两类数据的合并后进行关联分析为例:

(1) 比较第一类 a 的频繁项目集 A 和第二类 b 的频繁项目集 B,找出其中相同部分,将其中相同部分放入合并后的新数据类 c 的频繁项目集 C 中。

(2) 对频繁项目集 Ia,Ia ∈ A 但 Ia ∉ B,则扫描 b,获得在 b 中的支持度 supx(为了方便,将支持度改为支持 Ia 的事务数,即原来的支持度 * 总事务数,下同),supx 加上在 A 中的支持度 sup,若两者之和大于或等于最小支持度,则放入 C 中。

(3) 与(2)相类似,对频繁项目集 Ia,Ia ∈ B 但 Ia

∈ A,则扫描 a,重新计算其支持度,若大于或等于最小支持度,则放入 C 中。

算法基本框架描述如下:

```
/ * 对频繁项目集 la,la ∈ A 且 la ∈ B * /
For all la ∈ A do begin
  If la ∈ B then do begin
    C = { la} ;
    Delete la from A;
    Delete la from B;
  End;
End
/ * 对频繁项目集 la,la ∈ A 且 la ∈ B * /
For all la ∈ A do begin
  For all transaction t ∈ b do begin
    If la t then
      La. count + + ;
  End;
  C = C + la{ la. count ≥ minsup} ;
End;
/ * 对频繁项目集 la,la ∈ B 且 la ∈ A * /
For all la ∈ B do begin
 For all transaction t ∈ a do begin
  If la ∈ t then
   La. count + + ;
 End;
  C = C + la{ la. count ≥ minsup} ;
End;
```

则最后符合条件的频繁项目集就是 C。利用频繁项目集 C 生成所需要的关联规则,就得到了将两类数据(a 和 b)合并后得到的类 c 的关联规则。

3 系统的设计与实现

笔者在 Windows 2000 professional 系统下进行了该系统的设计,数据库平台选用 SQL Server 2000。开发工具为 Jbuilder9。系统是在基于园区网络的教务管理数据仓库平台上进行的,目的是得到课程之间的相关信息,起到指导学生选课的作用。

3.1 数据预处理

本阶段又可以进一步细分为两步:数据集成;数据选择和预分析。

(1) 集成(Integration)。在对教学系统数据的处理中,这一步主要是消除二义性,统一数据类型。基于园区网络的数据仓库系统研究设计了"学生"、"课程"这两个主题的数据集市,为数据挖掘系统提供集成的数据源,主要对学生成绩进行分析。

(2) 数据选择和预分析。采用 K – MEANS 算法将数据进行聚类,具体操作如下:首先构造一个数据矩阵,用 p 个变量来表示 n 个对象。在设计中,用课程名,课程代号,学分,课时,课程性质(必修或选修),开课时间,授课老师,授课地点来表现对象课程。然后随机选择 K 个代表点,其余目标根据到代表点的距离划分到 K 个代表点,其余目标根据到代表点的距离划分到 K 个类中。然后用每个类的中心代表这个类,对目标进行重新分割,这一过程迭代进行,直到收敛。在本实例中,选取 K 初始值为 10,随机选择 10 个代表点对"课程"为主题的数据集市进行分析,采用欧几里德距离公式:

$$d(i,j) = \sqrt{w_1(x_{i1} - x_{j1})^2 + w_2(x_{i2} - x_{j2})^2 + \cdots + w_p(x_{ip} - x_{jp})^2}$$

进行分析,直到没有对象的重新分配发生时,使不同学院的课程各成一类。再采用同样的方法使得每个院的各个专业课程各成一类,逐步缩小数据值。

3.2 数据挖掘阶段

不同的用户根据其需求或兴趣分别选择特定的数据类进行关联规则分析,输出相应的关联规则。在关联规则分析中用到学生的考试成绩,学生所属学院等学生基本信息以及在聚类分析过程中形成的各门课程所属的课程分类信息。根据学生的考试成绩分析出各门课程之间的内在联系,输出关联规则。本实例是对某校信息学院的学生成绩进行分析,首先是对各不同专业的学生成绩进行分析,得到专业内各课程之间的相关信息。由于各专业各学科存在交叉性,可以利用改进后的关联规则进行了综合分析。得到学院内课程之间的相关性分析。

3.3 结果描述

在本系统中采用表格形式学生成绩挖掘结果。设置支持度为 0.1,置信度为 0.8,得到了有关课程的关联规则。部分关联规则表示如下,各项分别为(no,front,rear,C),其含义分别是(产生关联规则的序号,关联规则前件,关联规则后件,可信度),其中关联规则项需要通过查询组合第二列和第三列得出来。部分结果如表 1 所示:

表 1 某学院课程关联规则表

no	Front	rear	C	S
1	离散数学	数据结构	0.76	0.04
2	数据结构	数据库基础	0.34	0.03
3	算法分析与设计	C 语言程序设计	0.83	0.05

通过分析以上结果可以得出加强《离散数学》的学习有助于对《数据结构》课程的学习,其他规则同样可按照这种方式分析。

3 4 评价

如果分析人员对分析结果不满意,可以调整支持度,可信度及兴趣度阈值,递归地执行上述三个过程,直到满意为止。

为了验证该改进算法的效率,同时也采用经典 Apriori 算法进行了关联规则的挖掘,得到了相同的挖掘结果,但耗费的时间相对较长,因此本改进算法是合理的,不仅可以使得用户只分析其感兴趣的数据,而且在进行各类之间的分析时,确实可以提高效率。

4 结束语

本文提出了一种新的提高数据挖掘效率的新方法——采用基于聚类的关联规则算法实现算法的挖掘,在预处理阶段采用聚类以缩小数据规模。并将该技术应用到教学系统中,解决了以前在教学系统中存在的问题,较客观实时地反映教学系统中存在的问题,为决策提供重要依据。这一研究也对实际的教学管理提出了很好的建议,给多年来的计算机教学管理工作又添上了新的内容。

参考文献

1 Agrawal R, Imielinski T, Swami A. Mining Association Rules between Sets of Ltems in Large Database[M]. In SIGMOD" 93, Washington, DC, May 1993. 207 – 216.

2 范明、孟小峰等译,数据挖掘概念与技术[M],机械工业出版社,2003.3.150 – 221.

3 贾彩燕、倪现君,关联规则挖掘研究述评[J],计算机科学,2003,30(4):145 – 148。

4 朱明,数据挖掘[M],中国科学大学出版社,2002,5:129 – 140。

5 马光志、龙硕柱,基于聚类和分类的自学习系统模型[J],计算机工程与应用,2003,39(10):83 – 84。

6 罗可、蔡碧野、吴一帆、谢中科、张丽,数据挖掘中聚类的研究[J],计算机工程与应用,2003,(39)20:182 – 184。