数据挖掘技术在教学评价中的应用研究

罗美淑[1]，刘世勇[2]，夏春艳[1]
（1.牡丹江师范学院 工学院；2.黑龙江幼儿师范高等专科学校，黑龙江 牡丹江 157011）

摘 要：数据挖掘技术是在海量数据中提取有用信息的有效手段，而教学评价是对教学工作质量所做的测量、分析和评定，是教学过程中的重要环节。将数据挖掘技术应用到教学评价数据分析过程中，验证了基于该技术的属性约简算法的正确性和有效性，从多角度对教学评价数据进行更深层次的分析和处理，从而挖掘出更多、更有价值的数据和信息，提供了更多的方法和措施以改进和提高教学的质量。

关键词：数据挖掘技术；属性约简算法；教学评价；应用
中图分类号：G40-058.1 **文献标志码**：A **文章编号**：1002-0845(2013)02-0081-02

一、数据挖掘技术的基本概念

定义1：信息系统S=(U,A,f,R)，通常略写为S=(U,A)。其中，U是对象的非空有限集；A是由条件属性集C和决策属性集D构成的非空有限属性集，即A=C∪D，C∩D=Φ；R为A的值域；f:A→V为从属性到值域的映射。

定义2：对一个知识系统S=(U,A,f,R)，P⊆R，可以用下面的公式表示不可区分关系：

$IND(P)=\{(x,y)\in U\times U\mid \forall a\in P, f(x,a)=f(y,a)\}$

IND(P)在U上导出的划分U/IND(P)={$[x]_p \mid x\in U$}，记为U/P。$[x]_p=\{y\in U\mid \forall a\in P, f(x,a)=f(y,a)\}$称为包含x的P等价类。

定义3：给定决策表S=(U,A,f,R)，其中A=C∪D，C为条件属性集，D为决策属性集，则区分矩阵$M_D=\{m_{ij}\}$定义为：

$$m_{ij}=\begin{cases}a\in C; f(x_i,a)\neq f(x_j,a)\}, & 当f(x_i,D)\neq f(x_j,D)时,\\ \Phi, & 其他\end{cases}$$

定义4：可形式化地定义知识的依赖性为：令S=(U,R)为一个知识库，P⊆R，Q⊆R，则知识Q依赖于知识P(记作P⇒Q当且仅当IND(P)⊆IND(Q)。

当$k=\gamma_p(Q)=|POS_P(Q)|/|U|$，则称知识Q是k度依赖于知识P，记作$P\Rightarrow_k Q$。Q和P间的依赖度为系数$\gamma_p(Q)$。

定义5：设C为条件属性集，D为决策属性集，已知属性R，则一个属性a∈C-R关于D的重要度可定义为：

$SGF(a,R,D)=\gamma_{R\cup\{a\}}(D)-\gamma_R(D)$

定义6：设R为一族等价关系，p∈R，若IND(R)=IND(R-{p})，则称p是R中不必要的；否则称p是R中必要的。若对每个p∈R都是R中必要的，则称R是独立的；否则，称R是依赖的。

设P⊆R，若P是独立的，且IND(P)=IND(R)，则称P是R的一个约简。

定义7：CORE(R)=∩RED(R)，其中RED(R)表示P的所有约简。

需要指出的是：一般属性约简不是唯一的，而属性核则是唯一的。其中，最小约简是指包含关系最小的约简。

二、源于数据挖掘技术的属性约简算法

在数据挖掘的研究方法中，大部分研究者普遍重视对不确定性问题的处理，并在人工智能的研究领域中，提出了许多关于不确定性问题的处理方法，其中模糊集、概率论、粗糙集理论、证据理论等被普遍应用。

在数据挖掘领域中，粗糙集理论有普遍的应用前景和实用价值，它是进行数据挖掘的主要方法之一。在数据挖掘领域中应用粗糙集理论，可以提高分析大型数据库中不完整数据的能力。属性约简的意义：在知识库的分类能力保持不变的情况下，对知识进一步化简，删除知识库中不相关或不重要的冗余的知识信息，推导出问题的决策规则，生成简化的知识规则库，使人们通过约简后的知识库能够准确地把握问题的决策方向，并快速做出对问题的处理决策。

发现属性集中的最小约简是属性约简的最终目的，而目前大多数属性约简算法没有考虑到属性间的影响度，仅考虑了条件属性对决策属性的依赖度。在可辨识矩阵中，当一个属性反复出现很多次时，我们就认为该属性是非常重要的。但是，在可辨识矩阵的属性项中，如果同时反复出现的约简集中的两个属性项彼此有很高的影响度，则说明这两个属性项对决策属性有相似的分类能力，我们应先计算出某属性对约简集中属性的影响度，然后判断该属性是否加入约简集，若影响度很高，则确定该属性为冗余属性，不必加入约简集。

属性加权频率应遵循的思想是：属性在可辨识矩阵里出现的次数越多，或可辨识矩阵中的属性项的长度越短，则属性的重要性越大。

属性重要性函数可定义如下：

$$SIG(a)=\sum_{i=1}^{n} count(a_i)/i$$

其中，a为可辨识矩阵属性项中的属性，i是含有a的属性项的长度，n是可辨识矩阵中含有a的所有属性项中最长项的属性个数，count(a_i)是指在可辨识矩阵中含有a的属性项的长度i出现的次数。

定义属性a相对于约简集中属性的影响度函数为：

$$IMP(a)=\frac{count(Red\cap a)}{count(a)}$$

其中，count(a)同上述说明，count(Red∩a)是属性a与约简集中属性同时出现的项的个数。

算法的实现过程可描述如下：

收稿日期：2012-11-01
基金项目：牡丹江师范学院教改项目(12-XJ14023)
作者简介：罗美淑(1981-)，女(朝鲜族)，哈尔滨人，讲师，从事数据库技术、程序设计与数据挖掘技术研究；刘世勇(1978-)，男，哈尔滨人，讲师，从事网络技术、数据库技术与软件开发研究；夏春艳(1980-)，女，黑龙江桦川人，讲师，硕士，从事数据挖掘与信息处理研究。

输入：决策系统IS=(U,C∪D)，其中，U、C和D依次是对象集、条件属性集和决策属性集；

输出：C的约简集Red；

步骤：1 CORE(A)=[Φ];

2 M=DisMat(S); //生成M，即可辨识矩阵

{for(i=0;i<n;i++)

for(j=i+1;j<n;j++)

for(k=1;k<=|c|;k++)

if((Ck(Xi)≠Ck(Xj))&&D(Xi)≠D(Xj))

$m_{ij}=m_{ij}\cup\{c_k\}$;}

3 Core=GeneCore(M, count);//根据M计算Core，M为可辨识矩阵，Core为属性核

{for(i=0;i<n;i++)

for(j=i+1;j<n;j++)

if(|m_{ij}|==1)

$C_0=C_0\cup m_{ij}$;}

4 通过公式$SIG(a)=\sum_{i=1}^{n} count(a_i)/i$求出各属性的重要度；

5 while (γ(Red, D)≠γ(Red, C))

{selecr(a)=max(SIG(a_i));

$p=IMP(a)=\frac{count(Red\cap a)}{count(a)}$

if(p<0.5)

Red=Red∪{a}

else

C=C-{a};}

6 程序运行后得到属性约简集Red，并输出Red。

三、数据挖掘技术在教学评价中的应用

传统的教学评价是对调查得出的数据进行量化分析，然后得出结论，并由此做出判断，在形式与内容上显得比较单一，往往局限于学生打分、教师评、学生互评等，然而这样做并不能发现数据中深层次的内容。因此，从原始数据中很难找出有关教学质量的一些规律，对提高教师教学的质量和水平起不到有效的帮助作用。而数据挖掘作为一种深层次的数据分析方法和有效地解决这一问题的新技术，它可以对教学的质量和水平与各因素之间隐藏的内在联系进行全面透彻的分析。我们可利用数据挖掘技术分析已有的教学评价数据，并对评价数据进行合理的处理，从中发现类似“可能对教师教学水平产生影响的因素”等这样的问题，以及在什么条件下教师的教学质量和水平是“高的”或“不高的”，帮助教师改进教学的方法，进而提高教学的质量和水平。

每学期结束，笔者所在学校都会对任课教师的教学进行评价，本文选用了其中某一个学期一门计算机课程的评价数据作为挖掘对象。先对数据进行预处理，删除异常数据后，提取各项评价数据（见表1）。

表1 教师评价数据信息表

教师编号	师德风范（15分）	教书育人（15分）	课堂教学（50分）	教学效果（20分）	总分
1	14.61	13.85	44.10	18.11	90.67
2	14.81	14.87	47.52	18.96	96.16
3	14.21	13.59	43.45	16.24	87.49
4	13.80	13.74	35.64	16.13	79.31
5	14.55	14.38	42.81	16.42	88.16
6	13.69	13.77	35.55	16.20	79.21

任课教师的人事信息汇总情况见表2。

表2 教师人事信息表

教师姓名	教师编号	性别	年龄	学历	职称
aaa	1	女	29	博士	讲师
bbb	2	男	37	博士	副教授
ccc	3	女	29	本科	讲师
ddd	4	女	26	本科	助教
eee	5	男	30	硕士	讲师
fff	6	男	27	本科	讲师

对表1、表2的数据进行处理，得到表3。其中，C为条件属性集，C={学历、职称、师德风范与教书育人、课堂教学、教学效果}，决策属性D为评价结果，令C_1=学历，C_2=职称，C_3=师德风范与教学育人，C_4=课堂教学，C_5=教学效果。

表3 信息决策表

U	C					D
	C_1	C_2	C_3	C_4	C_5	D
U_1	3	2	3	3	2	2
U_2	3	3	3	3	3	3
U_3	1	2	3	2	1	1
U_4	1	1	2	1	1	1
U_5	2	2	3	2	1	1
U_6	1	2	2	1	1	1

表4 传统区分矩阵

U	1	2	3	4	5	6
1	Φ					
2	C_5	Φ				
3	$C_1C_4C_5$	$C_1C_4C_5$	Φ			
4	$C_1C_3C_4C_5$	$C_1C_3C_4C_5$	C_3C_4	Φ		
5	$C_1C_4C_5$	$C_1C_4C_5$	C_1C_4	C_1C_5	Φ	
6	$C_1C_4C_5$	$C_1C_4C_5$	Φ	C_5C_4	C_1C_4	Φ

表5 简化区分矩阵

U	1	2	3	4	5	6
1	Φ					
2	C_5	Φ				
3	Φ	Φ	Φ			
4	Φ	Φ	C_5C_4	Φ		
5	Φ	Φ	C_1C_4	C_1C_5	Φ	
6	Φ	Φ	Φ	C_5C_4	C_1C_4	Φ

从以上简化区分矩阵表得出，{C_5}为属性的核集，即初始Red={C_5}，而且从简化区分矩阵表看其余属性的频率也很清晰。依据属性的频率及其依赖度，得到{C_1C_5}、{C_3C_5}、{C_4C_5}为最终的约简集。在评价规则中，属性的重要性依次为：1）课堂教学；2）学历、教书育人与师德风范、教学效果；3）职称。从上述结果可以看出本算法是可行的、有效的。同时，通过对表4和表5的比较，可以看出与传统区分矩阵相比，简化区分矩阵在空间和时间上的复杂度都有所降低。另外，本算法随着数据量的不断增大，还能够使空间及时间的复杂度大幅度降低。

参考文献：

[1]PAWLAK Z. Rough sets[J]. Information and Computer Science.198211(5):341-356.

[2]张文修，吴伟志，梁吉业. 粗糙集理论与方法[M]. 北京：科学出版社，2001：123-131.

[3]胡可云. 基于概念格和粗糙集的数据挖掘方法研究[D]. 北京：清华大学，2001：36-52.

[4]夏春艳，李树平，宋志超. 基于粗糙集理论属性约简的改进算法[J]. 微计算机信息，2010，12(36)：282-283.

[5]付海艳，符谋松，张诚一. 粗糙集理论在高校教学质量评价分析中的应用[J]. 计算机工程与应用，2007，43(36)：214-216.

〔责任编辑：张 华〕