模糊聚类最大树算法在教学质量评估中的应用*

卓景文[1,2],赵 鹏[1,2],李学俊[1,2],赵志伟[1,2]
(1.安徽大学 计算智能与信号处理教育部重点实验室,安徽 合肥 230039;
2.安徽大学 计算机科学与技术学院,安徽 合肥 230039)

摘 要:应用模糊聚类最大树算法对教学质量评估指标进行聚类以确定关键评估指标集,使用模糊相似关系挖掘出大量数据中教学质量评估指标与评估等级之间的规则,并以本校数据实例为对象建立教学质量评估模糊数据挖掘验证了该方法的有效性。

关键词:模糊数据挖掘;模糊聚类;教学质量评估;模糊相似矩阵
中图分类号:TP391 **文献标识码**:A **文章编号**:1674-7720(2012)05-0060-03

Research on the application of maximal tree method based on fuzzy clustering in teaching quality evaluation

Zhuo Jingwen[1,2], Zhao Peng[1,2], Li Xuejun[1,2], Zhao Zhiwei[1,2]
(1.Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, Anhui University, Hefei 230039, China;
2.School of Computer Science and Technology, Anhui University, Hefei 230039, China)

Abstract: This paper applies fuzzy clustering maximal tree method to cluster the evaluation indexes in order to define the key evaluation index sets. Using fuzzy similarity relation to mine the rules between the evaluation index and the evaluation level of teaching quality from mass of data. And we also use the data of our school as an example to build the fuzzy data mining of teaching quality evaluation to prove the effectiveness of the proposed method.

Key words: fuzzy data mining; fuzzy clustering; teaching quality evaluation; fuzzy similar matrix

教学管理是为了实现教学目标,按照教学规律和特点对教学过程进行的全面管理。通过不断改善影响学校教学质量的内部因素和外部因素,建立科学的评价体系来提高教学质量,达到最佳教学效果。

数据挖掘是从大量数据中提取或"挖掘"知识(即数据中的知识发现),并以这些知识为基础,自动做出决策和预测。数据挖掘已经应用于众多领域,如金融数据分析、零售业、信息检索等。随着信息技术的发展和高等教育体制改革的不断深入,高校实现了教育信息化,大大提高了工作效率。将数据挖掘技术应用于高校教务管理中,可以挖掘出重要的对决策或者预测有用的信息和知识,利用分析结果辅助教学,帮助教学管理者做出科学的决策。然而数据库或者数据仓库的容量越大,系统复杂性越高,相应的精确化能力就越低,也就是说模糊性越强,因而仅仅依靠复杂算法和推理并不能完全发现隐藏知识,因此,考虑将模糊数学、模糊逻辑和数据挖掘结合起来的模糊数据挖掘技术引入到教学质量评估中。

1 相关基本知识概论

模糊集是用来表达模糊性概念的集合[1]。

定义1:设 X 为论域,$x \in X$,设 $\widetilde{A}$ 是论域 X 到[0,1]的一个映射,即 $\widetilde{A}(X): X \to [0,1]$, $x \to \widetilde{A}(x)$,称 $\widetilde{A}$ 是 X 上的模糊集,而函数 $\widetilde{A}(X)$ 称为模糊集 $\widetilde{A}$ 的隶属函数,$\widetilde{A}(x)$ 称为 x 对模糊集 $\widetilde{A}$ 的隶属度。

定义2:公式 $A \to B$ 的逻辑含义称为决策规则,A 称为规则的前件,B 称为规则的后件,它们表达一种因果关系。其中公式 A 中所包含的原子公式只有决策表中的条件属性,B 中所包含的原子公式只有决策表中的决策属性。

* 基金项目:安徽省教育厅重点项目(KJ2009A001Z);
安徽大学211工程三期教学质量工程项目(39030051)

2 基于模糊聚类最大树算法的模糊数据挖掘

聚类是一种无监督的学习过程，把具有类似属性的个体聚成一类。从聚类的角度出发，由于客观世界中大量存在着界限并不分明的聚类问题，模糊聚类应运而生。模糊聚类是基于模糊等价关系分类的，模糊等价关系往往由模糊相似矩阵产生。

定义3：假设有 N 个要分类的样本，记为集合 $X=(x_1,x_2,x_3,\cdots,x_n)$，每个样本有 m 个量化指标，记为 $Y=(y_1,y_2,y_3,\cdots,y_m)$，则可以列出样本-指标原始数据矩阵 M，其中 x_{ij} 表示第 i 个对象相应于第 j 个指标的数值[1]。

在教学质量评估中，评估对象的某些评估因子往往会带有一定程度的模糊性。所以用模糊理论来进行聚类分析，然后再进行模糊数据挖掘，依据挖掘结果进行预测，得到有利于领导决策的有用规则。基于模糊聚类最大树算法的模糊数据挖掘算法如下：

(1)由定义3确定聚类分析的对象，得到原始矩阵：

$$M=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

(2)由于采集到的数据往往不在[0,1]内，根据模糊矩阵的要求，通过下面两步将数据压缩到区间[0,1]上：

①标准差变换：

$$x'_{ik}=\frac{x_{ik}-\overline{x_k}}{S_k}\quad (i=1,2,\cdots,n;k=1,2,\cdots,m) \tag{1}$$

其中 $\overline{x_k}=\frac{1}{n}\sum_{i=1}^{n}x_{ik}$ 是平均值，$S_k=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ik}-\overline{x_k})^2}$ 是标准差。此时得到的标准化数据 x'_{ik} 也不一定均在[0,1]内，还必须进行下面的变换。

②极差变换：

$$x''_{ik}=\frac{x'_{ik}-\min\limits_{1\leqslant i\leqslant n}x'_{ik}}{\max\limits_{1\leqslant i\leqslant n}x'_{ik}-\min\limits_{1\leqslant i\leqslant n}x'_{ik}}\quad (k=1,2,\cdots,m) \tag{2}$$

(3)建立模糊相似矩阵。模糊相似矩阵用来描述样本之间的相关程度，即标出衡量被分类对象间相似程度的统计量 $r_{ij}(i,j=1,2,\cdots,n)$。设论域 $U=\{u_1,u_2,u_3,\cdots,u_n\}$，其中每个元素为一个样本，建立 U 上的模糊相似矩阵：

$$R=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix}\quad 0\leqslant r_{ij}\leqslant 1,i,j=1,2,\cdots,n \tag{3}$$

考虑到正确性原则、不变性原则和可区分性原则，使用绝对值倒数法[2]计算 r_{ij}，即：

$$r_{ij}=\begin{cases} 1 & i=j \\ \dfrac{c}{\sum\limits_{k=1}^{m}|x_{ik}-x_{jk}|} & i\neq j \end{cases} \tag{4}$$

其中 c 为相似系数，反映样本之间相对于某些属性的相似程度。

(4)聚类分析。常用的模糊聚类分析方法有三种：传递闭包法、最大树法和编网法。考虑到计算量，本文采用最大树算法进行模糊聚类。构造最大树的算法如下：

①将模糊相似关系矩阵中的 r_{ij} 由大到小排序：$\beta_1>\beta_2>\cdots>\beta_h$，其中 $\beta_k(k=1,2,\cdots,h)$ 为某 r_{ij}；

②以被分类的对象为顶点，依据模糊相似矩阵将关联程度为 β_1 的顶点连接，并在相应的线段上标明 β_1，若在连接某两个顶点时出现回路，则不画此线；

③依次对 $\beta_2,\beta_3,\cdots,\beta_k(k\leqslant h)$ 按照上步重复，直到所有顶点构成一个无向连通赋权图(不一定到 h 步)，即得到最大树 $G=(X,r_{ij})$。

(5)得到聚类结果。首先确定截割水平 λ，然后根据 λ 值对最大树进行切割[3]。分别比较 λ 与最大树各边的权值之间的大小。当 $\lambda>r_{ij}$ 时，将 r_{ij} 对应的边截断，这样剩余的并且还相互连通的顶点就构成一类。

3 模糊数据挖掘在教学质量评估中的应用

3.1 建立教学质量评价指标体系

课堂教学质量测评工作是教学质量评估体系的重要组成，是加强教学管理、提高教学质量的重要手段。为使课堂教学质量、学生测评工作科学化和规范化，教务处制定了完善的课堂教学评价指标体系，其中第 n 条是整体评价。如表1所示。

3.2 教学质量评价中的模糊数据挖掘

通过科学评估教师的课堂教学质量，为学校教学管理提供决策的信息与依据，促使形成一套较为完整的教学评价机制。每门课程的学生测评成绩(统计时自动剔除5%的最高分和最低分)由教务管理系统自动生成。教师的学期测评成绩为其该学期所承担的各门课程学生测评成绩的平均值。年度测评成绩为两学期的平均值。如教师只承担一个学期的课程，则以该学期测评成绩为其该年成绩。教师年度学生测评成绩以70%计入教师当年教学考核总评成绩。表2所示为我校10名教师的学生测评成绩。

表1 课堂教学质量评价指标表

评价号	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n
评价指标	在教学过程中认真负责，遵纪守时，注重为人师表	教学准备工作充分，讲课熟练自如	教学中做到条理清晰，重点突出，语言生动，表述准确	教学中板书安排(或者使用多媒体等现代化教学手段)合理	辅导答疑及作业批改认真负责	讲授内容充实、新颖	教学中注重启迪思维，激发兴趣，联系实际，培养能力	对本课程教学质量的整体评价

表 2 课堂教学质量评分数据表

	n_1	n_2	n_3	n_4	n_5	n_6	n_7	n
1	89	82	91	83	81	86	90	优秀
2	85	83	89	84	80	72	74	良好
3	77	71	72	69	76	80	83	中等
4	74	69	78	79	73	88	69	中等
5	83	79	86	81	78	84	88	良好
6	61	76	64	62	66	63	71	及格
7	93	85	96	88	82	82	92	优秀
8	86	87	89	78	71	81	76	良好
9	94	93	96	89	89	83	95	优秀
10	87	76	84	78	74	77	83	良好

对表 2 中的数据应用基于模糊聚类的最大树算法找出影响教学质量的主要因素。

(1)由表 2 得到原始矩阵 $\boldsymbol{M}_{7\times10}=\begin{bmatrix}89 & 85 & \cdots & 87\\ 82 & 83 & \cdots & 76\\ \vdots & \vdots & \vdots & \vdots\\ 90 & 74 & \cdots & 83\end{bmatrix}$

(2) 由于得到的原始矩阵不是模糊矩阵，先由式(1)进行标准差变换，再由式(2)进行极差变换后的矩阵即为模糊矩阵：

$$\boldsymbol{M}'_{7\times10}=\begin{bmatrix}0.80 & 0.76 & \cdots & 1\\ 0.10 & 0.65 & \cdots & 0.15\\ \vdots & \vdots & \vdots & \vdots\\ 0.90 & 0.88 & \cdots & 0.69\end{bmatrix}$$

(3)由式(3)建立模糊相似矩阵 $\boldsymbol{R}_{7\times7}$：

$$\boldsymbol{R}_{7\times7}=\begin{bmatrix}1 & 0.18 & 0.47 & 0.18 & 0.13 & 0.17 & 0.71\\ 0.18 & 1 & 0.15 & 0.33 & 0.32 & 0.18 & 0.18\\ 0.47 & 0.15 & 1 & 0.15 & 0.12 & 0.14 & 0.97\\ 0.17 & 0.33 & 0.15 & 1 & 0.33 & 0.25 & 0.17\\ 0.13 & 0.32 & 0.12 & 0.33 & 1 & 0.23 & 0.13\\ 0.17 & 0.18 & 0.14 & 0.25 & 0.23 & 1 & 0.17\\ 0.71 & 0.18 & 0.97 & 0.17 & 0.13 & 0.17 & 1\end{bmatrix}$$

(4) 由上一步得到的模糊相似矩阵 $\boldsymbol{R}_{7\times7}$，按照算法步骤(4)最大树的构造算法建立的最大树如图 1 所示。

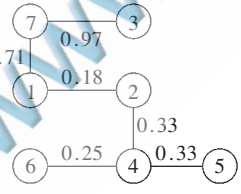

图 1 最大树

(5)从最大树中可以得出：当 $\lambda=1$ 时，分为 7 类：$\{n_1\},\{n_2\},\{n_3\},\{n_4\},\{n_5\},\{n_6\},\{n_7\}$；当 $\lambda\geqslant0.71$ 时，分为 5 类：$\{n_1,n_3,n_7\},\{n_2\},\{n_4\},\{n_5\},\{n_6\}$；当 $\lambda\geqslant0.33$ 时，分为 3 类：$\{n_1,n_3,n_7\},\{n_2,n_4,n_5\},\{n_6\}$；当 $\lambda\geqslant0.25$ 时，分为 2 类：$\{n_1,n_3,n_7\},\{n_2,n_4,n_5,n_6\}$；当 $\lambda\geqslant0.18$ 时，分为 1 类：$\{n_1,n_2,n_3,n_4,n_5,n_6,n_7\}$。用 F-统计量确定最佳划分阈值为 $\lambda\geqslant0.71$，评价指标被分为 $\{n_1,n_3,n_7\},\{n_2\},\{n_4\},\{n_5\},\{n_6\}$，对课堂教学质量评估数据应用最大树算法聚类得到 n_1,n_3,n_7，即为影响课堂教学质量的关键评价指标集。

用基于模糊相似关系的规则获取方法可以进一步由关键评价指标集得到分类规则[4]。课程整体评价指标 n 的取值分为 4 类：优秀、良好、中等、及格，分别用 m_1、m_2、m_3、m_4 表示，i_n 表示编号为 n 的教师，$m_1=\{i_1,i_7,i_9\}$，$m_2=\{i_2,i_5,i_8,i_{10}\}$，$m_3=\{i_3,i_4\}$，$m_4=\{i_6\}$。

对于表 1 中的评价指标属性评价结果数值划分为 5 个区间，分别为 y_1：90~100 分；y_2：80~89 分；y_3：70~79 分；y_4：60~69 分；y_5：小于 60 分。得到关键评价指标集的评价结果划分为 5 个区间以后的数据如表 3 所示。

表 3 关键指标集的评价等级划分表

	1	2	3	4	5	6	7	8	9	10
n_1	y_2	y_2	y_3	y_3	y_2	y_4	y_1	y_2	y_1	y_2
n_3	y_1	y_2	y_3	y_3	y_2	y_4	y_1	y_2	y_1	y_2
n_7	y_1	y_3	y_2	y_4	y_2	y_5	y_1	y_3	y_1	y_2
n	优秀	良好	中等	中等	良好	及格	优秀	良好	优秀	良好

基于关键评价指标集可将表 3 中的数据划分为 7 类：$k_1=\{i_7,i_9\}$，$k_2=\{i_1\}$，$k_3=\{i_5,i_{10}\}$，$k_4=\{i_2,i_8\}$，$k_5=\{i_3\}$，$k_6=\{i_4\}$，$k_7=\{i_6\}$。

将 $k_i(i=1,2,\cdots,7)$ 作为条件，$m_i(i=1,2,3,4)$ 作为结论，归纳总结可以得到如下规则：

规则 1：$(n_1=y_1)\wedge(n_3=y_1)\wedge(n_7=y_1)\Rightarrow m_1$

规则 2：$(n_1=y_2)\wedge(n_3=y_2)\wedge(n_7=y_2\vee y_3)\Rightarrow m_2$

规则 3：$(n_1=y_3)\wedge(n_3=y_3)\wedge(n_7=y_4\vee y_2)\Rightarrow m_3$

以上规则分析，当在教学过程中认真负责，遵纪守时，注重为人师表；教学中做到条理清晰，重点突出，语言生动，表述准确；教学中注意启迪思维，激发兴趣，联系实际，培养能力三条都≥90 分时，教学质量整体评价一定为优秀；当这三个指标都为[80,89)分或者后一指标为[70,79)分时，教学质量整体评价一定为良好；当前两个指标为[70,79)分且后一指标为[80,89)或[70,79)分时，教学质量整体评价一定为中等。基于以上分析可以看出，上述三条标准为影响课堂教学质量的关键因素。总之教师在上课过程中要注意做到端正教学态度，授课中要有条理、重点突出、表述准确，另外教学过程不能忽略学生这个主体，要激发学生兴趣，培养其独立思考和解决问题的能力。

本文使用基于模糊聚类最大树算法的模糊数据挖掘发现教学质量评估数据库中教师课堂教学质量评估等级同评估指标之间的规则知识，依据该规则知识对挖掘结果进行有效的评价，并且在分析、预测方面有着很大的优势，从而帮助决策者做出决策。当然对于该教学质量数据挖掘来说，这只是一部分工作，如何进一步优化该系统是下一步研究的主要工作。

参考文献

[1] 刘琦，林怀忠，陈纯.模糊聚类的最大树算法在 Web 页面分类中的应用[J].计算机应用研究，2004，21(11)：286-287.

[2] 王新洲，舒海翅.模糊相似矩阵的构造[J].吉首大学学报

(自然科学版),2003,24(3):37-41.

[3] Zhan Liqiang,Liu Daxin.Fuzzy clustering method for web user based on pages classification[J].Wuhan University Journal of Natural Sciences,2004,9(5):553-556.

[4] 冯源.基于模糊相似矩阵与粗糙集的规则获取[J].太原师范学院学报(自然科学版),2008,7(1):26-30.

(收稿日期:2011-11-18)

作者简介:

卓景文,男,1985年生,硕士研究生,主要研究方向:数据挖掘。

赵鹏,女,1976年生,博士,副教授,硕士生导师,主要研究方向:智能信息处理。

李学俊,男,1976年生,博士,副教授,硕士生导师,主要研究方向:数据库,数据挖掘。