FO-CA:一种基于距离差异度组合权重的多属性数据分类方法

龚 安[1] 高海康[1] 徐加放[2] 马兴敏[1]
(中国石油大学(华东)计算机与通信工程学院 青岛 266580)[1]
(中国石油大学(华东)石油工程学院 青岛 266580)[2]

摘 要 为了解决多属性数据分类问题,提出了一种基于模糊优选模型与聚类分析的分类方法(FO-CA)。首先由模糊优选模型得到有序综合指标数据集,其中在权重阶段提出了距离差异度并以此为依据构建了一种组合主客观权重的赋权方法;然后采用聚类分析将有序综合指标数据集聚类为几个簇进而分类;最后选取UCI中的Iris、Wine和Ruspini 3个数据集进行仿真实验。实验结果表明,该分类方法相比模糊优选方法及K-Means算法能获得更好的分类结果,对决策者有一定的参考价值。

关键词 模糊优选,聚类分析,距离差异度,组合权重,分类
中图法分类号 TP391 **文献标识码** A

FO-CA:A Multiple Attribute Data Classification Method Based on Distance Difference Degree Combination Weighting

GONG An[1] GAO Hai-kang[1] XU Jia-fang[2] MA Xing-min[1]
(School of Computer & Communication Engineering, China University of Petroleum, Qingdao 266580, China)[1]
(School of Petroleum Engineering, China University of Petroleum, Qingdao 266580, China)[2]

Abstract In order to solve the problem of multi-attribute data classification, we proposed a classification method based on fuzzy optimization and clustering analysis (FO-CA). First, we used fuzzy optimization model to get one dimensional composite indicator data set. Meanwhile, according to the distance difference degree, we established a combination weighting method to integrate subjective weights and objective weights in weighting stage. Second, we used hierarchical cluster analysis method to divide the composite indicator data set into several clusters, and then classified the clusters. Finally, we selected Iris, Wine and Ruspini datasets from UCI Machine Learning Repository for simulation experiments. The experiment results show that the proposed method achieves better results than fuzzy optimization method and K-Means algorithm, and provides an effective approach for data classification.

Keywords Fuzzy optimization, Clustering analysis, Distance difference degree, Combination weight, Classification

1 引言

多属性数据分类方法利用一套指标体系,以全面、综合的观点来研究方案的优劣性,避免认识上的片面性和单一性,提高分类准确度。近几年多属性数据分类问题研究取得长足进展,同时出现了诸多方法模型,但是目前的分类方法尚存在一些亟需改进的问题。陈守煜、赵英琪[1]在模糊综合评判的基础上,提出模糊优选理论并建立相应的数学模型。模糊优选模型具有明确的物理含义、缜密的数学思路等优点,在方案优选与分类评价中得到广泛应用,但在分类时需要专家对综合指标进行人为划分类范围,不能很好地体现类内数据对象的特性。俞文彬[2]等人提出了一种基于数据属性值的数据挖掘方法,根据ID3分类技术通过数据采集、整理及形成规则并化简规则,从众多指标中合理地选出影响分类的属性,其不足是没有充分考虑各参数的权重。A. Sharma、S. Srinivasan[3]提出了一种基于数据挖掘手段的多属性数据分类方法,并从聚类分析、朴素贝叶斯、多元线性回归分析多角度对储层多属性数据进行了分类与预测。

多属性数据分类方法多为定性或半定量的专家评价方法,其存在以下不足:①主观经验性较强;②很难对众多的参数做到比较科学的综合考虑,往往是顾此失彼;③权重系数不易确定,仅仅单一采用主观权重或者客观权重的做法是比较片面的。为了克服以上问题,本文提出了一种基于模糊优选模型与聚类分析的多属性数据分类方法FO-CA,其中针对权

到稿日期:2013-06-27 返修日期:2013-10-21 本文受2012山东省自然科学基金:含酸性气体甲烷气水合物生成机理及防治技术研究(ZR2012EEM020)资助。

龚 安(1971—),男,博士生,副教授,硕士生导师,主要研究方向为数据挖掘与知识发现,E-mail:gongan0328@sohu.com;高海康(1987—),男,硕士,主要研究方向为数据挖掘;徐加放(1973—),男,博士,副教授,硕士生导师,主要研究方向为钻井液完井液;马兴敏(1989—),男,硕士,主要研究方向为数据挖掘。

重问题提出了基于距离差异度的组合权重方法。

2 模糊优选理论数学模型

模糊优选理论[1,4]是以最小二乘法为基础建立的优选判别准则，根据优选判别准则分别量化数据集中的每个点(量化值表示数据点的优劣程度)，进而将量化值作为分类的依据。基于模糊优选理论的分类方法的主要步骤为：

步骤 1　设有 n 个数据点组成的数据集 A，每个数据点包含 m 个指标，则用矩阵表示为：

$$A_{n\times m}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}=(a_{ij}) \tag{1}$$

根据指标的性质，一般可以分为两类，一类是越大越优的指标，另一类是越小越优的指标。通过式(2)和式(3)分别对越大越优的指标和越小越优的指标进行数据预处理得到矩阵 R。

$$r_{ij}=\frac{a_{ij}}{\operatorname{Max} a_j},i=1,2,\cdots,n;j=1,2,\cdots,m \tag{2}$$

$$r_{ij}=1+\frac{\operatorname{Min} a_{ij}-a_{ij}}{\operatorname{Max} a_j},i=1,2,\cdots,n;j=1,2,\cdots,m \tag{3}$$

其中，$r_{ij}(i=1,2,\cdots,n;j=1,2,\cdots,m)$ 为数据集中第 i 个数据点第 j 个指标的优属度。

步骤 2　根据取大法确定越大越优指标的最优值为最优点 G；根据取小法确定越小越劣指标的最劣值为最劣点 B，即

$$\begin{aligned}\vec{G}&=(g_1,g_2,\cdots,g_m)\\&=(r_{11}\vee r_{21}\vee\cdots\vee r_{n1},r_{12}\vee r_{22}\vee\cdots\vee r_{n2},\cdots,r_{1m}\vee r_{2m}\\&\quad\vee\cdots\vee r_{nm})\end{aligned} \tag{4}$$

$$\begin{aligned}\vec{B}&=(b_1,b_2,\cdots,b_m)\\&=(r_{11}\wedge r_{21}\wedge\cdots\wedge r_{n1},r_{12}\wedge r_{22}\wedge\cdots\wedge r_{n2},\cdots,r_{1m}\wedge r_{2m}\\&\quad\wedge\cdots\wedge r_{nm})\end{aligned} \tag{5}$$

步骤 3　根据赋权方法求解权重 W，用权向量表示为：

$$\vec{W}=(w_1,w_2,\cdots,w_m) \tag{6}$$

满足：$\sum_{j=1}^{m}w_j=1$，w_j 为第 j 个指标的权重值。

步骤 4　模糊优选数学计算模型：

$$u_i=\left[1+\sum_{j=1}^{m}(w_j|r_{ij}-g_j|)^2/\sum_{j=1}^{m}(w_j|r_{ij}-b_j|)^2\right]^{-1},i=1,2,\cdots,n \tag{7}$$

u_i 表示第 i 个数据点隶属于最优点的隶属度。

步骤 5　已知 n 个数据点的优属度，按隶属度最大原则对系统中的方案进行排序，通过人为规定各类别的范围对其进行分类。

3 基于距离差异度的组合赋权方法

解决多属性数据分类问题时，如何确定指标的权重是核心问题之一，权系数是否合适严重影响后续的工作。目前，权重方法一般可分为主观赋权法和客观赋权法。主观赋权法是根据决策者的知识经验或偏好确定每个指标的重要性比例并计算得到权重的方法，其主要特点：1)主观性强，单纯地体现决策者的经验与偏好；2)权重确定过程可再现性比较差、易受干扰；3)权重计算简单、粗略。客观赋权法是根据数据特征确定指标权重的方法，其主要特点：1)客观性强，不依赖决策者的主观经验与偏好；2)权重确定过程可再现性比较强；3)权重计算方法多为完善的数学理论，计算过程较为复杂。主观赋权法和客观赋权法均存在不足和局限性，为了使计算权重的方法更合理、科学，提出组合优化规则来组合主客观权重的模型，使其既能客观反映指标的重要程度，又能体现决策者的主观偏好。

文献[5]提出了“在主客观权重下使所有数据与最优点的偏离越小越好”的组合规则；文献[6]提出了“在主客观权重下使所有数据与最劣点的偏离越大越好”的组合规则。为了直观地表示多属性数据，将数据集、最优点 G 和最劣点 B 映射到二维平面空间，如图 1 所示。如果采用前者计算组合权重，则有 $R_2\succ R_3$；采用后者求解组合权重，则有 $R_2\prec R_3$。由此可见，仅仅采用“与最优点偏离最小”或“与最劣点偏离最大”的组合规则很难判断出某个数据点的优劣性，甚至误导决策者的取向。因此，提出一种新的组合规则：在主客观权重下使数据点与最优点和最劣点之间的距离差异度(见定义 2)越大越好。这样就能够保证数据点与最优点的偏离越小时与最劣点的偏离越大，进而避免因采取不同准则而造成决策结果的不统一，消除其歧义性。

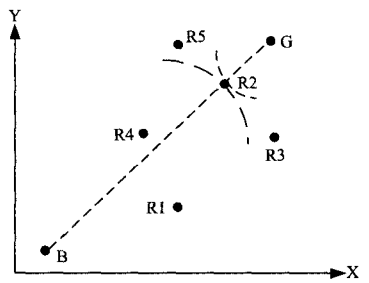

图 1　二维平面映射图

设决策者通过 P 种主观赋权法得到 p 个主观权重 $W_s=(w_{s1},w_{s2},\cdots,w_{sm}),(s=1,2,\cdots,p)$，通过 Q 种客观赋权法得到 q 个客观权重 $W_o=(w_{o1},w_{o2},\cdots,w_{om}),(o=1,2,\cdots,q)$，其中 w_{sj} 和 w_{oj} 分别表示指标的主观、客观权重，并且满足 $\sum_{j=1}^{m}w_{sj}=1,(s=1,2,\cdots,p)$ 和 $\sum_{j=1}^{m}w_{oj}=1,(o=1,2,\cdots,q)$。令 $W^*=(w_1^*,w_2^*,\cdots,w_m^*)$ 为上述 $l=p+q$ 种权重经组合后的组合权重，用向量表示为：

$$w_j^*=\sum_{k=1}^{l}\theta_k w_{kj},j=1,2,\cdots,m \tag{8}$$

其中，θ_k 表示权重组合系数，且 $\sum_{k=1}^{l}\theta_k=1,\theta_k\geqslant 0$。

定义 1　设数据点 R_i 与最优点 G 和最劣点 B 的加权广义距离为 D_i^+ 和 D_i^-，其定义为：

$$\begin{aligned}D_i^+&=\sum_{j=1}^{m}w_j^*|r_{ij}-g_j|\\&=\sum_{j=1}^{m}\sum_{k=1}^{l}\theta_k w_{kj}|r_{ij}-g_j|,i=1,2,\cdots,n\end{aligned} \tag{9}$$

$$\begin{aligned}D_i^-&=\sum_{j=1}^{m}w_j^*|r_{ij}-b_j|\\&=\sum_{j=1}^{m}\sum_{k=1}^{l}\theta_k w_{kj}|r_{ij}-b_j|,i=1,2,\cdots,n\end{aligned} \tag{10}$$

定义 2　数据点 R_i 与最优点和最劣点之间的加权广义距离差异度为：

$$\begin{aligned}Div&=D_i^--D_i^+\\&=\sum_{j=1}^{m}\sum_{k=1}^{l}\theta_k w_{kj}(|r_{ij}-b_j|-|r_{ij}-g_j|)\\&\quad i=1,2,\cdots,n\end{aligned} \tag{11}$$

由定义可知，加权广义距离差异度越大越能够体现距离劣等方案越远越好且距离优等方案越近越好的组合最优化规则，为此构造如下最优化模型：

$$\begin{cases}\text{Max } F_1=\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l}\theta_k w_{kj}(|r_{ij}-b_j|-|r_{ij}-g_j|)\\ \text{s. t. }\sum_{k=1}^{l}\theta_k=1\\ 0\leqslant\theta_k\leqslant 1,k=1,2,\cdots,l\end{cases}\tag{12}$$

为了尽可能地消除组合系数 θ_k 的不确定性影响，根据 Jaynes 最大熵理论[5]，组合系数 θ_k 应该使熵值最大，即

$$\begin{cases}\text{Max } F_2=-\sum_{k=1}^{l}\theta_k\ln\theta_k\\ \text{s. t. }\sum_{k=1}^{l}\theta_k=1\\ 0\leqslant\theta_k\leqslant 1,k=1,2,\cdots,l\end{cases}\tag{13}$$

综合以上两个模型构造如下最优化模型：

$$\begin{cases}\text{Max } F=\mu\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l}\theta_k w_{kj}(|r_{ij}-b_j|-|r_{ij}-g_j|)-\\ \qquad(1-\mu)\sum_{k=1}^{l}\theta_k\ln\theta_k\\ \text{s. t. }\sum_{k=1}^{l}\theta_k=1\\ 0\leqslant\theta_k\leqslant 1,k=1,2,\cdots,l\end{cases}\tag{14}$$

其中，μ 是反映决策者对最优化理论偏好程度的偏好因子，为了求解该最优化模型构造 Lagrange 函数

$$L=\mu\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{l}\theta_k w_{kj}(|r_{ij}-b_j|-|r_{ij}-g_j|)-(1-\mu)\sum_{k=1}^{l}\theta_k\ln\theta_k+2\lambda(\sum_{k=1}^{l}\theta_k-1)\tag{15}$$

其中，λ 是 Lagrange 因子，令

$$\frac{\partial L}{\partial\theta_k}=\mu\sum_{i=1}^{n}\sum_{j=1}^{m}w_{kj}(|r_{ij}-b_j|-|r_{ij}-g_j|)-(1-\mu)(1+\ln\theta_k)+2\lambda=0\tag{16}$$

$$\frac{\partial L}{\partial\lambda}=2(\sum_{k=1}^{l}\theta_k-1)=0\tag{17}$$

经计算解得

$$\begin{cases}\theta_k=\dfrac{\alpha_k}{\sum_{k=1}^{l}\alpha_k},k=1,2,\cdots,l\\ \alpha_k=\exp(\mu\sum_{i=1}^{n}\sum_{j=1}^{m}w_{kj}(|r_{ij}-b_j|-|r_{ij}-g_j|)/(1-\mu)-1)\end{cases}\tag{18}$$

将 θ_k 代入组合权重公式求得 W^*。

4 FO-CA 多属性数据分类方法

模糊优选理论与模型是综合评判理论、模型的一个新途径，能够将高维数据转换为一维有序综合评判指标集，具有思路清晰、计算模型简便实用的优点，但是在最后分类阶段需要人为地划分各类界限，类内各对象之间不一定有较好的相似特性，无法得到理想的分类结果。聚类分析算法较好地解决了低维数据的聚类分类问题[8,9]。鉴于模糊优选模型和聚类分析算法的优缺点，提出了一种基于模糊优选模型与聚类分析算法的多属性数据分类方法(Multiple Attribute Classification Method Based on Fuzzy Optimization and Clustering Analysis，简称 FO-CA)。首先采用模糊优选模型对多属性数据集进行计算得出相对优与次的一维有序综合指标数据集，即对多属性数据集降维；然后采用聚类算法对一维有序综合指标数据集进行聚类，最后分类。该多属性数据分类方法不仅弥补了聚类算法处理高维数据的不足，还避免了模糊优选分类的人为定界的缺陷。FO-CA 分类方法的流程见图 2，具体描述如下：

步骤 1　预处理。数据集的无量纲化。

步骤 2　组合权重。由主观赋权法和客观赋权法得到主客观权重，通过基于距离差异度的组合赋权法求解权重。

步骤 3　有序综合指标集。在模糊优选方法中结合步骤 2 所得组合权重，通过计算得到一维综合指标结果集。

步骤 4　聚类分析。有序综合指标集作为聚类分析的输入，根据综合指标的最大值和最小值平均将其分为 k 段，并选取各段的平均值作为初始中心；然后利用 K-Means 聚类分析法进行聚类。K-Means 法的过程如下[10-12]：

(1)将每个对象视为一个类簇，且每个类簇仅一个对象，计算它们之间的最短距离，类与类之间的距离就是每个类内所包含对象之间的距离，得到初始化距离矩阵；

(2)将距离最近的两个类簇合并成一个新的类簇；

(3)重新计算所有类簇之间的距离；

(4)重复(2)和(3)，直到所有类簇最后合并成一个类簇为止或者达到终止条件，终止条件一般为类簇的数量达到规定的数量或者两个相近的类簇超过阈值。

步骤 5　分类。统计分析类簇内综合指标的分布情况，确定分类范围。

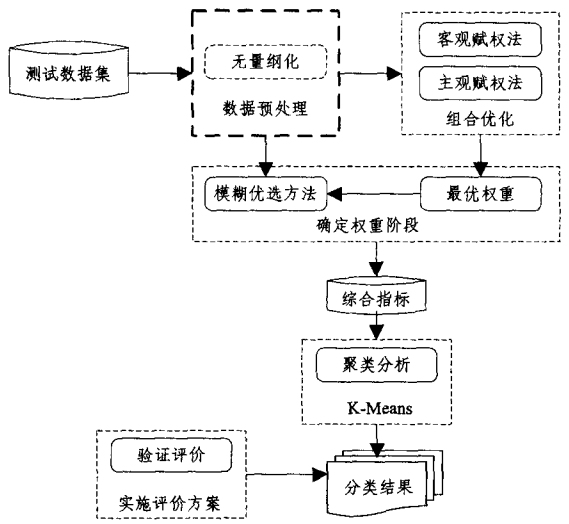

图 2　FO-CA 方法流程图

5 实验与分析

5.1 基于距离差异度的组合权重方法实验分析

实例描述(市民购房的多属性决策问题)：在问题中共有 4 处房源，其方案集用 $P=\{p_1,p_2,p_3,p_4\}$ 表示，房价总额(单位：万元)、全部面积(单位：m^2)、与上班地点的距离(单位：km)、住房设施(%)和住房周围环境(%)构成指标集，用 $S=\{s_1,s_2,s_3,s_4,s_5\}$ 表示，其中第一、三个指标是越小越优型的，第二、四、五个指标是越大越优型的，由熵权法和专家赋权法得到客观、主观权重，实例矩阵 $A=(a_{ij})_{4\times5}$ 为：

$$A=\begin{bmatrix}30&100&10&7&7\\25&80&8&3&5\\18&50&20&5&10\\22&70&12&5&9\end{bmatrix}$$

本文选取基于主客观权重离差最小化的组合赋权方法、基于标准差修正的组合赋权方法、基于综合评价目标值最大的组合权重、基于组合权重与综合评价目标最大的组合权重与基于距离差异度的组合赋权方法进行实验对比(见表 1)。实验结果表明该组合权重方法与其他学者所得出的结果一致。从方法的数学计算模型上进行分析,发现其不仅体现了主观权重和客观权重的组合优化规则,而且具有计算简便、实用的特点。

表 1 不同组合权重下的决策排序

组合赋权方法	权重 w*	决策排序
基于主客观权重离差最小化的集成权重	(0.2420,0.1906,0.1998,0.1686,0.1990)	$p_4 \succ p_1 \succ p_3 \succ p_2$
基于标准差修正的集成权重	(0.2300,0.1910,0.1978,0.1799,0.2013)	$p_4 \succ p_1 \succ p_3 \succ p_2$
基于综合评价目标值最大的组合权重	(0.2400,0.1910,0.2003,0.1690,0.1997)	$p_4 \succ p_1 \succ p_3 \succ p_2$
基于组合权重与综合目标最大的组合权重	(0.2130,0.1896,0.2014,0.1720,0.2240)	$p_4 \succ p_1 \succ p_3 \succ p_2$
基于距离差异度的组合权重	(0.2232,0.1950,0.2017,0.1968,0.1833)	$p_4 \succ p_1 \succ p_3 \succ p_2$

5.2 FO-CA 方法实验分析

本文选取 UCI 中的 Iris、Wine 和 Ruspini 3 个通用数据集作为测试集:Iris 数据集有花萼长度、花萼宽度、花瓣长度、花瓣宽度 4 个指标,包含 Setosa、Versicolour、Virginica 3 种花共 150 项数据,每种类型有 50 个数据对象;Wine 数据集有 Alcohol、Malic acid、Ash 等 13 个指标,包含 3 类酒共 178 项数据(一类 59 项,二类 71 项,三类 48 项);Ruspini 数据集有 x、y 两个指标共 75 项二维点数据。

本文选用错误率作为标准对比分类结果,错误率是被错误分类的数目所占数据集总数的百分比,其计算公式如下:

$$Error=\frac{\varepsilon}{n}\times 100$$

其中,ε 是被错误分类的数目,n 是数据集的数据总数。

在 3 个测试数据集上,将本文提出的 FO-CA 方法与 K-Means 算法和模糊优选方法进行对比实验,如图 3 所示。由于 FO-CA 分类方法通过模糊优选方法得到有序的综合指标数据集作为后续 K-Means 算法的输入,因此该分类算法只需运行一次即可。经典的 K-Means 算法采取随机策略确定初始中心可能会导致聚类结果不一致,在此采用 20 次重复运算取其平均值。

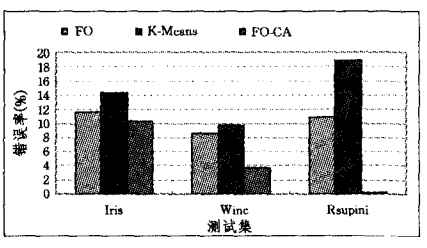

图 3 不同数据集上 FO-CA 与其他两种方法的错误率对比

从分类效果角度分析,在 Iris 和 Wine 数据集中模糊优选方法和 K-Means 算法的错误率相差不大,FO-CA 方法的错误率较低,在 Ruspini 数据集中表现最为明显,FO-CA 方法的错误率最低,与 K-Means 算法的错误率相差近 18%。由于 K-Means 算法的初始阶段的随机性导致了分类错误率偏高,模糊优选方法分类时受到人为划分界限的影响也导致了较高的分类错误率。FO-CA 方法利用两种方法优势互补的特点,消除了上述的负面影响,从而在分类准确率上有了较大幅度的提升。

从算法时间角度分析,由于模糊优选方法是线性时间,因此其算法时间开销最小;K-Means 算法在初始阶段任意选取初始聚类中心导致迭代次数较多,算法时间开销最大;FO-CA 方法在聚类分类阶段弥补了 K-Means 的随机性,所以在时间开销上较小。

结束语 本文在综合主观权重与客观权重优点的基础上,结合指标的特点以"距离差异度"作为约束性条件,通过组合优化思想构建相应的数学模型来求解更接近实际情况的最优权重;同时提出了 FO-CA 分类方法,由于模糊优选方法得出的综合指标值是具有优劣顺序的,该方法将其作为 K-Means 算法的输入,解决了其对数据集输入顺序敏感的问题,同时利用模糊优选方法的降维能力与聚类分析的数据识别能力优势互补,使该分类模型在多属性数据分类问题上有比较好的适用性,相比其它的多属性数据分类方法更为实用。

参 考 文 献

[1] 陈守煜,赵英琪. 模糊优选理论与模型[J]. 模糊系统与数学,1990,4(2):87-91

[2] 俞文彬,谢康林,张忠能. 基于属性分类的数据挖掘方法[J]. 小型微型计算机系统,2000,21(3):305-308

[3] Sharma A,Srinivasan S,Lake L W. Classification of Oil and Gas Reservoirs Based on Recovery Factor:A Data-Mining Approach [C]//SPE. 2010

[4] 纪崑,郑文瑞. 多维多目标模糊优选动态规划及其在资源分配中的应用[J]. 模糊系统与数学,2006,20(2):103-108

[5] 姜昱汐,迟国太,严丽俊. 基于最大熵原理的线性组合赋权方法[J]. 运筹与管理,2011,20(1):103-108

[6] Niu Xin-sheng,Lei Ming,Fu Lei,et al. A combined weighting method for power system restoration decision making[C]//2011 Seventh International Conference on Natural Computation. 2011,3:1223-1227

[7] 王中兴,李桥兴. 依据主客观权重集成最终权重的一种方法[J]. 应用数学与计算数学学报,2006,20(1):87-92

[8] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):49-50

[9] Huang Zhen-ping,Liu Ke-lin,Li Jing. Combination weight fuzzy recognition model and its application in the assessment of water resource renew ability[J]. Computer Application and System Modeling,2010,15:256-259

[10] 伍育红. 浅议聚类分析方法[J]. 计算机科学,2012,39(6A):325-327

[11] Zhou Yong,Xia Shi-xiong. A Novel Clustering Algorithm Based on Hierarchical and K-means Clustering[C]// Control Conference,2007:605-609

[12] Saunders D G O,Win J,Liliana M,et al. Cano,Using Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate Effectors of Rust Fungi[J]. PLOS ONE,2012,7(1):1-6