计算机研究与发展 ISSN 1000-1239/CN 11-1777/TP
Journal of Computer Research and Development 47(10): 1749-1755, 2010

基于新的距离度量的 *K*-Modes 聚类算法

梁吉业[1,2] 白 亮[1] 曹付元[1,2]
[1](山西大学计算机与信息技术学院 太原 030006)
[2](计算智能与中文信息处理教育部重点实验室(山西大学) 太原 030006)
(ljy@sxu.edu.cn)

K-Modes Clustering Algorithm Based on a New Distance Measure

Liang Jiye[1,2], Bai Liang[1], and Cao Fuyuan[1,2]
[1](*School of Computer and Information Technology, Shanxi University, Taiyuan* 030006)
[2](*Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan* 030006)

Abstract The leading partitional clustering technique, *K*-Modes, is one of the most computationally efficient clustering methods for categorical data. In the traditional *K*-Modes algorithm, the simple matching dissimilarity measure is used to compute the distance between two values of the same categorical attributes. This compares two categorical values directly and results in either a difference of zero when the two values are identical or one if otherwise. However, the similarity between categorical values is not considered. In this paper, a new distance measure based on rough set theory is proposed, which overcomes the shortage of the simple matching dissimilarity measure and is used along with the traditional *K*-Modes clustering algorithm. While computing the distance between two values of the same categorical attributes, the new distance measure takes into account not only their difference but also discernibility of other relational categorical attributes to them. The time complexity of the modified *K*-Modes clustering algorithm is linear with respect to the number of data objects which can be applied for large data sets. The performance of the *K*-Modes algorithm with the new distance measure is tested on real world data sets. Comparisons with the *K*-Modes algorithm based on many different distance measures illustrate the effectiveness of the new distance measure.

Key words clustering algorithm; categorical data; rough set; rough membership degree; distance measure

摘 要 传统的 *K*-Modes 聚类算法采用简单的 0-1 匹配差异方法来计算同一分类属性下两个属性值之间的距离, 没有充分考虑其相似性. 对此, 基于粗糙集理论, 提出了一种新的距离度量. 该距离度量在度量同一分类属性下两个属性值之间的差异时, 克服了简单 0-1 匹配差异法的不足, 既考虑了它们本身的异同, 又考虑了其他相关分类属性对它们的区分性. 并将提出的距离度量应用于传统 *K*-Modes 聚类算法中. 通过与基于其他距离度量的 *K*-Modes 聚类算法进行实验比较, 结果表明新的距离度量是更加有效的.

关键词 聚类算法; 分类属性数据; 粗糙集; 粗糙隶属度; 距离度量
中图法分类号 TP181

收稿日期: 2008-06-23; 修回日期: 2010-02-05
基金项目: 国家"八六三"高技术研究发展计划基金项目(2007AA01Z165); 国家自然科学基金项目(60773133, 70971080); 山西省自然科学基金项目(2008011038); 山西省高校科技开发项目(2007103).

0 引 言

聚类分析[1]是数据挖掘中的一个重要研究领域,由于聚类分析不对数据作任何统计假设,在模式识别和人工智能等领域,聚类分析常被称为一种无监督的学习.目前聚类分析已被广泛应用于金融欺诈、医疗诊断、图像处理、信息检索和生物信息学等研究领域[1-6].

K-Means 聚类算法[7]是目前较流行的一种聚类方法,由于其简单、易实现,被广泛应用于各个领域,但它仅仅局限于对数值属性数据聚类.然而在现实生活中的数据含有大量分类属性数据(即符号属性数据),因此,分类属性数据的聚类已成为一个重要研究课题.1997 年 Huang[8-9]对 *K*-Means 聚类算法进行了扩展,提出了针对分类属性数据的 *K*-Modes 聚类算法,该算法采用简单 0-1 匹配方法来计算同一分类属性下两个属性值之间的距离,虽然此距离度量由于其简单被广泛使用,但同时也弱化了类内的相似性,没有充分反映同一分类属性下两个属性值之间的距离.目前许多学者从不同角度对此进行了研究.Ng 等人[10]和 San 等人[11]通过属性值在类内出现的频率来计算同一属性下两个属性值之间的距离.Li 等人[12]提出了基于生物特征的距离度量.但以上两种方法都没有充分考虑同一属性下不同属性值之间的距离.Hsu 等人[13-14]提出一种基于概念层次的距离度量,但过多依赖于用户的经验和专业知识.Ganti 等人[15]和 Ahmad 等人[16-17]等人通过属性值之间的共现程度来反映同一属性下两个属性值之间的距离,但也因此忽略了它们本身的异同.上述方法都不同程度地对同一属性下两个属性值之间的距离度量进行了改善.

由波兰学者 Pawlak 提出的粗糙集理论[18-20]是一种处理不精确、不确定和模糊知识的软计算工具,在处理分类属性数据方面有着独特的优势[18-22].许多学者将其用于分类属性数据聚类[23-27].本文基于粗糙集理论,提出了一种新的距离度量.此距离度量既体现了同一属性下两个属性值本身的异同,也体现了其他相关属性对它们的可区分性.通过与基于其他距离度量的 *K*-Modes 聚类算法实验比较,表明基于新距离度量的 *K*-Modes 聚类算法更加有效.

1 粗糙集的基本概念

设四元组 $S=(U, A, V, f)$ 是一个分类属性信息系统,其中 U 是对象的非空有限集合,称为论域;A 是分类属性的非空有限集合;$V=\bigcup_{a\in A} V_a$, V_a 是属性 a 的值域;$f: U\times A \to V$ 是一个信息函数,即对 $\forall a\in A, x\in U, f(x, a)\in V_a$;通常 $S=(U, A, V, f)$ 也简记为 $S=(U, A)$.

令 $B\subseteq A$, $x, y\in U$,定义属性集 B 的不可区分关系 $IND(B)$ 为:$IND(B)=\{(x, y)\in U\times U \mid \forall a\in B, f(x, a)=f(y, a)\}$. $[x]_B$ 表示包含 $x\in U$ 的 B 等价类.

给定分类属性信息系统 $S=(U, A)$,对于每个子集 $X\subseteq U$ 和 $B\subseteq A$,称子集

$$\underline{B}(X)=\{x\in U \mid [x]_B\subseteq X\}$$

和

$$\bar{B}(X)=\{x\in U \mid [x]_B\cap X\neq\varnothing\}$$

分别为 X 的 B 下近似集和上近似集.

集合 $bn_B(X)=\bar{B}(X)-\underline{B}(X)$ 称为 X 的 B 边界域;$pos_B(X)=\underline{B}(X)$ 称为 X 的 B 正域;$neg_B(X)=U-\bar{B}(X)$ 称为 X 的 B 负域.显然,$\bar{B}(X)=pos_B(X)\cup bn_B(X)$,且 $\underline{B}(X)\subseteq X\subseteq\bar{B}(X)$.

元素 $x\in U$ 关于集合 $X\subseteq U$ 的粗糙隶属度为

$$\mu_X^B(x)=\frac{|[x]_B\cap X|}{|[x]_B|},$$

其中,$0\leqslant\mu_X^B(x)\leqslant 1$.

当 B 中只包含一个元素 a 时,为了书写方便,通常用 a 代替 B,并将 $IND(B)$, $\underline{B}(X)$, $\bar{B}(X)$, $bn_B(X)$, $\mu_X^B(x)$ 记为 $IND(a)$, $\underline{a}(X)$, $\bar{a}(X)$, $bn_a(X)$, $\mu_X^a(x)$.

由于 *K*-Modes 聚类算法是一种专门针对分类属性数据的算法,因此,可以将 *K*-Modes 聚类算法中处理的数据形式化地描述为分类属性信息系统,此表述等同于 Huang 在文献[8]中分类属性数据的描述.

2 传统 *K*-Modes 聚类算法

K-Modes 聚类算法是通过对 *K*-Means 聚类算法的扩展,使其应用于分类属性数据聚类.它采用简单匹配方法度量同一分类属性下两个属性值之间的距离,用 Mode 代替 *K*-Means 聚类算法中的 Means,通过基于频率的方法在聚类过程中不断更新 Modes.

定义 1[8]. 设 $S=(U, A)$ 是一个分类信息系统,$U=\{x_1, x_2, \cdots, x_n\}$, $A=\{a_1, a_2, \cdots, a_m\}$, $x_i, x_j\in U$ $(1\leqslant i, j\leqslant n)$, x_i 和 x_j 分别被 A 描述为 $x_i=(f(x_i,$

$a_1), f(x_i, a_2), \cdots, f(x_i, a_m))$ 和 $x_j = (f(x_j, a_1), f(x_j, a_2), \cdots, f(x_j, a_m))$, x_i 与 x_j 距离定义为

$$d(x_i, x_j) = \sum_{l=1}^{m} \delta(f(x_i, a_l), f(x_j, a_l)),$$

其中:

$$\delta(f(x_i, a_l), f(x_j, a_l)) = \begin{cases} 1, & f(x_i, a_l) \neq f(x_j, a_l); \\ 0, & f(x_i, a_l) = f(x_j, a_l). \end{cases}$$

Huang[8] 为实现 K-Modes 聚类算法定义目标函数为

$$F(\boldsymbol{W}, Z) = \sum_{l=1}^{k} \sum_{i=1}^{n} w_{il} d(x_i, z_l),$$

其中:

$$w_{il} \in \{0, 1\}, 1 \leqslant l \leqslant k, 1 \leqslant i \leqslant n; \quad (1)$$

$$\sum_{l=1}^{k} w_{il} = 1, 1 \leqslant i \leqslant n; \quad (2)$$

$$0 < \sum_{i=1}^{n} w_{il} < n, 1 \leqslant l \leqslant k; \quad (3)$$

$\boldsymbol{W}$ 是一个 $n \times k$ 的 $\{0, 1\}$ 矩阵, n 表示对象集 U 所包含的对象个数, k 表示聚类的个数, $w_{il} = 1$ 表示第 i 个对象被划分到第 l 类中, $Z = \{z_1, z_2, \cdots, z_k\}$, $z_l (1 \leqslant l \leqslant k)$ 是第 l 类的中心.

为了使目标函数 F 在满足约束条件式 (1)~(3) 下达到极小化, K-Modes 聚类算法基本步骤如下:

Step1. 从数据集中随机选择 k 个对象作为初始类中心, 其中 k 表示聚类个数;

Step2. 应用简单匹配方法计算对象与类中心 (Modes) 之间的距离, 并将每个对象分配到离它最近的类中去;

Step3. 基于频率方法重新计算各类的类中心 (Modes);

Step4. 重复上述 Step3, Step4 过程, 直到目标函数 F 不再发生变化为止.

3 基于粗糙集的距离度量

由于数值属性具有天然几何性, 可以通过属性值之间的差值来反映其距离. 而分类属性却缺乏这样的几何性, 但这并不能说明同一分类属性下两个属性值之间没有相似性. 客观上, 同一分类属性下的两个值之间的相似程度既取决于它们本身, 又取决于它们所处的环境, 即其他相关属性对它们的影响. 因此, 本文基于粗糙集理论, 提出一种新的距离度量. 此距离度量既体现了同一属性下两个属性值本身的异同, 又体现了其他相关属性对它们的区分性.

定义 2. 设 $S = (U, A)$ 是一个分类信息系统, $A = \{a_1, a_2, \cdots, a_m\}$, 对于任意 $a_i \in A$, 设 $p, q \in V_{a_i}$, p 和 q 相对于 a_i 的内部距离定义为:

$$\delta_{a_i}(p, q) = \begin{cases} 1, & p \neq q; \\ 0, & p = q. \end{cases}$$

定义 3. 设 $S = (U, A)$ 是一个分类信息系统, $A = \{a_1, a_2, \cdots, a_m\}$, 对于任意 $a_i \in A$, 设 $p, q \in V_{a_i}$, p 和 q 相对于属性 $a_j (j \neq i)$ 的外部距离定义为

$$\delta_{a_j}(p, q) = \frac{1}{|U|} \sum_{x \in U} | \mu_X^{a_j}(x) - \mu_Y^{a_j}(x) |,$$

其中, $X = \{x | f(x, a_i) = p, x \in U\}$, $Y = \{x | f(x, a_i) = q, x \in U\}$.

定义 3 表明从属性 a_j 所提供的信息来看, 可以将 U 划分为 3 部分, 即 $\underline{a_j}(X) \cup \underline{a_j}(Y)$, $bn_{a_j}(X) \cup bn_{a_j}(Y)$, $U - \overline{a_j}(X) \cup \overline{a_j}(Y)$.

当 $x \in \underline{a_j}(X) \cup \underline{a_j}(Y)$ 时, 从对象 x 角度来看, 完全可以将 p 和 q 区分开来, 即 $\mu_X^{a_j}(x) = 1$ 或 $\mu_Y^{a_j}(x) = 1$; 当 $x \in U - \overline{a_j}(X) \cup \overline{a_j}(Y)$ 时, 对于 x 来说, p 和 q 是相同的, 即 $\mu_X^{a_j}(x) = 0$ 且 $\mu_Y^{a_j}(x) = 0$; 当 $x \in bn_{a_j}(X) \cup bn_{a_j}(Y)$ 时, $| \mu_X^{a_j}(x) - \mu_Y^{a_j}(x) |$ 越大, 则表明通过 x 将 p 和 q 区分开的可能性越大.

性质 1. $0 \leqslant \delta_{a_j}(p, q) \leqslant 1$. 对于任意 $x \in U$, 当 $\mu_X^{a_j}(x) = \mu_Y^{a_j}(x)$ 时, $\delta_{a_j}(p, q) = 0$; 当 $\underline{a_j}(X) \cup \underline{a_j}(Y) = U$, $\delta_{a_j}(p, q) = 1$.

性质 2. 当 $x \in \underline{a_j}(X) \cup \underline{a_j}(Y)$ 时, $| \mu_X^{a_j}(x) - \mu_Y^{a_j}(x) | = 1$.

性质 3. 当 $x \in U - \overline{a_j}(X) \cup \overline{a_j}(Y)$ 时, $| \mu_X^{a_j}(x) - \mu_Y^{a_j}(x) | = 0$.

性质 4. 当 $\overline{a_j}(X) \cap \overline{a_j}(Y) = \varnothing$ 时, $\delta_{a_j}(p, q) = \dfrac{| \overline{a_j}(X) \cup \overline{a_j}(Y) |}{|U|}$.

性质 5. 设 $S = (U, A)$ 是一个分类信息系统, $A = \{a_1, a_2, \cdots, a_m\}$, $C_t \in U/IND(a_j)$, $0 < t \leqslant |V_{a_j}|$, 对于任意 $a_i \in A$, 设 $p, q \in V_{a_i}$, p 和 q 相对于属性 $a_j (j \neq i)$ 的外部距离可等价地定义为

$$\delta_{a_j}(p, q) = \sum_{t=1}^{|V_{a_j}|} \frac{|C_t|}{|U|} \times \left| \frac{|C_t \cap X|}{|C_t|} - \frac{|C_t \cap Y|}{|C_t|} \right| = \frac{1}{|U|} \sum_{t=1}^{|V_{a_j}|} | |C_t \cap X| - |C_t \cap Y| |,$$

其中, $X = \{x | f(x, a_i) = p, x \in U\}$, $Y = \{x | f(x, a_i) = q, x \in U\}$.

定义 4. 设 $S = (U, A)$ 是一个分类信息系统, $A = \{a_1, a_2, \cdots, a_m\}$, 对于任意 $a_i \in A$, 设 $p, q \in V_{a_i}$, p 和 q 关于属性集 A 的距离定义为

$$\delta(p,q)=\frac{1}{m}\sum_{j=1}^{m}\delta_{a_j}(p,q).$$

性质 **6.** $0\leqslant\delta(p,q)\leqslant 1$. 当 $p=q$ 时, $\delta(p,q)=0$; 当 $p\neq q$ 时, $\frac{1}{m}\leqslant\delta(p,q)\leqslant 1$.

定义 **5.** 设 $S=(U,A)$ 是一个分类信息系统, $U=\{x_1,x_2,\cdots,x_n\}$, $A=\{a_1,a_2,\cdots,a_m\}$, $x_i,x_j\in U$ $(1\leqslant i,j\leqslant n)$, x_i 和 x_j 分别被 A 描述为 $x_i=(f(x_i,a_1),f(x_i,a_2),\cdots,f(x_i,a_m))$ 和 $x_j=(f(x_j,a_1),f(x_j,a_2),\cdots,f(x_j,a_m))$, x_i 与 x_j 之间的距离定义为

$$d_1(x_i,x_j)=\sum_{l=1}^{m}\delta(f(x_i,a_l),f(x_j,a_l)).$$

容易验证, 对于任意 $x,y,z\in U$, d_1 有以下性质:

① 对称性. $d_1(x,y)=d_1(y,x)$.

② 非负性. $d_1(x,y)\geqslant 0$.

③ 三角不等式. $d_1(x,y)+d_1(y,z)\geqslant d_1(x,z)$.

显然, d_1 是分类属性数据对象集上一个度量空间.

例 **1.** 表 1 是由 UCI 中的 Mushroom 的部分数据组成的一个关于蘑菇特征的分类属性信息系统 $S=(U,A)$.

Table 1 An Information System About Mushroom Features
表 **1** 一个关于蘑菇特征的信息系统

Object	*cap-shape*	*cap-surface*	*cap-color*	*odor*	*class*
x_1	convex	smooth	buff	fishy	Poisonous
x_2	convex	scaly	red	fishy	Poisonous
x_3	knobbed	scaly	white	none	Poisonous
x_4	knobbed	scaly	red	fishy	Poisonous
x_5	sunken	fibrous	gray	none	edible
x_6	bell	smooth	white	anise	edible
x_7	bell	smooth	white	almond	edible
x_8	bell	fibrous	gray	none	edible
x_9	convex	smooth	gray	none	edible

其中, $U=\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8,x_9\}$, $A=\{cap\text{-}shape, cap\text{-}surface, cap\text{-}color, odor\}$, *class* 表示数据对象所属的类.

应用简单 0-1 匹配方法计算得到 $d(x_4,x_5)=d(x_7,x_5)=4$. 此时, 无法判断 x_5 与 x_4,x_7 中哪个对象更接近. 当应用新的距离度量时可以有效地作出判断. 下面给出计算过程:

根据定义 2, 可计算得出 $\delta_{cap\text{-}shape}(\text{knobbed},\text{sunken})=1$.

根据定义 3, 可计算得出:

$$\delta_{cap\text{-}surface}(\text{knobbed},\text{sunken})=\frac{1}{9}\times(4\times|0-0|+3\times|\frac{2}{3}-0|+2\times|0-\frac{1}{2}|)=\frac{1}{3}.$$

$$\delta_{cap\text{-}color}(\text{knobbed},\text{sunken})=\frac{1}{9}\times(1\times|0-0|+3\times|\frac{1}{3}-0|+2\times|\frac{1}{2}-0|+3\times|0-\frac{1}{3}|)=\frac{1}{3}.$$

$$\delta_{odor}(\text{knobbed},\text{sunken})=\frac{1}{9}\times(3\times|\frac{1}{3}-0|+4\times|0-\frac{1}{4}|+1\times|0-0|+1\times|0-0|)=\frac{2}{9}.$$

根据定义 4, knobbed 与 sunken 的关于属性集 A 的距离为

$$\delta(\text{knobbed},\text{sunken})=\frac{1}{4}\times(1+\frac{1}{3}+\frac{1}{3}+\frac{2}{9})=\frac{17}{36}.$$

同理, 可计算 scaly 与 fibrous 的关于属性集 A 的距离为

$$\delta(\text{scaly},\text{fibrous})=\frac{1}{4}\times(\frac{5}{9}+1+\frac{5}{9}+\frac{1}{3})=\frac{11}{18}.$$

red 与 gray 的关于属性集 A 的距离为

$$\delta(\text{red},\text{gray})=\frac{1}{4}\times(\frac{1}{3}+\frac{5}{9}+1+\frac{5}{9})=\frac{11}{18}.$$

fishy 与 none 的关于属性集 A 的距离为

$$\delta(\text{fishy},\text{none})=\frac{1}{4}\times(\frac{1}{3}+\frac{1}{3}+\frac{7}{9}+1)=\frac{11}{18}.$$

因此, x_4 与 x_5 的距离为

$$d_1(x_4,x_5)=\frac{17}{36}+\frac{11}{18}+\frac{11}{18}+\frac{11}{18}=\frac{83}{36}.$$

同理, 可计算 x_7 与 x_5 的距离为

$$d_1(x_7,x_5)=\frac{5}{12}+\frac{1}{2}+\frac{7}{12}+\frac{1}{2}=2.$$

显然, $d_1(x_4,x_5)>d_1(x_7,x_5)$, 说明在 d_1 度量下 x_7 较 x_4 更接近于 x_5, 这与实际问题的描述也是相符的.

4 基于新距离度量的 *K*-Modes 聚类算法

设 $S=(U,A)$ 是一个分类信息系统. 下面将新的距离度量应用于传统的 *K*-Modes 聚类算法. 定义目标函数 F_1 为

$$F_1(\boldsymbol{W},Z)=\sum_{l=1}^{k}\sum_{i=1}^{n}w_{il}d_1(x_i,z_l),$$

其中, F_1 同样满足约束条件式(1)~(3).

基于新距离度量的 *K*-Modes 聚类算法步骤如下:

Step1. 根据定义 4, 应用新距离度量公式计算每个属性下任意两个属性值之间的距离;

Step2. 从 U 中随机选择 k 个对象作为初始类中心, 其中 k 表示聚类个数;

Step3. 根据定义 5, 计算对象与每个类中心的距离, 并将每个对象分配到离它最近的类中去;

Step4. 应用与传统 K-Modes 聚类算法同样的方法去更新各类中心;

Step5. 重复上述 Step3, Step4 过程, 直到目标函数 F_1 不再发生变化为止.

通过分析可知, 基于新距离度量的 K-Modes 聚类算法的时间复杂度为 $O(nms+m^2s^3+knmt)$, 传统的 K-Modes 聚类算法的时间复杂度为 $O(knmt)$, 其中 n 表示数据集 U 中包含的对象数, m 为属性个数, s 为 $\max_{a\in A}|V_a|$, k 为聚类个数, t 为最大迭代次数. 由于在现实数据中 $n\gg m$, $n\gg s$, 所以当 n 足够大时, $O(knmt)$ 和 $O(nms+m^2s^3+knmt)$ 相对于 n 来说都是线性的.

5 实验分析

下面分别从分类正确率(accuracy)、类精度(precision)、召回率(recall) 和迭代次数(iteration) 4 个方面[28] 来分析算法的聚类质量: Accuracy(AC), Precision(PE), Recall(RE) 分别定义如下:

$$AC=\frac{\sum_{i=1}^{k}a_i}{n};\ PE=\frac{\sum_{i=1}^{k}\frac{a_i}{a_i+b_i}}{k};\ RE=\frac{\sum_{i=1}^{k}\frac{a_i}{a_i+c_i}}{k};$$

其中, n 表示数据集的对象数, a_i 表示正确分到第 i 类的对象数, b_i 表示误分到第 i 类的对象数, c_i 表示应该分到第 i 类却没有分到的对象数, k 表示聚类个数.

为了测试新距离度量的有效性, 从 UCI 数据集中挑选了 3 组数据 Vote, Breast Cancer 和 Mushroom, 并将基于新距离度量的 K-Modes 聚类算法分别与 Huang 的 K-Modes[8] 和 Ahmad 的 K-Modes[17] 进行比较. 3 组数据描述如表 2 所示:

Table 2 Description of Data Sets

表 2 数据描述

Data Set	Samples	Attributes	The Class	
			I	II
Vote	435	16	267	168
Breast cancer	699	16	458	241
Mushroom	8124	22	3916	4208

由于 K-Modes 聚类算法的聚类结果受初始类中心的选择的影响, 不同的初始类中心可能有不同的聚类结果, 所以对于数据 Vote, Breast Cancer 和 Mushroom 分别随机选择 100 组类中心, 使每个算法运行 100 次, 通过计算平均聚类质量来验证算法的有效性. 表 3~5 是基于不同的距离度量的 K-Modes 聚类算法性能比较.

Table 3 Comparison with Algorithms on the Vote Data

表 3 在 Vote 下算法的性能比较

Validation Measure	Huang's K-Modes	Ahmad's K-Modes	Proposed K-Modes
AC	0.8602	0.8751	0.8782
PE	0.8561	0.8700	0.8730
RE	0.8743	0.8890	0.8921
Iteration	3.5300	3.5000	3.6200

Table 4 Comparison with Algorithms on the Breast Cancer Data

表 4 在 Breast Cancer 下算法的性能比较

Validation Measure	Huang's K-Modes	Ahmad's K-Modes	Proposed K-Modes
AC	0.8538	0.9307	0.9396
PE	0.8697	0.9360	0.9452
RE	0.7967	0.9254	0.9303
Iteration	3.6300	3.5000	3.6200

Table 5 Comparison with Algorithms on the Mushroom Data

表 5 在 Mushroom 下算法的性能比较

Validation Measure	Huang's K-Modes	Ahmad's K-Modes	Proposed K-Modes
AC	0.7176	0.7592	0.7745
PE	0.7453	0.7786	0.7926
RE	0.7132	0.7526	0.7703
Iteration	5.4100	4.8500	5.3600

通过分析表 3~5, 在数据 Vote, Breast Cancer 和 Mushroom 上, 基于新距离度量的 K-Modes 聚类算法得到了较好的聚类效果, 优于基于其他距离度量的 K-Modes 聚类算法. 以上实验结果表明新距离度量是有效的, 基于新距离度量的 K-Modes 聚类算法能在较少的迭代中得到较高的聚类精度.

6 结 论

本文基于粗糙集理论, 提出一个新的距离度量. 此距离度量从属性本身和其他相关属性两个角度对同一属性下两个属性值之间的距离进行度量. 将此距离度量应用于传统的 K-Modes 聚类算法, 通过与基于其他距离度量的 K-Modes 聚类算法进行实验

比较,结果表明基于新距离度量的 K-Modes 聚类算法总体上优于其他的 K-Modes 聚类算法,并且能在较少的迭代中得到较高的聚类精度.

参 考 文 献

[1] Han Jiawei, Kamber M. Data Mining Concepts and Techniques[M]. San Francisco: Morgan Kaufmann, 2001

[2] Brendan J F, Delbert D. Clustering by passing messages between data points[J]. Science, 2007, 315(16): 972-976

[3] Zhang Jiangshe, Liang Yi, Xu Zongben. Clustering methods by simulating visual systems [J]. Chinese Journal of Computers, 2001, 24(5): 496-501 (in Chinese)
(张讲社, 梁怡, 徐宗本. 基于视觉系统的聚类算法[J]. 计算机学报, 2001, 24(5): 496-501)

[4] Zhang Jiangshe, Liang Yiuwing. Improved possibilistic c-means clustering algorithms [J]. IEEE Trans on Fuzzy Systems, 2004, 12(2): 209-217

[5] Yu Jian. On the fuzziness index of the FCM algorithms[J]. Chinese Journal of Computers, 2003, 26(8): 968-973 (in Chinese)
(于剑. 论模糊 c 均值算法的模糊指数. 计算机学报[J], 2003, 26(8): 968-973)

[6] Chen Zonghai, Wen Feng, Nie Jianbin, et al. A reinforcement learning method based on node-growing K-means clustering algorithm [J]. Journal of Computer Research and Development, 2006, 43(4): 661-666
(陈宗海, 文锋, 聂建斌, 等. 基于节点生长 K-均值聚类算法的强化学习方法[J]. 计算机研究与发展, 2006, 43(4): 661-666)

[7] Mac Q J. Some methods for classification and analysis of multivariate observation[C]//Proc of the 5th Berkley Symp on Mathematical Statistics and Probability. Berkley, California: University of California Press, 1967: 281-297

[8] Huang Zhexue. Clustering large data sets with mixed numeric and categorical values [C]//Proc of PAKDD'97. Singapore: World Scientific, 1997: 21-35

[9] Huang Zhexue. Extensions to the K-means algorithm for clustering large data sets with categorical values [J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304

[10] Ng M K, Li Junjie, Huang Zhexue, et al. On the impact of dissimilarity measure in K-modes clustering algorithm [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2007, 29(3): 503-507

[11] San O M, Huynh V N, Nakamori Y. An alternative extension of the K-means algorithm for clustering categorical data[J]. Int Journal Application Mathematic and Computer Science, 2004, 14(2): 241-247

[12] Li Cen, Biswas G. Unsupervised learning with mixed numeric and nominal data [J]. IEEE Trans on Knowledge and Data Engineering, 2002, 14(4): 673-690

[13] Hsu Chungchian, Chen Chinlong, Su Yuwei. Hierarchical clustering of mixed data based on distance hierarchy [J]. Information Sciences, 2007, 177(20): 4474-4492

[14] Hsu Chungchian. Generalizing self-organizing map for categorical data [J]. IEEE Trans on Neural Network, 2006, 17(2): 294-304

[15] Ganti V, Gehrke J, Ramakrishnan R. CACTUS, clustering categorical data using summaries [C]//Proc of the 5th Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 1999: 73-83

[16] Ahamad A, Dey L. A K-mean clustering algorithm for mixed numeric and categorical data [J]. Data & Knowledge Engineering, 2007, 63(2): 503-527

[17] Ahamad A, Dey L. A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set [J]. Pattern Recognition Letters, 2007, 28(1): 110-118

[18] Pawlak Z. Rough Sets—Theoretical Aspects of Reasoning About Data [M]. London: Kluwer Academic Publishers, 1991

[19] Zhang Wenxiu, Wu Weizhi, Liang Jiye, et al. Rough Set Theory and Approach[M]. Beijing: Science Press, 2001 (in Chinese)
(张文修, 吴伟志, 梁吉业, 等. 粗糙集理论与方法[M]. 北京: 科学出版社, 2001)

[20] Liang Jiye, Li Deyu. Uncertainty and Knowledge Acquisition in Information Systems [M]. Beijing: Science Press, 2005 (in Chinese)
(梁吉业, 李德玉. 信息系统中的不确定性与知识获取[M]. 北京: 科学出版社, 2005)

[21] Liang Jiye, Shi Zhongzhi, Li Deyu, et al. Information entropy, rough entropy and knowledge granulation in incomplete information systems [J]. Int Journal of General Systems, 2006, 35(6): 641-654

[22] Qian Yuhua, Liang Jiye, Li Deyu, et al. Measures for evaluating the decision performance of a decision table in rough set theory [J]. Information Sciences, 2008, 178(1): 181-202

[23] Parmar D, Wu T, Blackhurst J. MMR: An algorithm for clustering categorical data using rough set theory [J]. Data & Knowledge Engineering, 2007, 63(3): 879-893

[24] Chen Chunbao, Wang Liya. Rough set-based clustering with refinement using Shannon's entropy theory [J]. Computer and Mathematics with Application, 2006, 52(10/11): 1563-1576

[25] Kumar P, Krishna P R, Bapi R S, et al. Rough clustering of sequential data [J]. Data & Knowledge Engineering, 2007, 63(2): 183-199

[26] Jiang Feng, Sui Yuefei, Cao Cungen. A rough set approach to outlier detection [J]. Int Journal of General Systems, 2008, 37(5): 519-536

[27] An Qiusheng, Shen Junyi, Wang Guoyin. A clustering method based on information granularity and rough sets[J]. Pattern Recognition and Artificial Intelligence, 2003, 16(4): 412-417 (in Chinese)
(安秋生, 沈钧毅, 王国胤. 基于信息粒度与粗糙集的聚类方法研究[J]. 模式识别与人工智能, 2003, 16(4): 412-417)
[28] Yang Yiming. An evaluation of statistical approaches to text categorization[J]. Journal of Information Retrieval, 1999, 1(1/2): 67-88

Liang Jiye, born in 1962. Professor and PhD supervisor. Senior member of China Computer Federation. His main research interests include rough set theory, data mining, artificial intelligence, etc.
梁吉业, 1962 年生, 教授, 博士生导师, 中国计算机学会高级会员, 主要研究方向为粗糙集理论、数据挖掘、人工智能等.

Bai Liang, born in 1982. PhD candidate. His main research interests include machine learning.
白 亮, 1982 年生, 博士研究生, 主要研究方向为机器学习.

Cao Fuyuan, born in 1974. PhD and lecturer. His main research interests include cluster analysis and machine learning.
曹付元, 1974 年生, 博士, 讲师, 主要研究方向为聚类分析、机器学习.

Research Background

This work is supported by the High Technology Research and Development Program of China (No. 2007AA01Z165), the National Natural Science Foundation of China (No. 60773133, 70971080), the Natural Science Foundation of Shanxi (No. 2008011038), and the Technology Research Development Projects of Shanxi (No. 2007103).

Since first published in 1997, the *K*-Modes algorithm has become a popular technique in solving categorical data clustering problems in different application domains. However, the distance between values of the same categorical attributes computed with the simple matching similarity measure is either 0 or 1. This often results in clusters with weak intra similarity. To overcome the shortage, a new distance measure based on rough set theory is proposed in this paper, which is implemented to traditional *K*-Modes clustering algorithm. The distance measure takes into account not only the difference of two values themselves of the same attributes but also other attributes' discernible degree to them. The computational cost and the performance of the *K*-Modes clustering algorithm with the new distance measure is analyzed. Experimental comparisons with the *K*-Modes clustering algorithms based on other distance measures on standard data sets, which are taken from UCI repository, illustrate the effectiveness of the *K*-Modes clustering algorithm with the new dissimilarity measure.