基于粗糙集的改进 K-Modes 聚类算法

白 亮 梁吉业 曹付元
(计算智能与中文信息处理教育部重点实验室 太原 030006)
(山西大学计算机与信息技术学院 太原 030006)

摘 要 传统的 K-Modes 算法采用简单匹配的方法来计算对象之间的距离,并没有充分考虑同一属性下的两个不同值之间的相似性。基于粗糙集中的上、下近似,提出了一种新的距离度量,并重新定义了类中心,对传统 K-Modes 算法进行了改进。与其他改进 K-Modes 算法进行了比较,实验结果表明,基于粗糙集的改进 K-Modes 算法有效地提高了聚类精度。

关键词 聚类算法,粗糙集,距离度量

Improved K-Modes Clustering Algorithm Based on Rough Sets

BAI Liang LIANG Ji-ye CAO Fu-yuan
(Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Taiyuan 030006, China)
(School of Computer and Information Technology, Shanxi University, Taiyuan030006, China)

Abstract Traditional K-Modes clustering algorithm uses a simple matching dissimilarity measure to compute the distance between two objects. However, the similarity between two values of the same attributes is not considered. A new distance measure based on upper and lower approximations in rough set theory was proposed, and a new description of cluster center was defined. Traditional K-Modes clustering algorithm was improved. By comparing with other improved K-Modes algorithms, experimental results illustrate that the improved K-Modes clustering algorithm based on rough sets increases the clustering accuracy.

Keywords Clustering algorithm, Rough sets, Distance measure

1 引言

聚类分析[1]是数据挖掘研究和应用中的一个重要部分,由于聚类算法不对数据作任何统计假设,在模式识别和人工智能等领域,聚类算法常被称为一种无监督的学习。聚类分析是将数据对象分组成多个类或多个簇,在同一个簇中的对象具有较高的相似度,而不同簇中的对象差别较大。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。

K-Means[2]算法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型数据聚类。然而在现实生活数据中含有大量的字符型数据,因此,字符型数据的聚类已成为一个重要的研究内容。1997 年 Huang[3,4]对 K-Means 算法进行了扩展,提出了针对字符型数据的 K-Modes 聚类算法和针对混合数据的 K-Prototypes 聚类算法。K-Modes 是用简单匹配方法度量对象之间的距离,用 mode 代替 K-Means 算法中的均值,通过基于频率的方法在聚类过程中不断更新 mode 使目标函数最小化。但是它有两个不足,一是它采用简单匹配方法来计算对象之间距离,认为同一属性下的两个值,如果相同距离为 0,反之为 1,弱化了其相似性;二是用 mode 表示一个类的中心,没有充分考虑属性在类上的分布,且不具有唯一性。

目前许多学者对同一属性下的两个值的相似性进行了研究。Ng[5]基于频率方法定义了同一属性下的两个值之间的相似性。Li[6]提出了基于生物特征的属性值之间的距离(相似性)度量。虽然文献[5,6]提高了聚类精度,但其并没有充分考虑不同属性值之间的相似性。Hsu[7,8]等人提出一种基于概念层次的方法来计算字符型属性值之间的距离,虽然其考虑了属性之间的相似性,但过多依赖于用户的经验和专业知识。Ganti[9]的 CACTUS 方法是通过计算某一属性的两个值与其它属性值的同现次数来决定相似度。Amir Ahmad[10,11]提出了一种基于条件概率的距离度量来测试字符型属性值之间的距离。尽管在符号值之间的距离度量已经有许多,但目前仍然没有一种能普遍接受的方法。

粗糙集理论[12]在处理字符值方面有着独特的优势,因此,本文应用粗糙集理论中的上、下近似,从属性本身和其它

到稿日期:2008-04-30 本文受国家 863 计划项目(2007AA01Z165),国家自然科学基金(60773133),高等学校博士学科点专项科研基金(20050108604),教育部科学技术研究重点项目(206017),山西省重点实验室开放基金(200603023),山西省高校科技开发项目(2007103)和太原市科技局科技兴市专项项目(07010724)资助。

白 亮(1982—),男,硕士研究生,主要研究方向为数据挖掘、机器学习;梁吉业(1962—),男,教授,博士生导师,主要研究方向为粗糙集理论、数据挖掘、人工智能;曹付元(1974—),男,博士研究生,主要研究方向为数据挖掘、机器学习。

相关属性两个角度出发，提出了一种新的同一属性下的两个不同值之间的相似性度量。此外，改进了类中心表示方法，使其能够体现属性在该类上的分布，保证其唯一性。与其他改进 K-Modes 算法进行了比较，实验结果表明，基于粗糙集的改进 K-Modes 算法有效地提高了聚类精度。

2 粗糙集的基本概念

定义 1[12] 设四元组 $S=(U,A,V,f)$ 是一个信息系统，其中 U 是对象的非空有限集合，称为论域；A 是字符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 $S=(U,A)$。

定义 2[12] 设 $S=(U,A)$ 是一个信息系统，$B\subseteq A$，$x,y\in U$，B 上的不可区分关系定义为：

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

对于任意 $x\in U$ 的等价类记为：$[x]_B=\{y\mid \forall y\in U,(x,y)\in IND(B)\}$。

定义 3[12] 设 $S=(U,A)$ 是一个信息系统，$X\subseteq U$，$B\subseteq A$，X 关于 B 的上近似和下近似分别为：

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

$\overline{B}(X)=\{x\in U\mid [x]_B\cap X\neq\emptyset\}$，

$bn_B(X)=\overline{B}(X)-\underline{B}(X)$称为 X 的 B 边界域；$pos_B(X)=\underline{B}(X)$称为 X 的 B 正域；$neg_B(X)=U-\overline{B}(X)$称为 X 的 B 负域。显然，$\overline{B}(X)=pos_B(X)\cup bn_B(X)$。集合 X 在上近似和下近似之间：$\underline{B}(X)\subseteq X\subseteq\overline{B}(X)$。

当 $|B|=1$ 时，为了书写方便，我们用 $a\in B$ 代替 B，将 $\underline{B}(X)$，$\overline{B}(X)$，$bn_B(X)$ 记为 $\underline{a}(X)$，$\overline{a}(X)$，$bn_a(X)$。

3 基于粗糙集的对象之间的距离公式

在实际数据中，同一属性下两个值之间是否相似，既取决于属性值之间本身，又取决于所处的环境，即其它相关属性。下面从两个方面对属性值相似性进行定义。

定义 4 设 $S=(U,A)$ 是一个信息系统，$A=\{a_1,a_2,\cdots,a_m\}$，对于任意 $a_i\in A$，设 $p,q\in V_{a_i}$，p 和 q 的内部相似度可以定义为：

$$ISim(p,q)=\begin{cases}1,(p=q)\\0,(p\neq q)\end{cases}$$

定义 5 设 $S=(U,A)$ 是一个信息系统，$A=\{a_1,a_2,\cdots,a_m\}$，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 的外部相似度定义为：

$$ESim_{a_j}(p,q)=\frac{|\overline{a_j}(X)\cap\overline{a_j}(Y)|}{|\underline{a_j}(X)\cup\underline{a_j}(Y)|}$$

其中 $X=\{x\mid f(x,a_i)=p, x\in U\}$ 和 $Y=\{x\mid f(x,a_i)=q, x\in U\}$。$p$ 和 q 相对于其它所有属性 $a_j(i\neq j)$ 的平均外部相似度定义为

$$SESim(p,q)=\frac{1}{m-1}\sum_{j=1,j\neq i}^{m}ESim_{a_j}(p,q)$$

性质 1 设 $S=(U,A)$ 是一个信息系统，当 $X\cap Y=\emptyset$ 时，有 $\underline{a_j}(X)\cap\underline{a_j}(Y)=\emptyset$，且 $\overline{a_j}(X)\cap\overline{a_j}(Y)=bn_{a_j}(X)\cap bn_{a_j}(Y)$，即

$$ESim_{a_j}(p,q)=\frac{|bn_{a_j}(X)\cap bn_{a_j}(Y)|}{|\overline{a_j}(X)\cup\overline{a_j}(Y)|}$$

性质 2 当 $IND(\{a_i\})\supseteq IND(\{a_j\})$，$(i\neq j)$ 时，$ESim_{a_j}(p,q)=0$；当对于 $\forall a_j\in A-\{a_i\}$，$IND(\{a_i\})\supseteq IND(\{a_j\})$ 时，$SESim(p,q)=0$。

定义 6 设 $S=(U,A)$ 是一个信息系统，$A=\{a_1,a_2,\cdots,a_m\}$，对于任意 $a_i\in A$，设 $p,q\in V_{a_i}$，p 和 q 的相似度定义为：

$$Sim(p,q)=\frac{ISim(p,q)+SESim(p,q)}{2}$$

p 和 q 的距离为 $\delta(p,q)=1-Sim(p,q)$，其中 $0\leqslant Sim(p,q)\leqslant 1$。

定义 7 设 $S=(U,A)$ 是一个信息系统，$A=\{a_1,a_2,\cdots,a_m\}$，$x,y\in U$，x 和 y 距离定义为：

$$d(x,y)=\sum_{i=1}^{m}\delta(f(x,a_i),f(y,a_i))。$$

例 1 一个信息系统 $S=(U,A)$，如表 1 所示。

表 1

	a_1	a_2	a_3	a_4
x_1	L	B	E	F
x_2	L	C	E	F
x_3	M	C	E	G
x_4	M	D	E	H
x_5	M	D	E	H

计算 L 和 M 的相似度如下：

根据定义 4，L 和 M 的内部相似度为：$ISim(L,M)=0$；从属性 a_1 所提供的信息来看，L 和 M 可以完全区分开来。

根据定义 5，L 和 M 相对于属性 a_2 的外部相似度为：

$$ESim_{a_2}(L,M)=\frac{|\{1,2,3\}\cap\{2,3,4,5\}|}{|\{1,2,3\}\cup\{2,3,4,5\}|}=\frac{2}{5}$$

从属性 a_2 所提供的信息来看，当对象在属性 a_2 上的值为 B，那么该对象的属性 a_1 的值一定为 L；当对象在属性 a_2 上的值为 D，那么该对象的属性 a_1 的值一定为 M，但当属性 a_2 的取值为 C 时无法将 L 和 M 区分，说明通过属性 a_2 可以将 L 和 M 部分分开。

L 和 M 相对于属性 a_3 的外部相似度为：

$$ESim_{a_3}(L,M)=\frac{|\{1,2,3,4,5\}\cap\{1,2,3,4,5\}|}{|\{1,2,3,4,5\}\cup\{1,2,3,4,5\}|}=1$$

从属性 a_3 所提供的信息来看，L 和 M 无法区分，所以从属性 a_3 的角度来看，L 和 M 是相同的。

L 和 M 相对于属性 a_4 的外部相似度为：

$$ESim_{a_4}(L,M)=\frac{|\{1,2\}\cap\{3,4,5\}|}{|\{1,2\}\cup\{3,4,5\}|}=0$$

从属性 a_4 所提供的信息来看，L 和 M 完全可以区分开，当对象在属性 a_4 上的值为 F，那么该对象的属性 a_1 的值一定为 L；当对象在属性 a_4 上的值为 G,H 时，那么该对象的属性 a_1 的值一定为 M，所以从属性 a_4 的角度来看，L 和 M 是完全不同的。

L 和 M 相对于其他所有属性的平均外部相似度为：

$$ESim(L,M)=\frac{1}{3}\sum_{j=2}^{4}ESim_{a_j}(L,M)=\frac{1}{3}\times(\frac{2}{5}+0+1)=\frac{7}{15}$$

根据定义 6，L 和 M 的相似度为：

$$Sim(p,q)=\frac{ISim(p,q)+SESim(p,q)}{2}=\frac{1}{2}\times(0+\frac{7}{15})=\frac{7}{30}$$

L 和 M 的距离为：

$$\delta(L,M)=1-\mathrm{Sim}(L,M)=1-\frac{7}{30}=\frac{23}{30}$$

4 类中心的描述

对于数值型属性，我们可以求出它在某一类中的均值(mean)来表示该类的类中心。而对于字符型属性，我们无法求出它在某一类中的均值。K-Modes 算法用 mode 代替均值，也就是选择该属性在类中出现频率最高的值来表示该类的类中心在其上的取值。但此方法没有充分考虑属性在该类上的分布，这样得到的类中心很难准确地反映该类的特征。

基于文献[10]的思想，我们对类中心给出新的描述，使其充分反映属性在该类上的分布。

定义 8[10] 设 $S=(U,A)$ 是一个信息系统，$A=\{a_1,a_2,\cdots,a_m\}$，$X\subseteq U$，X 的类中心为向量 $C=[C_1,C_2,\cdots,C_m]$，表示形式如下：

$$C_i=\frac{n_1}{n}a_{i,1}+\frac{n_2}{n}a_{i,2}+\cdots+\frac{n_t}{n}a_{i,t}$$

其中 $a_i\in A$，它有 t 个属性值，分别为 $a_{i,1},a_{i,2},\cdots,a_{i,t}$，$C_i$ 是对类 X 在属性 a_i 取值情况的描述，n 表示类 X 所包含的数据对象的个数，n_j 表示类 X 中对象在属性 a_i 取值为 $a_{i,j}$ 时的对象个数。

定义 9[10] $S=(U,A)$ 是一个信息系统，$A=\{a_1,a_2,\cdots,a_m\}$，$X\subseteq U$，向量 $C=[C_1,C_2,\cdots,C_m]$ 是 X 的类中心，$x\in U$，x 与 C 在属性 $a_i\in A$ 上距离公式为：

$$\delta(f(a_i,x),C_i)=\sum_{j=1}^{t}\frac{n_j}{n}\delta(f(a_i,x),a_{i,j})$$

其中 $f(a_i,x)$ 表示 x 在属性 $a_i\in A$ 上的取值，t 表示 a_i 有 t 个属性值，分别为 $a_{i,1},a_{i,2},\cdots,a_{i,t}$，$n$ 表示类 X 所包含的数据对象的个数，n_j 表示类 X 中数据对象的属性 a_i 取值为 $a_{i,j}$ 时的对象个数。

例 2 假设颜色属性有 3 个取值分别是 *red*，*yellow*，*blue*。这三个值在某一类中分布如下：类中颜色属性值为 *red* 的对象占了 40%，*yellow* 占了 39%，*blue* 占了 21%。传统的 K-Modes 选择 *red* 作为 mode 在颜色属性上的取值，此时的 mode 并没有充分反映该类在颜色属性上的取值，丢失了信息。根据定义 9 我们给出类中心在颜色属性上的描述为：

$0.4Red+0.39Yollow+0.21Blue$

此描述反映了颜色属性在类上的分布。

设对象 x 在颜色属性上取值为 *red*，根据定义 10，x 与该类的类中心在颜色属性上距离为：

$0.4\times\delta(red,red)+0.39\times\delta(red,yollow)+0.21\times\delta(red,blue)$

5 基于粗糙集的改进 K-Modes 聚类算法

本文所建议的算法是在传统 K-Modes 算法的基础上，改进了对象之间的距离度量和类中心的表示。基本步骤如下：

1. 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；

2. 根据第 3 部分给定的距离公式，计算对象与每个类的类中心的距离，分配该对象到离它最近的类中心所代表的类中；

3. 根据第 4 部分，重新计算每个类的类中心；

4. 重复上述 2，3 过程，直到类中对象不再发生变化。

6 实验分析

为了评价聚类质量，我们引入了文献[10]所提及的聚类正确率(Micro-precision)，正确率定义为：

$$Micro-p=(\sum_{i=1}^{k}b_i)/n$$

其中 n 表示数据集的对象数，b_i 表示正确分到第 i 类的对象数，k 表示聚类个数。正确率越高，那么聚类结果越好。

为了测试该算法的有效性，我们从 UCI 数据集中挑选了 3 组数据，分别为 soybean，vote 和 mushroom，3 组数据描述如表 2 所示。

表 2 数据描述

Data Set	Samples	Attributes	I 类	II 类	III 类	IV 类
Soybean	47	35	10	10	10	17
Vote	435	16	267	168	0	0
mushroom	8124	22	3916	4208	0	0

表 3 和表 4 是数据集 soybean 在选择第 1，2，3 和 4 条记录为初始类中心时，基于粗糙集的改进 K-Modes 和传统 K-Modes 的聚类结果。

表 3 在 soybean 上传统 K-Modes 聚类结果

Cluster Number	D	C	R	P	b_i
1	10	10	0	0	10
2	0	0	10	0	10
3	0	0	0	10	10
4	0	0	0	7	7
Micro-p					0.787

表 4 在 soybean 上基于粗糙集的改进 K-Modes 聚类结果

Cluster Number	D	C	R	P	b_i
1	10	0	0	0	10
2	0	10	0	0	10
3	0	0	0	17	17
4	0	0	10	0	10
Micro-p					1.000

表 5 和表 6 是 vote 选择第 1 和第 2 条记录为初始类中心时，基于粗糙集的改进 K-Modes 和传统 K-Modes 的聚类结果。

表 5 在 vote 上传统 K-Modes 聚类结果

Cluster Number	D	R	b_i
1	223	15	223
2	44	153	153
Micro-p			0.864

表 6 在 vote 上基于粗糙集的改进 K-Modes 聚类结果

Cluster Number	D	R	b_i
1	221	7	221
2	46	161	161
Micro-p			0.878

表 7 和表 8 是 mushroom 选择第 151 和第 3708 条记录为初始类中心时，基于粗糙集的改进 K-Modes 和传统 K-Modes 的聚类结果。

(下转第 176 页)

面得到应用，当然值得进一步探询。

参 考 文 献

[1] 王国俊. 数理逻辑引论与归结原理[M]. 北京：科学出版社，2006

[2] 闫林. 数理逻辑基础与粒计算[M]. 北京：科学出版社，2007

[3] 刘清，黄兆华. G-逻辑及其归结推理. 计算机学报[J]，2004，27(7)：865-873

[4] Yan Lin，Liu Qing. A Logical Method of Formalization for Granular Computing[C]// Proceedings of 2007 IEEE International Conference on Granular Computing. Silicon Valley，California，USA，2007：22-27

[5] Yan Lin，Wang Sui-hua，Zhang Xue-dong. Semantic Reasoning Study for Rough Logic About n-ary Formulas[C]// Proceedings of 2006 IEEE International Conference on Granular Computing. Atlanta，Georgia，USA，2006：381-384

[6] Liu Qing，Wang Ji-yi. Semantic Analysis of Rough Logical Formulas Based on Granular Computing[C]// Proceedings of 2006 IEEE International Conference on Granular Computing. Atlanta，Georgia，USA，2006：393-396

[7] Pawlak Z. Rough Logic. Bulletin of Polish Academy of Sciences Technical Sciences[J]，1987，35(5/6)：253-258

(上接第 164 页)

表 7 在 mushroom 上传统 K-Modes 聚类结果

Cluster Number	E	P	b_i
1	285	1996	1996
2	3923	1920	3923
Micro-p			0.729

表 8 在 mushroom 上基于粗糙集的改进 K-Modes 聚类结果

Cluster Number	E	P	b_i
1	96	3100	3100
2	4112	816	4112
Micro-p			0.888

由于 K-Modes 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，所以我们对于数据 soybean，vote 和 mushroom 随机选择 100 组类中心，并将基于粗糙集的改进 K-Modes 算法分别与传统 K-Modes[3]，Ahmad 的 K-Modes[10] 和 Ng 的 K-Modes[5] 进行比较，使每个算法分别运行 100 次，通过计算平均聚类正确率，来验证基于粗糙集的改进 K-Modes 算法的有效性。表 9 是不同的 K-Modes 算法的聚类性能比较。

表 9 在 3 种不同的数据集下算法的性能比较

	传统 K-Modes	Ahmad 的 K-Modes	Ng 的 K-Modes	基于粗糙集的改进 K-Modes
soybean	86%	90%	93%	92%
vote	86%	87%	86%	88%
mushroom	71%	77%	79%	81%

通过以上实验表明，基于粗糙集的改进 K-Modes 算法有效地提高了聚类效果。

结束语 本文利用粗糙集中的上、下近似，提出了一种新的距离度量，该距离公式既考虑了属性值本身的不同，又考虑了属性值相对于其它相关属性的相似度。此外，对类中心进行了重新定义，使其充分反映类的特征。与其他改进 K-Modes 算法进行了比较，实验结果表明，基于粗糙集的改进 K-Modes 算法有效地提高了聚类精度。

参 考 文 献

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

[2] MacQueen J B. Some methods for classification and analysis of multivariate observation // Proceeding 5th Berkley Symposium. on Mathematical Statistics and Probability. 1967，I：281-297. University of California Press，1967，Xvii，666

[3] Huang Zhexue. Clustering Large Data Sets with Mixed Numeric and Categorical Values// PAKDD'97. Singapore，World Scientific，1997：21-35

[4] Huang Zhexue. Extensions to the k-Means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery，1998，2：283-304

[5] Michael K，Ng M，Li Junjie，et al. On the impact of dissimilarity measure in K-Modes clustering algorithm. IEEE Transaction on Pattern Analysis and Machine Intelligence，2007，29(3)：503-507

[6] Li Cen，Biswas Gautam. Unsupervised learning with mixed numeric and nominal data. IEEE Transactions on Knowledge and Data Engineering，2002，14：673-690

[7] Hsu C C，Chen Chinlong，Su Yuwei. Hierarchical clustering of mixed data based on distance hierarchy. Information Sciences，2007：4474-4492

[8] Hsu C C. Generalizing self-organizing map for categorical data. IEEE Transaction on Neural Network，2006，17(2)：294-304

[9] Ganti V，Ramakrishnanz J G R. CACTUS，clustering categorical data using summaries // Proceedings of the 5th International Conference on Knowledge Discovery and Data Mining. San Diego：ACM Press，1999：73-83

[10] Ahmad A，Dey L. A k-mean clustering algorithm for mixed numeric and categorical data. Data & Knowledge Engineering，2007，63：503-527

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

[12] 张文修，吴伟志，梁吉业，等. 粗糙集理论与方法. 科学出版社，2003

[13] 张敏，于剑. 基于划分的模糊聚类算法. 软件学报，2004，15(6)：858-868

[14] 陈宁，陈安，周龙骧. 数值型和分类型混合数据的模糊 K-Prototypes 聚类算法. 软件学报，2001，12(8)：1107-1119

[15] 郭建生，赵奕，施鹏飞. 一种有效的用于数据挖掘的动态概念聚类算法. 软件学报，2001，12(4)：582-591