文章编号:1001 - 9081(2008)06 - 1441 - 03

适用于区间数据的基于相互距离的相似性传播聚类

谢信喜,王士同
(江南大学 信息工程学院,江苏 无锡 214122)
(xiexinxi@163.com)

摘 要:符号聚类是对传统聚类的重要扩展,而区间数据是一类常见的符号数据。传统聚类中使用的对称性度量不一定适用于度量区间数据,且算法初始化也一直是干扰聚类的严重问题。因此,提出了一种适用于区间数据的度量——相互距离,并在此度量的基础上采用了一种全新的聚类方法——相似性传播聚类,解决了初始化干扰问题,从而得出了适用于区间数据的基于相互距离的相似性传播聚类。通过理论阐述和实验比较,说明了该算法比基于欧氏聚类的 K-均值算法要好。

关键词:符号聚类;区间数据;相互距离;相似性传播;K-均值
中图分类号: TP301 **文献标志码**:A

Affinity propagation clustering for symbolic interval data based on mutual distances

XIE Xin-xi, WANG Shi-tong
(*School of Information and Technology, Jiangnan University, Wuxi Jiangsu* 214122, *China*)

Abstract: Clustering for symbolic data is an important extension of conventional clustering, and interval representation for symbolic data is often used. The symmetrical measures in conventional clustering algorithms are sometimes not fit to interval data and the initialization is another severe problem that can affect the clustering algorithms. One metric called mutual distances for interval data was proposed; based on the metric, a new clustering method named affinity propagation clustering that could solve the problem initialization was used. Then, affinity propagation clustering for symbolic interval data based on mutual distance was given. Theoretical explanation and experiments indicate that the proposed algorithm outperforms K-means based on Euclidean distances for the interval symbolic data.

Key words: clustering of symbol; interval data; mutual distance; affinity propagation; K-means

0 引言

聚类在模式识别、数据挖掘、机器学习、图像处理等领域发挥着重要作用。所谓的聚类就是根据若干个属性,按照一定的相似性或不相似性度量,把数据集分成若干类,其目的是要让同类的数据间的相似度尽量大,不同类的数据间相似度尽量小。文献[1]中对聚类算法的各个方面进行了总结。聚类算法基本上可分为层次聚类、划分聚类、基于密度的聚类、基于模型的聚类和基于网格的聚类。其中,划分聚类中的K-均值聚类算法应用广泛。但是传统聚类大都针对连续的数值型数据,如果将它们直接应用到符号数据(符号数据包括区间数据、属性多值数据、带权重的属性多值数据、布尔数据、名词性数据等)中,效果往往不尽人意;而现实中,符号数据常常出现,特别是区间数据(例如:人的正常血压范围、一个地区一个月内的气温变化范围等),所以需要提出新的有效算法或对原有算法进行一定的改进。

度量是聚类的基础,根据数据的性质和结构选择一个合适的度量是保证良好聚类效果的根本。对于传统数据,常用的度量有 Minkowski 距离、Euclidean 距离、城市距离、无穷 Minkowski 距离、Mahalanobis 距离、Pearson 相关性、点对称距离、余弦相似性等[1]648。文献[2]采用可调整的欧氏距离对区间数据进行度量;文献[3-4]基于 position、span、content 提出了针对符号数据的不相似性度量和相似性度量;文献[5]给出了适用于混合数据(符号数据和传统数据的混合,即既有符号属性又有传统的属性)的 Minsowski 距离度量;文献[6]针对区间数据,提出了可调整的 Hausdoff 距离。但这些度量都默认了符号数据之间的距离(不相似度或相似度)是对称的,即 $d(X,Y) = d(Y,X)$,可事实并非都是如此。

表 1 赠血相似度表格

捐赠者血型	受赠者血型 A	B	AB	O
A	Yes	No	Yes	No
B	No	Yes	Yes	No
AB	No	No	Yes	No
O	Yes	Yes	Yes	Yes

表 1 是赠血时,不同血型之间的相似度矩阵,当符号数据 X 到符号数据 Y 的相似度为 *Yes* 时表示 X 型血可以输给 Y 型血,从表中可以看出,X 到 Y 的相似度并不总等于 Y 到 X 的相似度,即相似度(不相似度或距离)不对称。因此,基于符号数

收稿日期:2007 - 12 - 06;**修回日期**:2008 - 01 - 16。 **基金项目**:国家 863 计划项目(2007AA1Z158;2006AA10Z313);国家自然科学基金资助项目(60773206/F020106;60704047/F030304);2004 年教育部跨世纪优秀人才支持计划基金项目(NCET-04-0496);2005 年教育部科学研究重点基金项目(105087);中国科学院自动化所模式识别国家重点实验室开放课题。

作者简介:谢信喜(1981 -),男,福建宁德人,硕士研究生,主要研究方向:人工智能、模式识别; 王士同(1964 -),男,江苏扬州人,教授,博士生导师,主要研究方向:人工智能、模式识别、神经模糊系统、生物信息学。

据度量中的不对称情况，本文提出了一种适用于区间数据的度量——相互距离(Mutual Distance，MD)，它分别考虑X到Y的距离和Y到X的距离，最后综合决定X与Y之间的距离。

大部分聚类算法包括K-均值算法的前提是要指定最终的聚类数目以及对应数量的初始聚类中心，但算法对这些初始化很敏感。以K-均值算法为例，要想得到较满意的聚类结果，初始化就要合理，为了降低初始化不合理所造成的不良影响，通常采用的方法是使用不同的初始化，多次运行算法，最终取平均效果或最佳的那个，这样的做法代价太大。最近，文献[7]提出了一种全新的相似性传播聚类(Affinity Propagation Clustering)方法，成功地解决了这些初始化问题。它不需要指定聚类数目和初始聚类中心，只需要给出数据之间的相似度矩阵，最终产生各个聚类以及各类的中心点，摆脱初始化对算法的干扰。本文将该聚类方法应用于区间数据。

1 相互距离

度量是聚类的基础，度量选择的好坏直接影响聚类的效果，应该根据数据本身的性质以及数据集的结构选用合适的度量。在传统的聚类中，最常用的度量是欧氏距离。但用欧氏距离度量区间数据的效果不如在常规数据中的应用，而且欧氏距离默认数据之间的距离是对称的。为此，本文提出了一种新的度量概念，称之为相互度量。

设区间数据集为$D = \{X_1, X_2, \cdots, X_i, \cdots, X_N\}$，数据点$X_i = (x_i^1, x_i^2, \cdots, x_i^k, \cdots, x_i^P)$，属性$x_i^k = [a_i^k, b_i^k]$$(a_i^k \leqslant b_i^k)$，其中$N$为数据集大小，$P$为数据点维数。

1.1 欧氏距离

$$D(X_i, X_j) = \sqrt{\sum_{k=1}^{P} \left((a_i^k - a_j^k)^2 + (b_i^k - b_j^k)^2 \right)} \tag{1}$$

其中$X_i, X_j \in D$，并且$D(X_i, X_j) = D(X_j, X_i)$。

1.2 相互距离

$$d_{i \to j}^k = \begin{cases} 0, & a_i^k \geqslant a_j^k \text{ 且 } b_i^k \leqslant b_j^k \\ \dfrac{(|a_i^k - a_j^k| + |b_i^k - b_j^k|)}{span_{ij}^k}, & \text{其他} \end{cases} \tag{2}$$

其中$d_{i \to j}^k$为区间$x_i^k = [a_i^k, b_i^k]$到区间$x_j^k = [a_j^k, b_j^k]$的距离，$span_{ij}^k$为包含这两个区间的最小区间，即$\max(b_i^k, b_j^k) - \min(a_i^k, a_j^k)$。当$x_i^k = [a_i^k, b_i^k]$包含于$x_j^k = [a_j^k, b_j^k]$时，显然$x_i^k = [a_i^k, b_i^k]$到$x_j^k = [a_j^k, b_j^k]$的距离为0；其余的情况下，应同时考虑两个区间的上限之间以及下限之间的关系，$span_{ij}^k$用来规范度量。

同理x_j^k到x_i^k的距离：

$$d_{j \to i}^k = \begin{cases} 0, & a_j^k \geqslant a_i^k \text{ 且 } b_j^k \leqslant b_i^k \\ \dfrac{(|a_j^k - a_i^k| + |b_j^k - b_i^k|)}{span_{ji}^k}, & \text{其他} \end{cases} \tag{3}$$

其中$span_{ji}^k = span_{ij}^k$。

由于x_i^k到x_j^k的距离并非总等于x_j^k到x_i^k的距离，在度量x_i^k与x_j^k的距离时要同时考虑这二者。因此，我们定义x_i^k与x_j^k之间的距离为：

$$d(x_i^k, x_j^k) = \frac{(d_{i \to j}^k + d_{j \to i}^k)}{2} \tag{4}$$

则最终数据点X_i与X_j的相互距离为：

$$MD(X_i, X_j) = \frac{1}{P} \sum_{k=1}^{P} d(x_i^k, x_j^k) \tag{5}$$

2 基于相互距离的相似性传播聚类

2.1 K-均值聚类

K-均值算法是具有代表性的传统聚类算法，其原理比较简单：

输入：数据集D；

输出：各个类及其中心。

算法步骤：

1)初始化：指定聚类数目C及其各个类的初始中心；

2)归类：按照一定的度量准则，将各个数据点归入与其最相似的中心所在的类；

3)确定中心：根据当前的归类，利用各个类中的数据重新确定该类的中心；

4)循环执行步骤2)与3)，直到算法收敛或迭代停止的条件被满足。

K-均值算法的特点是算法简单，执行速度也比较快，因此被广泛地应用。但是跟其他很多传统聚类算法一样，聚类前要指定聚类数目和初始中心，且算法对这些初始化很敏感，而实际中要做到合理初始化比较困难，虽然也存在一定的方法来比较合理地确定聚类数目[8-9]和初始化中心，但是均以牺牲效率为代价。

2.2 相似性传播聚类

相似性传播聚类[7]972-973一开始就把所有数据点都看做可能的中心，任意两个数据点i和j之间存在两个信息量$a(i, j)$和$r(i, j)$，前者表示i可以选择j作为其中心点的累积证据，后者表示j可以作为i中心点的累积证据；每个点可以看做网络中的一个节点，网络节点之间不断地传播前面的两个信息量，直到一个好的聚类结果(各个类及其对应中心)出现。

相似性传播聚类算法不要求指定聚类数和初始化中心，只需要输入数据点之间的相似性矩阵S。$s(i, j)$表示i归入以j为中心的类的可能性，该值越大，i越有可能归入该类中。当$i = j$时，$s(j, j)$表示j作为类中心的可能性，显然$s(j, j)$越大，j越有可能成为中心。通常根据一定的度量，计算出所有$s(i, j)$$(i \neq j)$后，$s(j, j)$可以取一个常量，如果所有的数据点一开始作为中心点的可能性一样，则取相同的常量，需要注意的是，$s(j, j)$的值直接影响着聚类的结果，一般情况下，$s(j, j)$越大，最后产生的聚类越多。

$a(i, j)$表示i可以选择j作为其中心点的累积证据，如果网络中越多的点选择j作为其中心点，这表明j越有可能成为中心，则i选择j可能性也就越大。在以往算法迭代过程中，通常由数据点去选择中心(或者向中心靠近)，中心是被动的。但相似性传播聚类考虑双向的因素，即i主观地想选择j作为其中心。然而j并不一定可以作为i的中心点，还得考虑$r(i, j)$，即j可以作为i中心点的累积证据。在j想争做i中心点的同时，网络中还存在着其他候选的中心点也想成为i的中心点，j要与这些候选的中心点通过竞争来得到i。这样，在算法的迭代过程中，网络中的每个数据点都在不断地扮演选择与被选择的角色，最终确定谁来充当中心以及各个非中心点的最终归宿。算法迭代过程中，数据点之间通过如下的迭代公式来传播信息：

$$r(i, j) = s(i, j) - \max_{j' \neq j} \{a(i, j') + s(i, j')\} \tag{6}$$

$$a(i,j) = \min\left\{0, r(j,j) + \sum_{i' \notin \{i,j\}} \max\{0, r(i',j)\}\right\} \quad (7)$$

当 $i = j$ 时,$a(i,j)$ 的迭代式为:

$$a(i,j) = \sum_{i' \neq j} \max\{0, r(i',j)\} \quad (8)$$

初始化时 $a(i,j) = 0$,算法开始迭代时先更新 $r(i,j)$。停止迭代后,对于数据点 i,若 $a(i,j) + r(i,j)$ 最大,表明 i 的中心就是 j;若 $i = j$ 时,表明 j 本身是个中心点。

为了防止迭代过程中出现震荡,算法引入了防止震荡的因子 λ。r 与 a 按如下规则迭代:

$$\begin{cases} a_{new} = (1-\lambda) \cdot a_{new} + \lambda \cdot a_{old} \\ r_{new} = (1-\lambda) \cdot r_{new} + \lambda \cdot r_{old} \end{cases} \quad (9)$$

其中,a_{new} 与 r_{new} 为当前迭代得到的值,a_{old} 与 r_{old} 为上次迭代得到的值。λ 取 0 到 1 之间的值,本文实验中 λ 均取 0.5。

2.3 基于相互距离的相似性传播聚类

这里直接采用 2.2 节中的相似性传播算法,对于该算法要求输入的相似性矩阵 S,可以根据第 1 章给出的相互距离 MD 来计算,数据点之间的相似性取其相互距离的负值。这样就得到了适用于区间数据的基于相互距离的相似性传播聚类算法。

输入:数据集 D;

输出:各个类及其中心。

算法步骤:

1) 使用相互距离计算相似性矩阵 S,并指定其中的 $s(j,j)$($j = 1,2,\cdots,N$,N 为数据集大小);

2) 使用式(6)、(9) 更新 r;

3) 使用式(7) ~ (9) 更新 a;

4) 循环执行步骤 2) 与 3),直到算法收敛或迭代停止条件被满足为止;

5) 根据 $\max\limits_{1 \leq j \leq N}\{a(i,j) + r(i,j)\}$($i = 1,2,\cdots,N$) 来确定各个类以及各类的中心。

3 实验结果与分析

使用两个真实数据集(Fat Oil 数据集[10] 和 Temperature 数据集[11])和一个人工数据集,通过实验证明了本文算法比基于欧氏距离的 K-均值聚类算法有效。由于两个算法使用的度量不一样,所以本文跟文献[6]一样,采用一个外部评价准则——Corrected Rand Index(CR)来评价二者的聚类结果。

3.1 聚类结果的评价准则 CR

使用 CR 准则要已知标准的分类,本文使用数据的标准分类为 $U = \{u_1, u_2, \cdots, u_i, \cdots, u_R\}$,而聚类的结果为 $V = \{v_1, v_2, \cdots, v_j, \cdots, v_C\}$,则:

$$CR = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{N_{ij}}{2} - \binom{N}{2}^{-1}\sum_{i=1}^{R}\binom{N_{i.}}{2}\sum_{j=1}^{C}\binom{N_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{N_{i.}}{2} + \sum_{j=1}^{C}\binom{N_{.j}}{2}\right] - \binom{N}{2}^{-1}\sum_{i=1}^{R}\binom{N_{i.}}{2}\sum_{j=1}^{C}\binom{N_{.j}}{2}} \quad (10)$$

其中,N 为数据集中的数据个数,$N_{i.}$ 为 u_i 中的数据个数,$N_{.j}$ 为 v_j 中的数据个数,N_{ij} 为 u_i 与 v_j 的公共数据个数,$\binom{N}{2} = \frac{N(N-1)}{2}$。$CR$ 的取值为[-1,1],值越大表明聚类结果越接近于真实情况。

3.2 真实数据集

Fat oil 数据集[10] 中有 8 个数据点,每个数据点有 4 个区间属性,该数据的最佳分类为 {(1,2,3,4,5,6),(7,8)};Temperature 数据集[11] 中有 37 个数据点,每个数据点有 12 个区间属性,分别表示 1 月—12 月中每个月的最低和最高气温,该数据分为 8 类时的最佳分类为 {(3,4,5,6,7,9,13,16,18,20,23,24,30,32),(11),(12),(19),(21),(31),(33),(1,2,8,10,14,15,17,22,25,26,27,28,29,34,35,36,37)}。由于 K-均值算法选择初始中心的随机性,所以这两个实验中让 K-均值算法运行 100 次,取这 100 次聚类的 CR 的平均值作为最后结果,而相似性传播算法只运行 1 次,实验结果如表 2 所示。

表 2　K-均值与 APCIMD 在 Fat Oil 和 Temperature 上的 CR 比较

数据集	K-均值聚类	相似性传播聚类
Fat Oil 数据集	-0.0583	1.0000
Temperature 数据集	0.2988	0.3703

表 2 的数据结果表明,无论对区间数据集 Fat Oil 还是 Temperature,使用基于相互距离的相似性传播算法的聚类效果要优于基于欧氏距离的 K-均值算法。

3.3 人工数据集

分别随机生成三个服从如下分布的二维正态分布数据集,大小分别为 150、150、50。

1) $u_1 = 28, u_2 = 22, \sigma_1^2 = 100, \sigma_2^2 = 9$;

2) $u_1 = 60, u_2 = 30, \sigma_1^2 = 9, \sigma_2^2 = 144$;

3) $u_1 = 45, u_2 = 38, \sigma_1^2 = 9, \sigma_2^2 = 9$。

这三个数据集就构成 350 个数据点的二维数据集,取该数据集中的任意一点 $X_i = (x_i^1, x_i^2)$($i = 1,2,\cdots,350$),作如下转换:$\left(\left[x_i^1 - \frac{r_1}{2}, x_i^1 + \frac{r_1}{2}\right], \left[x_i^2 - \frac{r_2}{2}, x_i^2 + \frac{r_2}{2}\right]\right)$,其中,$r_1$ 与 r_2 随机取至一个正区间 $[\alpha,\beta]$,本实验中分别取区间为[1,8],[1,16],[1,24],[1,32],[1,40]。这样就得到了一个有 350 个数据点,每个数据点有两个区间属性的数据集。该区间数据集分成三类:1 ~ 150 为一类,151 ~ 300 为一类,301 ~ 350 为一类。由于数据集是随机产生的,所以对每个不同的 $[\alpha,\beta]$,我们让 K 均值和相似性传播分别运行 100 次,取平均 CR。

表 3　K-均值与 APCIMD 在人工数据集上的 CR 比较

$[\alpha,\beta]$	K-均值聚类	APCIMD 聚类
[1,8]	0.2892	0.3239
[1,16]	0.2868	0.3295
[1,24]	0.2606	0.3380
[1,32]	0.2799	0.3594
[1,40]	0.2872	0.3435

从表 3 中的运行结果可以看出,对于人工合成的区间数据,相对于基于欧氏距离的 K-均值的聚类算法,本文的聚类算法也表现出了它的优越性。

4 结语

针对符号数据中的区间数据,提出了基于相互距离的相似性传播聚类算法。跟传统的很多聚类算法比较,本算法的优点是聚类前不需要指定聚类数目和初始化聚类中心,只要给出相似性矩阵,算法就能够通过迭代产生各个聚类以及聚类中心,而且产生的聚类中心是数据集中存在的点,这样的中心我们可以称之为类的代表点。同时通过实验也证明了本文提出的相互距离比传统的欧氏距离更适合于度量区间数据。

(下转第 1493 页)

据后,所有相关的 C 结构以及其他实际数据(如 C 结构内指针所指向的字节串)都在独立内存区申请,L7 解码完成之后所有 C 结构已经填充好值,L7 可以进行进一步处理。如果 C 结构需要传递给设备应用层,设备应用层需要将此独立内存区释放;否则应由 L7 进行释放。

2.2.3 说明

由于编解码所占用的时间在 DSRC 中较多,"多次申请,一次释放"的方式提高了编解码的效率,解决了编解码内存使用的诸多问题。由于我们的应用是单任务的,所以内存为静态划分。在多任务环境下,编解码所需内存应动态分配,申请释放都在此动态分配的内存区上进行。另外,如果编解码的数据量较大且内存容量充足,难以指定一次为编解码分配多大内存区,可以考虑使用链表将各内存块串联,进行动态管理。

在小设备中,资源相对紧张,内存使用相对比较频繁,在可以预知模块内存使用上限的情况下,使用单独的内存管理方式往往对整体性能的提高有较大作用。以上所述的内存管理机制比较通用,适用于多种应用场景。

2.3 其他辅助函数

其他辅助函数功能包括:根据约束的上界和下界计算表示此字段的最小位数(这在 PER 编码中随处可见);提供编码运行时的缓冲区溢出检查功能;提供解码时向前探测所解码字段的长度以防止解码时超出所解码数据的缓冲区等。这些辅助函数能够简化编解码的错误处理,方便库函数的编写。

2.4 其他

由于各函数为嵌套调用,在设计各函数(如基本类型的编解码函数、辅助函数等)时应该从空间和时间两方面权衡。如果从节约代码空间考虑,应尽量将各个可能多处调用的功能设计为函数;若从时间考虑,应尽量设计为宏。某基本类型的函数可能被多处调用,如果将其设计为宏,产生的代码量将非常可观,小设备上实现时应加以注意。

3 运行结果

本文的实现已稳定运行于实际环境中,在此以 BST 和 VST 的编解码为例简单说明其运行性能。

运行环境:采用 STR712F 为 CPU,其自带 64 KB SRAM,设置其主频为 48 MHz;无操作系统支持,即单线程独占 CPU 运行。通过 Multi ICE 及 AXD 调试器运行编解码程序,通过计算编解码开始与结束时的时钟差值来计算编解码所需要的时间。

运行结果:解码 20 Byte 符合标准格式的 BST 平均开销为 150 μs,编码 90 Byte 符合标准格式的 VST 时平均开销为 350 μs,很好地符合 DSRC 中实时性要求。

4 结语

本文给出了 DSRC 中 ASN.1 模块的实现框架,提出了适合小设备应用的编解码内存管理机制,对其做出了实现。经验证,具有较理想的编解码效率。同时,本文所提出的 ASN.1 PER 编解码方法以及内存管理机制比较通用,适用于其他类型的编解码规则以及应用场景。

参考文献:

[1] 刘富强 项雪琴,邱冬. 车载通信 DSRC 技术和通信机制研究[J]. 上海汽车, 2007(8): 35-39.

[2] PANAYAPPAN R, TRIVEDI J M, STUDER A, *et al*. VANET-based approach for parking space availability [C]// International Conference on Mobile Computing and Networking: Proceedings of the fourth ACM International Workshop on Vehicular Ad Hoc Networks. New York: ACM Press, 2007: 75-76.

[3] GB/T 16263.2-2006, 信息技术 ASN.1 编码规则,第 2 部分:紧缩编码规则(PER)规范[S]. 2006.

[4] GB/T 20851.3-2007, 电子收费专用短程通信——应用层[S]. 2007.

[5] GB/T 20851.4-2007, 电子收费专用短程通信——设备应用[S]. 2007.

[6] ISO/IEC 8825-2, ITU-T-1997, Recommendation X.691 ASN.1 encoding rules specification of packed encoding rules [S]. 1997.

[7] ISO/IEC 8824-1, ITU-T-1997, Recommendation X.680 Abstract Syntax Notation (ASN.1) specification of basic notation [S]. 1997.

[8] DUBUISSON O. ASN.1 communication between heterogeneous systems [M]. San Francisco: Morgan Kaufmann, 2000.

(上接第 1443 页)

但相似性矩阵中 $s(j,j)$ 的确定是个有待解决的问题,怎样方便设定 $s(j,j)$,同时又能得到恰当的聚类数目,而且一开始各个数据点成为中心的可能性不一样时,又应如何处理才能不产生错误的中心,以及如何自动选择防震荡因子 λ,这些都将是我们接下来要研究的工作。

参考文献:

[1] RUI XU, WUNSCH D II. Survey of clustering algorithms [J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645-678.

[2] de CARVALHO F A T. A fuzzy clustering algorithm for symbolic interval data based on a single adaptive Euclidean distance [C]// ICONIP 2006, Part III, LNCS 4234. Berlin: Springer-Verlag, 2006: 1012-1021.

[3] GOWDA K C, DIDAY E. Symbolic clustering using a new dissimilarity measure [J]. Pattern Recognition, 1991, 24(6): 567-578.

[4] GOWDA K C, DIDAY E. Symbolic clustering using a new similarity measure [J]. IEEE Transactions on Systems, Man and Cybernetics, 1992, 22(2): 368-378.

[5] ICHINO M, YAGUCHI H. Generalized Minkowski metrics for mixed feature type data analysis [J]. IEEE Transactions on Systems, Man and Cybernetics, 1994, 24(4): 698-708.

[6] de CARVALHO F A T, de SOUZA R M C R, CHAVENT M, *et al*. Adaptive Hausdorff distances and dynamic clustering of symbolic interval data [J]. Pattern Recognition Letters, 2006, 27(3): 167-179.

[7] FREY B J, DUECK D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972-976.

[8] 洪志令. 模糊聚类中判别聚类有效性的新指标[J]. 计算机科学, 2004, 31(10): 121-125.

[9] HUBERT L, ARABIE P. Comparing partitions [J]. Classification, 1985, 2(1): 193-218.

[10] EL-SONBARY Y, ISMAIL M A. Fuzzy clustering for symbolic data [J]. IEEE Transactions on Fuzzy System, 1998, 6(2): 195-204.

[11] GURU D S, KIRANAGI B B. Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic [J]. Patterns Recognition, 2005, 38(1): 151-156.

适用于区间数据的基于相互距离的相似性传播聚类

作者： 谢信喜， 王士同， XIE Xin-xi， WANG Shi-tong
作者单位： 江南大学,信息工程学院,江苏,无锡,214122
刊名： 计算机应用 ISTIC PKU
英文刊名： JOURNAL OF COMPUTER APPLICATIONS
年，卷(期)： 2008,28(6)
被引用次数： 3次

参考文献(11条)

1. RUI XU;WUNSCH D II Survey of clustering algorithms[外文期刊] 2005(03)

2. de CARVALHO F A T A fuzzy clustering algorithm for symbolic interval data based on a single adaptive Euclidean distance 2006

3. GOWDA K C;DIDAY E Symbolic clustering using a new dissimilarity measure 1991(06)

4. GOWDA K C;DIDAY E Symbolic clustering using a new similarity measure 1992(02)

5. ICHINO M;YAGUCHI H Generalized Minkowski metrics for mixed feature type data analysis 1994(04)

6. de CARVALHO F A T;de SOUZA R M C R;CHAVENT M Adaptive Hausdorff distances and dynamic clustering of symbolic interval data[外文期刊] 2006(03)

7. FREY B J;DUECK D Clustering by passing messages between data points[外文期刊] 2007(5814)

8. 洪志令 模糊聚类中判别聚类有效性的新指标[期刊论文]-计算机科学 2004(10)

9. HUBERT L;ARABIE P Comparing partitions 1985(01)

10. EL-SONBARY Y;ISMAIL M A Fuzzy clustering for symbolic data[外文期刊] 1998(02)

11. GURU D S;KIRANAGI B B Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic[外文期刊] 2005(01)

本文读者也读过(10条)

1. 区间数据包络分析模型的求解及其有效性[期刊论文]-海军航空工程学院学报2006,21(4)

2. Kazuhiro Hayama Classi.cation of GravitationalWaves from Supernova Core Collapses Using Afinity Propagation [会议论文]-2008

3. 陈东升.陈明璨.张利利.CHEN Dong-sheng.CHEN Ming-can.ZHANG Li-li 基于区间值的聚类方法研究及应用[期刊论文]-数学的实践与认识2010,40(3)

4. 郭景峰.马鑫.代军丽.GUO Jing-feng.MA Xin.DAI Jun-li 基于文本-链接模型和近邻传播算法的网页聚类[期刊论文]-计算机应用研究2010,27(4)

5. Jian Liu.Na Wang Detecting Community Structure of Complez Networks by Affinity Propagation[会议论文]-2009

6. Shuzhong Yang.Siwei Luo Community Detection Based on Adaptive Kernel Affinity Propagation[会议论文]-2009

7. 张伟斌.刘文江.ZHANG Wei-bin.LIU Wen-jiang 区间型数据的模糊c均值聚类算法[期刊论文]-计算机工程2008,34(11)

8. 谢信喜.王士同.XIE Xin-xi.WANG Shi-tong 带特征权重的混合特征模糊C均值算法[期刊论文]-计算机工程与应用2008,44(6)

9. WAN Jie 基于Affinity Propagation 聚类方法的图像检索技术在数字图书馆中的应用[期刊论文]-计算机与现代化2008(8)

10. Yan Zhu.Jian Yu.Caiyan Jia Initializing K-means Clustering Using Affinity Propagation[会议论文]-2009

引证文献(3条)

1.张仁彦.赵洪亮.卢晓.曹茂永 基于相似性传播聚类的灰度图像分割[期刊论文]-海军工程大学学报 2009(3)

2.吴海华.李绍滋.林达真.柯逍.曹冬林 基于新型聚类算法Increase K-Means的Blog相似度分析[期刊论文]-厦门大学学报（自然科学版） 2009(2)

3.刘晓勇.付辉 一种快速AP聚类算法[期刊论文]-山东大学学报：工学版 2011(4)