基于维度属性距离的混合属性近邻传播聚类算法

黄德才 钱潮恺

(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 针对近邻传播聚类算法不能处理混合属性数据集的问题,提出了一种新的距离度量测度,并将其应用到近邻传播聚类算法中,提出了一种基于维度属性距离的混合属性近邻传播聚类算法。与传统聚类算法不同的是,该算法不需要计算虚拟的中心点,同时考虑了数据集整体分布对聚类结果的影响。将算法在UCI数据库的2个混合属性数据集上进行验证,同时对比了经典的K-Prototypes算法以及K-Modes算法。实验结果表明,改进后的算法具有更好的聚类质量以及执行效率,算法的优越性得到了验证。

关键词 属性距离,混合属性,近邻传播,聚类

中图法分类号 TP391.4 **文献标识码** A

Mixed Data Affinity Propagation Clustering Algorithm Based on Dimensional Attribute Distance

HUANG De-cai QIAN Chao-kai

(College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract A new distance measurement was raised because the affinity propagation cannot cluster mixed data sets. And this distance measurement was successfully applied into affinity propagation clustering algorithm. This new algorithm doesn't need to calculate the virtual cluster center points, and also considers the effect of diversity of whole data set. This algorithm was validated through two UCI data sets. And the clustering performance is better than K-Prototypes and K-Modes in both clustering entropy and execution efficiency.

Keywords Attribute distance, Mixed attributes, Affinity propagation, Clustering

1 引言

聚类分析[1]是数据挖掘领域的一个热门的研究课题,它在机器学习中被称为无监督学习。简单说来,聚类就是把相似的东西分到同一个类中,而并不关心某一类是什么。因此,一个聚类算法通常只要知道如何计算数据点之间的距离就可以开始工作了。现在,聚类分析广泛应用于金融数据的分类、信息检索、图像自动检测、卫星图片分析等。

K-Means算法[2]是一种经典的聚类算法,它因算法思想简单、易于实现、执行效率高等优点而应用于多个商业领域。但是,它仅限于数值型数据聚类问题。然而,随着社会的发展,许多应用领域会产生大量既包含数值型数据又包含分类型数据的混合属性数据[3]。所以,如何处理混合属性数据的聚类已经逐渐成为一个重要的研究课题。

Huang等人在1997年提出了一种能够处理混合属性数据的聚类算法,即K-Prototypes算法[4]。针对分类属性部分的数据,该算法用一种匹配的差异度来描述数据点之间的距离,并且采用模式的概念来替代传统的聚类中心;而对于数值属性部分的数据,则采用传统的欧氏距离来计算;同时用参数γ来控制数值型数据和分类型数据在聚类时的权重。虽然该算法的距离度量公式被广泛使用,但是它并不能真正地反映微簇内部的相似性,且由于聚类时虚拟中心点的使用而导致了空簇的产生。到目前为止,许多研究者在混合属性聚类方面进行了研究。比如,Chatzis提出的基于相对熵的模糊C均值聚类算法[5]及白天提出来的一种全局K-Prototypes聚类算法[6]。但是,现有的混合属性聚类算法仍然会用虚拟的聚类中心,因此,依然会存在产生空簇的可能。

众所周知,数值属性具有天然的几何性质,我们可以直接对其进行加减乘除的数学运算。但是,分类属性数据的特殊性,导致我们无法直接对其进行数学运算,且传统的距离计算公式依赖于虚拟的中心点,该中心点的某个属性并不一定能完全代表该维度下数据值的整体分布情况,这就会使得聚类的质量受到影响。本文针对上述问题,引出了新的距离计算公式,并将其应用于近邻传播聚类算法,提出了一种基于维度属性距离的混合属性近邻传播聚类算法(Mixed Data Affinity Propagation Clustering Algorithm Based on Dimensional Attribute Distance, APDA)。通过仿真实验对比,表明APDA算法在聚类纯度以及聚类效率方面较原有的混合属性聚类算法有一定的优越性。

2 近邻传播算法的相关研究

2.1 近邻传播算法(Affinity Propagation, AP)

经典的K-Means等聚类算法对初始聚类中心点的选择非常敏感。为了解决这一问题,Frey和Dueck提出了一种新

本文受水利部公益性行业科研专项(201401044)资助。

黄德才(1958-),男,博士,教授,主要研究方向为数据挖掘、人工智能,E-mail:hdc@zjut.edu.cn;钱潮恺(1990-),男,硕士生,主要研究方向为数据挖掘。

的聚类算法——近邻传播聚类算法[7]，它是一种基于近邻信息传播的聚类算法。与许多经典的聚类算法相比，最为突出的不同点是，它将数据集中所有的数据点都作为初始类代表点，而不需要指定初始的类中心点；同时，该算法对于数据集所生成的相似性矩阵没有对称性要求，并且在处理大规模数据时能够进行快速运算，因此性能更好[8-10]。

AP算法不必事先定义聚类个数，在迭代的过程中，算法会不停地搜索合适的代表点即聚类中心，从而识别出可以作为聚类中心的数据点，使得所有数据点到最近的聚类代表点的相似度之和达到最大。该算法先将数据集中的所有 N 个数据点均视作候选的聚类中心，然后为每一个数据点计算与其他 $N-1$ 个数据点之间的相似性程度信息并保存到矩阵 $s(i,k)$ 中来作为算法开始的输入参数。同时，算法需要每个数据点与其他数据点之间建立起一种吸引程度的信息，该吸引程度信息由如下两个矩阵组成。

吸引度(Responsibility)矩阵：$r(i,k)$ 是从数据点 X_k 指向数据点 X_i，它用于描述 X_k 适合作为 X_i 的聚类中心的程度；

归属度(Availability)矩阵：$a(i,k)$ 是从数据点 X_i 指向数据点 X_k，它用于描述 X_i 选择 X_k 作为其聚类中心的合适程度。

AP算法通过不停地从数据点中获取 $r(i,k)$ 和 $a(i,k)$，来确定合适的聚类中心。初始化时，$a(i,k)=0$，则 $r(i,k)$ 与 $a(i,k)$ 的更新公式为：

$$r(i,k)=s(i,k)-\max_{k'\neq k}\{a(i,k')+s(i,k')\} \tag{1}$$

$$a(i,k)=\begin{cases}\min(0,r(k,k)+\sum_{i'\neq i}\max(0,r(i',k))), & i\neq k\\ \sum_{i'\neq i}\max(0,r(i',k)), & i=k\end{cases} \tag{2}$$

AP算法主要根据上述两个公式不停地迭代来更新证据，从而确定聚类中心。上述两个矩阵只需在数据对之间传递信息，假设数据点 X_k 需要作为 X_i 的聚类中心，只要满足 $r(i,k)+a(i,k)$ 的值达到最大变化即可。当局部 $r(i,k)+a(i,k)$ 的值不再变化时，消息传递将会停止。另外，迭代更新的速度可以通过调节阻尼系数 λ 来实现。AP算法在迭代过程中不停地更新每个数据点的吸引度信息和归属度信息，从而产生 m 个高质量的聚类中心，然后将剩余数据点分配到最近的聚类中心去。

2.2 近邻传播算法的扩展

AP算法采用欧氏距离来计算各数据点之间的相似性，该度量方法是以数据点之间的绝对距离来作为数据点之间的相似程度，那么数据集的形状和密度对聚类结果会有较大的影响。Michele 等人提出了一种软约束的近邻传播聚类算法[11]，它使得算法能够处理环形的数据集。该算法通过引入一个软参数 β，使得算法在一种松弛的条件下聚类，且允许聚类中心在该类之外。该算法虽然能够解决环形数据集的聚类，但是仍然不能完全解决其他非凸形数据集的聚类问题。为了使 AP 算法能应用于大规模数据集，提升算法的处理效率，刘晓楠等人提出了面向大规模数据的分层近邻传播聚类算法[12]。该算法通过对大规模数据集进行划分，用 AP 聚类算法对划分的子集进行聚类，然后分别选出各个子集的聚类中心，并对所有子集的聚类中心执行 AP 算法得到全局的聚类中心，以此提升 AP 算法对大规模数据聚类的处理效率，但是该算法无法对数据集中的微型类进行有效聚类。Cyril 等人在文献[13]中，首次将 BP 神经网络与 AP 聚类算法相结合，提出了一种层次的 AP 聚类算法(Hierarchical Affinity Propagation，HAP)。该算法可以在略低于甚至不低于原始 AP 算法聚类精度的情况下，大大提高算法的执行速度。Zhang 等在顶级会议 PKDD (Principles and Practice of Knowledge Discovery in Databases)上针对数据流首次提出了 StrAP (Stream data streaming with affinity propagation)算法[14]。但是该算法存在聚类质量不高、处理离群点能力较差以及不能有效检测数据流变化等问题，因此，张建朋等人在 StrAP 算法的基础上提出一种密度与近邻传播融合的高效数据流聚类算法 StrDenAP[15]。为了提高 AP 聚类算法的准确性，王开军等人提出了一种自适应仿射传播聚类算法[16]，并讨论了如何设置合理的迭代次数和阻尼因子等参数。

上述算法均局限于数值属性数据的聚类问题，无法适用于混合属性数据集。本文将 AP 算法进行扩展，使之能够对混合属性数据进行聚类。

3 基于维度属性距离的混合属性 AP 聚类算法模型

如前文所述，传统的混合属性聚类算法需要定义一个虚拟的簇中心，这种方式会导致聚类精度不高。因此，本节提出了一种无中心的聚类算法。由于 AP 聚类算法需要在相似度矩阵的基础上进行聚类，因此这个相似度矩阵就显得尤为重要。然而原始的 AP 算法采用的是欧氏距离平方的负数来作为相似度，这就导致无法采用该算法对混合属性数据集进行聚类。

3.1 维度属性距离计算公式

设数据集 $D=\{X_1,X_2,\cdots,X_n\}$ 表示由 n 个数据点组成的数据集合，对于每个数据点 $X_i=\{A_1,A_2,\cdots,A_r\}$，每个属性 A_j 的值域由离散变量表示的分类型或由连续变量表示的数值型组成。为方便起见，设前 m 维的值为分类属性，从 $m+1$ 维开始到 r 维的值为数值属性，其中分类属性的值域 $Domain(A_j)=\{a_j^1,a_j^2,\cdots,a_j^h\}$，$h$ 为第 j 个属性可以取到的类别数目。

定义 1 对于任意分类属性维度 A_j，设 a_j^x,a_j^y 属于 $Domain(A_j)$，那么 a_j^x 与 a_j^y 基于 A_j 的内部维度属性距离定义为：

$$d\_Inner_{A_j}(a_j^x,a_j^y)=\begin{cases}0, & a_j^x=a_j^y\\ 1, & a_j^x\neq a_j^y\end{cases} \tag{3}$$

定义 2 对于任意分类属性维度 A_k，设 a_j^x,a_j^y 属于 $Domain(A_j)$，那么 a_j^x 与 a_j^y 基于 $A_k(j\neq k)$ 的外部维度属性距离定义为：

$$d\_Outer_{A_k}(a_j^x,a_j^y)=\frac{1}{|D|}\sum_{p\in D}|S(p,A_j)-S(p,A_k)| \tag{4}$$

其中，D 为任意一个混合属性数据集合，$S(p,A_i)=\{p\mid\varphi(p,A_i)=a_j^x,i\neq j\}$，$S(p,A_k)=\{p\mid\varphi(p,A_k)=a_j^y,j\neq k\}$，$p$ 为分类属性维度 A_i 的属性值，$S(p,A_i)$ 表示分类属性值 p 与 a_j^x 相同的值的集合。

定义 3 对于任意分类属性维度 A_j，设 a_j^x,a_j^y 属于 $Domain(A_j)$，数据点 X_i 和 X_j 之间分类属性关于 A_j 的维度属性距离可以定义为：

$$d_{A_j}(a_j^x,a_j^y)=d\_Inner_{A_j}(a_j^x,a_j^y)+\sum_{k=1}^{m}d\_Outer_{A_k}(a_j^x,a_j^y),\quad k\neq j \tag{5}$$

定义 4 对于任意分类属性维度 A_t，设 a_t^x, a_t^y 属于 $Domain(A_t)$，数据点 X_i 和 X_j 之间分类属性部分的维度属性距离可以定义为：

$$dc(X_i,X_j)=\frac{\sum_{t=1}^{m} d_{A_t}(a_t^x,a_t^y)}{\sum_{l=1}^{m} m} \tag{6}$$

其中，l 和 t 均为分类属性的维度。

定义 5 数据点 X_i 和 X_j 之间的维度属性距离公式可以定义为：

$$dist(X_i,X_j)=\sum_{d=1}^{m}(X_{id}-X_{jd})^2+\gamma * dc(X_i,X_j) \tag{7}$$

接下来用一个例子说明上述距离的计算过程。如表 1 所列，该表列出 UCI 数据库中 Census Income 的部分数据。为了使数据能够更加清楚地展示，对属性类别与分类属性值采用简化的字符表示。

表 1 Census Income 数据集

数据点	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	C
x_1	39	13	40	N	SG	B	AC	NF	N
x_2	50	13	13	M	SE	B	EM	H	N
x_3	38	7	40	D	P	H	HC	NF	N
x_4	26	13	40	D	SG	B	PS	H	Y
x_5	24	13	35	N	P	B	AC	OC	N
x_6	35	3	40	M	FG	N	FF	H	N
x_7	50	13	55	D	FG	B	EM	NF	Y
x_8	42	23	45	M	P	D	PS	H	Y
x_9	47	13	40	M	P	B	EM	W	Y
x_{10}	38	14	40	M	FG	M	PS	H	Y

表 1 中总共包含 10 个数据点，其中 A_1、A_2 和 A_3 为 3 个数值类型值的属性，A_4、A_5、A_6、A_7 以及 A_8 是 5 个分类属性，最后一列的 C 代表数据点所属的类，收入超过 5 万的用 Y 表示，否则用 N 表示。如果采用 K-Prototypes 算法中的公式，取分类属性的权重 $\gamma=8.2$，计算 X_1 与 X_4 之间的距离以及 X_1 与 X_5 之间的距离，那么 $d(X_1,X_4)=37.6$，$d(X_1,X_5)=37.6$，此时，X_4 到 X_1 的距离与 X_5 到 X_1 的距离相等。这种情况下就无法准确判断 X_4 与 X_5 的真正归属。事实上，这种情况不应存在，根据表中的数据可以看到 X_1 与 X_5 应该更为接近。因此，采用这种距离计算方式来聚类，其聚类精度不会很高。采用新的距离计算方式可以有效地避免这种情况，其计算过程如下：

首先，计算分类属性部分的距离。数据点 X_1 和 X_4 在 A_4 这个属性下的内部属性距离 $d\_Inner_{A4}(N,M)=1$。接着计算基于 A_4 的外部属性距离，由于剩下有 4 个维度的属性，那么基于 A_4 的外部属性距离就由 4 个部分组成。

$d\_Outer_{A5}(N,M)=(2\times|1/2-1/2|+1\times|0-1|+4\times|1/4-2/4|+3\times|0-2/3|)/10=0.4$；

$d\_Outer_{A6}(N,M)=(6\times|2/6-3/6|+1\times|0-0|+1\times|0-1|+1\times|0-1|+1\times|0-1|)/10=0.4$；

$d\_Outer_{A7}(N,M)=(2\times|2/2-0|+3\times|0-2/3|+1\times|0-0|+3\times|0-3/3|+1\times|0-1|)/10=0.8$；

$d\_Outer_{A8}(N,M)=(3\times|1/3-0|+5\times|0-5/5|+1\times|1-0|+1\times|1-0|)/10=0.8$；

$d_{A4}(N,M)=(1+0.4+0.4+0.8+0.8)/5=0.68$。

同理，可以计算出 $d_{A5}(SG,SG)=0$，$d_{A6}(B,B)=0$，$d_{A7}(AC,PS)=0.46$，$d_{A8}(NF,H)=0.6$。然后，可以计算出分类属性部分的距离 $dc(X_1,X_4)=0.68+0+0+0.46+0.6=1.74$，加上数值部分的属性距离可以得到数据点 X_1 与 X_4 之间的距离 $d(X_1,X_4)=8.2\times1.74+11=27.268$。

同样地，可以计算得到数据点 X_1 与 X_5 之间的距离 $d(X_1,X_5)=27.1$。可以看到 $d(X_1,X_4)>d(X_1,X_5)$，采用新的距离公式使精度得到了提高。

众所周知，两个数据点的距离越远，其相似度就越小，因此可以用距离的倒数作为两个数据点之间的相似度来作为聚类算法所需要的参数。

3.2 算法描述以及步骤

描述：基于加权维度相似度的混合属性聚类算法

输入：数据集 D，$max\_iterations$

输出：最优的聚类中心集合，数据集的聚类结果

步骤：

Step1 初始化 $a(i,j)=0$；

Step2 采用改进的距离公式计算任意两个数据点之间的相似性，并将此结果保存到矩阵 sim 中；

Step3 根据式(1)和式(2)分别计算数据点之间的 $Responsibility$ 值和 $Availability$ 值；

Step4 利用 AP 算法的原理更新 $r(i,j)$ 和 $a(i,j)$；

Step5 当算法的迭代次数超过了设定值，或者当聚类中心点在一定次数的迭代后仍未改变时，迭代终止，否则跳转至 Step3。

4 仿真实验

4.1 仿真实验条件

为了评估 APDA 聚类算法的效果与性能，在 UCI 中选取 zoo 和 Heart disease 这两个数据集来进行仿真实验。这两个数据集的具体情况如表 2 所列。

表 2 实验数据集

数据集	数据点个数	维度	数值型维度	分类型维度	类数
zoo	101	16	1	15	7
Heart disease	270	13	6	7	2

这两个数据集分别有 16 个维度和 13 个维度。聚类算法实验的硬件环境为 Windows 7 操作系统、Intel(R) Core(TM) i3-2330M 2.20GHz CPU、4GB 内存、750GB 硬盘，实验算法用 MATLAB 实现。仿真实验所采用的 zoo 数据集和 Heart disease 数据集来自 UCI 数据库。UCI 数据库是加州大学欧文分校提出来的适用于机器学习的数据库。仿真实验包括两部分，第一部分将改进的聚类算法与经典的混合属性聚类算法进行对比，第二部分是对算法的效率进行对比。

4.2 仿真实验结果分析

K-Modes 和 K-Prototypes 算法是解决混合属性数据聚类的经典算法，而且自从混合属性聚类提出来之后，很多的研究工作也都是基于这两种算法的改进。因此，本文将它们作为基准算法进行比较，以验证 APDA 算法在聚类效果上的改进。为评价各个算法的聚类效果，本文采用聚类熵来进行评价。针对每一个聚类簇 C_i，其熵(Entropy)的定义如下：

$$Entropy(C_i)=\frac{1}{\log k}\sum_{j=1}^{k}\frac{|C_i\cap P_j|}{|C_i|}\log\left(\frac{|C_i|}{|C_i\cap P_j|}\right) \tag{8}$$

(下转第 71 页)

算机工程与设计,2010,31(21):4679-4681
[8] 雷阳,雷英杰,冯有钱,等. 基于直觉模糊推理的目标识别方法[J]. 控制与决策,2011,26:1163-1174
[9] 林娟,米据生,解滨. 粗糙集的两种相似性度量[J]. 计算机科学,2015,42(6):97-100
[10] 范成礼,邢清华,邹志刚,等. 基于直觉模糊粗糙集相似度的多属性决策方法[J]. 计算机工程与应用,2014,50(7):121-124
[11] 张云霞,崔晓松,邹丽. 一种基于十八元语言值模糊相似矩阵的聚类方法[J]. 山东大学学报,2013,43(1):1-7
[12] 邹丽,谭雪微,张云霞. 语言真值直觉模糊逻辑的知识推理[J]. 计算机科学,2014,41(1):134-137
[13] Liu De-shan, Yin Ming'e, Zou Li. Reasoning rules of linguistic truth-valued intuitionistic fuzzy propositional logic system[J]. CEA,2011,47(33):62-64
[14] 邹丽. 基于语言真值格蕴涵代数的格值命题逻辑及其归结自动推理研究[D]. 成都:西南交通大学,2010
[15] 徐江,潘正华. 金融投资决策中的模糊知识及其不同否定的表示与推理[J]. 计算机应用与软件,2011,28(3):38-40
[16] 田野,陈东锋,雷英杰. 基于直觉模糊相似度量的近似推理方法[J]. 空军工程大学学报(自然科学版),2007,8(6):81-82

(上接第 57 页)

其中,k 为数据集中的准确的聚类数目,P_j 表示数据集中第 j 个标记簇。因此,k 和 P_j 都是预先就知道的。聚类熵是一个 0 到 1 的值,0 表示微簇 C_i 由一个 P_j 完整包含了;1 则表示微簇 C_i 均匀地包括了所有的人工标记簇,这种情况就是一个很差的结果。因此,聚类熵的值应该越接近于 0 越好。不同聚类算法的聚类熵值如表 3 所列。

表 3 不同聚类算法的聚类熵值

数据集	各聚类算法及相应的聚类熵值		
	K-Modes	K-Prototypes	APDA
zoo	0.161	0.15	0.0705
Heart Disease	0.283	0.22	0.201

从表 3 可以看出,3 个聚类算法的聚类熵值有所不同,本文提出的 APDA 算法的聚类熵值相比 K-Modes 和 K-Prototypes 算法在 zoo 数据集上提升了 50%以上,在 Heart Disease 数据集上提升了 10%以上,因此,APAD 算法更好地完成了数据聚类的任务。由于 K-Modes 和 K-Prototypes 算法在不同程度上忽略了分类属性值的整体分布情况,使得前两个算法的聚类结果没有本文的算法好。

另一方面,仿真实验对比了不同算法的运行时间。由于 K-Prototypes 算法的时间复杂度为 $O(nkt)$,其聚类速度较快,能够快速处理大型数据集,而且在执行效率方面,相对于许多目前的混合属性数据聚类算法有优势。因此,本文的算法选择 K-Prototypes 算法进行对比。两个算法的执行时间如表 4 所列。

表 4 不同算法的执行时间对比(单位:s)

数据集 \ 算法	K-Prototypes	APDA
zoo	2	1.5
Heart Disease	5.8	5.7

从表 4 中的数据可以看出,对于 zoo 数据集,APDA 算法在执行效率方面比 K-Prototypes 算法稍快,但是在 Heart Disease 数据集上两者的效率几乎一致,其稍稍好于 K-Prototypes 算法,这是由于 K-Prototypes 算法需要随机地选择 k 个数据点作为初始中心点,这个过程对于聚类算法的正确率和效率方面都有一定的影响。

结束语 本文提出了一种基于属性距离的混合属性近邻传播聚类算法,它改进了传统混合属性数据点之间的距离公式,消除了虚拟中心点对聚类结果带来的影响,提高了聚类分辨率,同时将近邻传播聚类算法由原来仅适用于数值属性聚类的数据集,扩展到了适用于混合属性数据集的聚类问题中。将本文算法与经典的混合属性聚类算法进行仿真对比,对不同聚类算法的聚类熵这一指标值进行比较,证明了基于维度属性距离的混合属性近邻传播算法确实较传统的混合属性聚类算法在聚类效果方面有了一定的提升。

参 考 文 献

[1] Tan P N, Steinbach M, Kumar V. 数据挖掘导论[M]. 范明,范宏建,等译. 北京:人民邮电出版社,2011
[2] Kaufan L, Rousseeuw P J. Finding Groups in Data: An Introduction to Cluster Analysis[M]. New York: John Wiley&Sons, 1990
[3] 黄德才,沈仙桥,陆亿红. 混合属性数据流的二重 k 近邻聚类算法[J]. 计算机科学,2013,40(10):226-230
[4] Huang Zhe-xue. Clustering Large Data Sets with Mixed Numeric and Categorical Values[C]// Proceedings of PAKDD'97. Singapore, World Scientific, 1997:21-35
[5] Chatzis S P. A Fuzzy C-Means-Type Algorithm for Clustering of Data with Mixed Numeric and Categorical Attributes Employing a Probabilistic Dissimilarity Functional[J]. Expert Systems with Applications, 2011, 38(7): 8684-8689
[6] 白天,冀进朝,何加亮,等. 混合属性数据聚类的新方法[J]. 吉林大学学报(工学版),2013,43(1):130-134
[7] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976
[8] Qian Y, Yao F, Jia S. Band selection for hyperspectral imagery using affinity propagation[J]. IET Computer Vision, 2009, 3(4): 213-222
[9] Li G, Guo L, Liu T. Grouping of brain MR images via affinity propagation[C]// IEEE International Symposium on Circuits and Systems, 2009(ISCAS 2009). IEEE, 2009: 2425-2428
[10] Dueck D, Frey B J, Jojic N, et al. Constructing treatment portfolios using affinity propagation[M]// Research in Computational Molecular Biology. Springer Berlin Heidelberg, 2008: 360-371
[11] Sumedha M L, Weigt M. Unsupervised and semi-supervised clustering by message passing: soft-constraint affinity propagation[J]. The European Physical Journal B-Condensed Matter and Complex Systems, 2008, 66(1): 125-135
[12] 刘晓楠,尹美娟,李明涛,等. 面向大规模数据的分层近邻传播聚类算法[J]. 计算机科学,2014,41(3):185-188
[13] Furtlehner C, Sebag M, Zhang X. Scaling analysis of affinity propagation[J]. Physical Review E, 2010, 81(6): 066102
[14] Zhang X, Furtlehner C, Sebag M. Data streaming with affinity propagation[M]// Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2008: 628-643
[15] 张建朋,陈福才,李邵梅,等. 基于密度与近邻传播的数据流聚类算法[J]. 自动化学报,2014,40(2):277-288
[16] 王开军,张军英,李丹,等. 自适应仿射传播聚类[J]. 自动化学报,2007,33(12):1242-1246