ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 4,Jul. 2016,pp. 767－774
DOI:10.16337/j.1004-9037.2016.04.015

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

属性样本同步粒化的 AP 熵加权软子空间聚类算法

朱　红[1]　丁世飞[2]

(1.徐州医科大学医学信息学院,徐州,221005; 2.中国矿业大学计算机科学与技术学院,徐州,221116)

摘　要:仿射传播(Affinity propagation,AP)聚类算法是将所有待聚类对象作为潜在的聚类中心,通过对象之间传递的可靠性和有效性信息找到合适的聚类中心,从而计算出相应的聚类结果,但不适用子空间聚类。将粒度计算引入到仿射传播聚类算法中,提出属性与样本同步粒化的 AP 熵加权软子空间聚类算法(Entropy weighting AP algorithm for subspace clustering based on asynchronous granulation of attributes and samples,EWAP)。EWAP 首先去除冗余属性,然后在每次聚类的迭代过程中修改属性的权重值。在满足一定条件迭代终止时,就会得到构成各兴趣度子空间的属性权重值,从而得到属性集的粒化结果以及相应的子空间聚类结果。理论与实验证明 EWAP 算法既保留了 AP 算法的优点,又克服了该聚类算法不能进行子空间聚类的不足。

关键词:聚类;熵;加权;子空间;AP 聚类

中图分类号:TP181　　**文献标志码**:A

Entropy Weighting AP Algorithm for Subspace Clustering Based on Asynchronous Granulation of Attributes and Samples

Zhu Hong[1], Ding Shifei[2]

(1. School of Medical Information, Xuzhou Medical University, Xuzhou, 221005, China; 2. School of Computer Science and Technology, China University of Mining and Technology, Xuzhou, 221116, China)

Abstract: Affinity propagation (AP) clustering algorithm considers all clustering objects as potential clustering centers, and messages of responsibility and availability are exchanged between objects until a high-quality set of clustering centers and corresponding clusters gradually emerge. But it is not appropriate for subspace clustering. To solve this problem, an entropy weighting AP algorithm for subspace clustering based on asynchronous granulation of attributes and samples (EWAP) is put forward through introducing the idea of granular computing into the affinity propagation clustering method. It removes the redundant attributes first, and then a step of modifying attribute weights is added to the clustering procedure for obtaining the exact weights value. At the end of iteration, the attribute weights of each subspace, an accurate result of attributes granularity and the corresponding clusters will be produced. The theory and practice prove that EWAP preserves the advantages of AP clustering and overcomes its shortage of unsatisfying subspace clustering.

Key words: clustering; entropy; weighting; subspace; AP clustering

---

基金项目:国家自然科学基金(61379101)资助项目;江苏省自然科学基金(BK20130209)资助项目;江苏省高校自然科学基金(14KJB520039)资助项目;江苏省高校优秀中青年教师和校长境外研修计划资助项目。
收稿日期:2015-06-05;**修订日期**:2015-06-30

引 言

聚类算法是人工智能、数据挖掘和机器学习等领域的关键技术之一,在数据处理 、图像分析、电子商务、预测和医疗卫生[1-4]等方面有着非常广泛的应用。但是随着大数据时代的到来,产生了大量不一致数据、混合类型数据、高维稀疏数据、非有益数据和部分值缺失的数据等。典型的聚类算法在对这些数据集进行聚类时遇到了难题。例如在高维稀疏数据中,簇类只存在于部分属性构成的兴趣度子空间中,传统的全维聚类算法却不能搜索到隐含在属性子集上的簇类(这些数据集从全维空间来讲根本不存在簇类)。为此,1998 年 Agrawal 提出了子空间聚类的概念[5]。在高维数据中挖掘可以聚类的属性子集中的簇类过程称为子空间聚类。子空间聚类的任务有两个:(1)发现可以聚类的子空间(属性子集);(2)在相应的子空间上聚类。依据属性对每个子空间的贡献,子空间聚类算法可分为硬子空间聚类算法和软子空间聚类算法。在硬子空间聚类算法中,一个属性必须且只能属于一个子空间,聚类则在这些子空间中进行,属性在每个子空间的权值要么是 0,要么是 1。软子空间聚类是在全维空间对整个数据集聚类,每个兴趣度子空间包含所有属性,但是每个属性在不同的兴趣度子空间被赋予[0,1]不同的权值,属性权值描述了属性与对应子空间之间的关联程度,权值越大说明该属性在这个子空间越重要,与该子空间的关联性也就越强。

依据不同的搜寻策略,子空间聚类又可以分成由底向上的聚类方法和由顶向下的聚类方法。这两种子空间聚类算法都要遍历所有维度的数据。由底向上的子空间聚类算法首先要在每一属性上寻找数据的密集单元,如果可以在两维属性空间合并某些密集单元,就会形成两维属性的密集单元,依此类推,形成 k 维属性的密集单元,最后形成在这 k 维子空间上的簇类。此方法的理论基础是关联规则中的先验性质:如果 $k-1$ 维属性空间是不密集的,那么包含它的 k 维属性空间必定也是不密集的;如果 k 维属性空间是密集的,则它所包含的 $k-1$ 维属性空间也必定是密集的。由顶向下的搜寻方法是首先将数据集分成 n 个簇类,给每个簇类赋予一定的权值,然后依据某种算法更新簇的权值,直到满足一定的终止条件。

依据属性与样本粒化的顺序子空间聚类算法可以分为同步和异步两种。这种分类方法认为子空间实际上是对数据集的全维属性集合粒化的结果,而聚类结果则是对样本或待聚类对象粒化的结果。异步方法是先通过对属性集粒化,找到兴趣度子空间,然后在各子空间内部对样本进行聚类。同步方法则是对属性和样本同时粒化,在搜索子空间的同时完成对子空间中样本的聚类。

Frey 等于 2007 年提出了仿射传播(Affinity propagation,AP)聚类算法[6],具有不需事先确定聚类中心与类数等优点,在搜索最优航线、识别基因外显子、识别手写的邮政编码和人脸识别等方面应用效果良好[7-10],但该算法所有属性在聚类过程中同等重要。本文提出的属性样本同步粒化的 AP 熵加权软子空间聚类算法,将属性权重矩阵引入到 AP 算法中,在每次聚类的迭代过程中修改属性的权重值。满足一定条件迭代终止时,就会得到构成各兴趣度子空间的属性权重值,从而得到属性集的粒化结果以及相应的聚类结果。

1 K-Means 熵加权软子空间聚类算法

1.1 W-K-Means 算法[11]

在软子空间聚类算法中使用权值来描述属性对每个子空间的贡献,在此基础上实现子空间的聚类。常用方法是在传统聚类算法的迭代过程中增加属性权值的计算,从而获取不同子空间属性权值的集合。

典型算法有:W-K-means [11],SCAD[12] 和 EWKM[13]。

属性权值对聚类结果的影响如图 1 所示。数据集有 3 个属性 a_1,a_2,a_3,在子空间 a_1,a_2 上的聚类如图 1(a),在子空间 a_1, a_3 上的聚类如图 1(b),在子空间 a_2,a_3 上的聚类如图 1(c)。可见在 a_1, a_3 和 a_2, a_3 上不能发现有效的聚类结果,数据集的簇类只存在于子空间 a_1,a_2 中。但若对 3 个属性赋权值 0.48, 0.45 和 0.07,然后在全维属性集 a_1,a_2 和 a_3 上聚类得到结果如图 1(d)。可见如果通过计算得到 a_1,a_2 和 a_3 的权值为 0.48,0.45 和 0.07,在此基础上聚类可以得到正确的聚类结果。这说明 a_1,a_2 在聚类中起重要作用,a_3 的作用则可以忽略,发现的子空间就是 a_1,a_2。这样通过计算属性的权值就可以发现子空间。

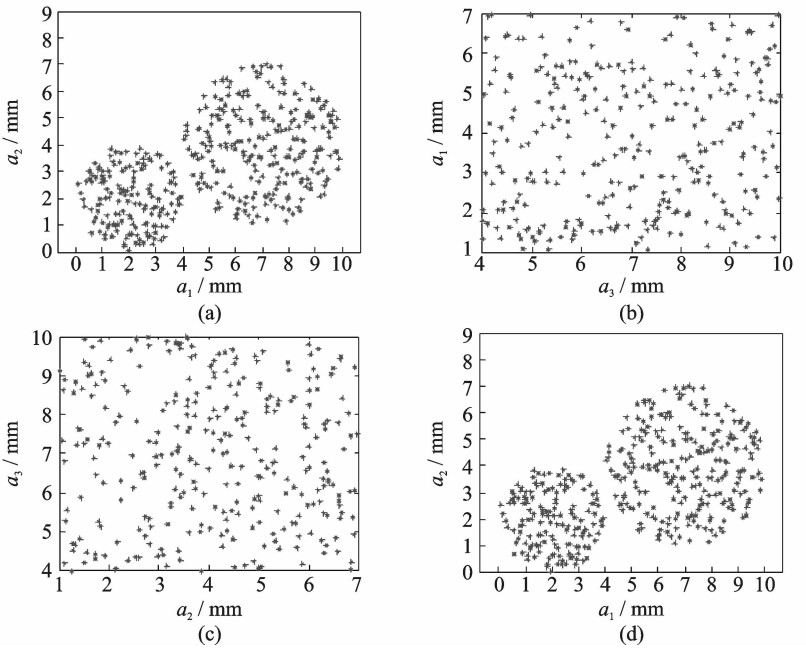

图 1　子空间聚类示意图

Fig. 1　The sketch of subspace clustring

设 $X=\{X_1,X_2,\cdots,X_n\}$ 是 n 个对象的集合,可以聚类成为 k 簇,其中 $X_i=\{x_{i1},x_{i2},\cdots,x_{im}\}$,$m$ 是每个对象所含的属性个数,$C=\{C_1,C_2,\cdots,C_k\}$ 是 k 个簇的聚类中心,$\boldsymbol{W}=\{w_1,w_2,\cdots,w_m\}$ 是 m 个属性的权重向量,β 属性权重的一个参数。W-K-Means 算法的目标函数是

$$J(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{il}w_j^{\beta}d(x_{ij},c_{lj}) \tag{1}$$

式中:$\boldsymbol{U}$ 为隶属度矩阵,$\boldsymbol{C}$ 为聚类中心矩阵,$\boldsymbol{W}$ 为属性权重矩阵,μ_{il} 表示对象 i 属于类 l 的隶属程度,$D_{il}=\sum_{j=1}^{m}d(x_{ij},c_{lj})$ 表示对象 X_i 到其聚类中心 C_l 的距离。可见 W-K-Means 算法对所有簇类都是一个属性权重向量。该算法认为权重非常小的属性是噪声,可以很好地去除噪声属性,从而提高聚类正确率,但算法往往不能发现隐含在子空间上的簇类。

1.2　EWKM 子空间聚类算法

EWKM 是一种通过熵加权对高维稀疏数据集进行子空间聚类的软子空间聚类算法。EWKM 算法

依然通过计算属性权重实现对属性的粒化，但不同的子空间属性集的权重向量不同。权重矩阵在EWKM算法开始时是随机指定的，但随着算法的迭代深入，权重矩阵不断修正，直到满足某种条件算法结束，从而得到各子空间属性的权重向量。

EWKM算法的的属性权重矩阵 $\boldsymbol{W}=\{W_1,W_2,\cdots,W_k\}$ 是 k 个簇的属性权重向量，其中 $W_l=\{w_{l,1},w_{l,2},\cdots,w_{l,m}\}$。目标函数是

$$J(\boldsymbol{U},\boldsymbol{C},\boldsymbol{W})=\sum_{l=1}^{k}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{li}w_{lj}d(x_{ij},c_{lj})+\gamma\sum_{j=1}^{m}w_{lj}\log w_{lj}\right] \tag{2}$$

其中限制条件是

$$\begin{cases}\sum_{l=1}^{k}\mu_{li}=1 \quad 1\leqslant i\leqslant n, \quad 1\leqslant l\leqslant k, \quad \mu_{li}\in\{0,1\}\\ \sum_{j=1}^{m}w_{lj}=1 \quad 1\leqslant l\leqslant k, \quad 1\leqslant j\leqslant m, \quad 0\leqslant w_{lj}\leqslant 1\end{cases} \tag{3}$$

根据目标函数最小化原则以及限制条件，在固定隶属度矩阵和类心矩阵的情况下，运用拉格朗日乘子技术，就能得到计算属性权值矩阵的迭代公式，这样每次迭代中重新计算属性权值，最终会得到兴趣度子空间及相应的聚类结果。

2 属性样本同步粒化的AP熵加权软子空间聚类算法

2.1 EWAP算法属性粒化的迭代算法

AP聚类算法没有考虑属性对整个空间或子空间的贡献，认为所有属性的重要度是一样的，所以无法去除噪声属性，不能搜索子空间及相应的簇类。EWAP算法将计算属性权重纳入到聚类之中，去除冗余属性后，在每次AP聚类的迭代过程中修改属性权重矩阵，最终完成属性集的迭代粒化，同时得到子空间及其聚类结果。

AP算法开始时所有数据点都被认为是可能的聚类中心，数据点 i 与 j 之间的相似度用 $s(i,j)=-\|x_i-x_j\|^2$ 来计算，其值存储在 $n\times n$ 的矩阵 $\boldsymbol{S}$ 中（n 为数据点的个数）。迭代开始后中，可靠性和有效性两种信息在任意两数据点之间传递，并形成 $\boldsymbol{R}$ 矩阵和 $\boldsymbol{A}$ 矩阵。算法的最终目标是目标函数 $J=\sum_{l=1}^{k}\sum_{i=1}^{n}d\cdot(X_i-C_l)$ 的值最小（k 为簇的个数，X_i 为第 i 个数据点，C_l 为第 l 个聚类中心）。在聚类迭代过程中，EWAP依据 $J=\sum_{l=1}^{k}\sum_{i=1}^{n}d(X_i-C_l)$ 最小的原则更改属性权值。

为达到搜寻兴趣度子空间的目的，给每个潜在的聚类中心赋予一个属性权值向量，每个对象则赋予一个隶属度，EWAP算法的目标函数变为

$$J=\sum_{l=1}^{k}\sum_{i=1}^{n}U_iW_ld(X_i-C_l) \tag{4}$$

为了计算方便，式(4)修改为

$$J=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{li}w_{lj}d(x_{ij}-c_{lj})^2 \tag{5}$$

考虑到算法针对稀疏矩阵也适用，采用熵加权的方法，式(5)修改为

$$J=\sum_{l=1}^{k}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{li}w_{lj}d(x_{ij}-c_{lj})^2+\gamma\sum_{j=1}^{m}w_{lj}\log w_{lj}\right] \tag{6}$$

约束条件为

$$\begin{cases}\sum_{l=1}^{k}\mu_{li}=1 \quad 1\leqslant l\leqslant n, \quad 1\leqslant i\leqslant n, \quad \mu_{li}\in\{0,1\}\\ \sum_{j=1}^{m}w_{lj}=1 \quad 1\leqslant l\leqslant k, \quad 1\leqslant j\leqslant m, \quad 0\leqslant w_{lj}\leqslant 1\end{cases} \tag{7}$$

用拉格朗日乘子法最小化式(6),求出属性权重矩阵的迭代公式,有

$$\min J(w,\lambda)=J=\sum_{l=1}^{k}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m}\mu_{li}w_{lj}d(x_{ij}-c_{lj})^2+\gamma\sum_{j=1}^{m}w_{lj}\log w_{lj}\Big]-\sum_{l=1}^{k}\lambda_l(\sum_{j=1}^{m}w_{lj}-1) \tag{8}$$

上式可以分解为 k 个独立的最小化问题,即

$$\min J=\sum_{i=1}^{n}\sum_{j=1}^{m}w_{lj}d(x_{ij}-c_{lj})^2+\gamma\sum_{j=1}^{m}w_{lj}\log w_{lj}-\lambda_l(\sum_{j=1}^{m}w_{lj}-1) \tag{9}$$

$$\frac{\partial J_l}{\partial\lambda_l}=(\sum_{j=1}^{m}w_{lj}-1)=0 \tag{10}$$

$$\frac{\partial J_l}{\partial w_{lt}}=\sum_{i=1}^{n}\mu_{li}d(x_{it}-c_{lt})^2+\gamma(1+\log w_{lt})-\lambda_l=0 \tag{11}$$

这样通过计算得到属性权值的更新公式为

$$w_{lt}=\frac{\exp(-D_{lt}/\gamma)}{\sum_{j=1}^{m}\exp(-D_{lt}/\gamma)} \tag{12}$$

式中 $D_{lt}=\sum_{i=1}^{n}\mu_{li}d(x_{it}-c_{lt})^2$。

隶属度矩阵迭代公式为

$$\begin{cases}\mu_{li}=1 \quad \sum_{j=1}^{m}w_{lj}(x_{ij}-c_{lj})\leqslant\sum_{j=1}^{m}w_{rj}(x_{ij}-c_{rj}),\ 1\leqslant r\leqslant k\\ \mu_{li}=0 \quad \text{其他}\end{cases} \tag{13}$$

聚类中心的选择依据是 $\boldsymbol{E}=\boldsymbol{A}+\boldsymbol{R}$ 的对角线与所给阈值的比较结果。由于引入属性加权系数,AP 聚类中计算 $\boldsymbol{A}$,$\boldsymbol{R}$ 的公式发生了变化。原有的 $\boldsymbol{A}$ 和 $\boldsymbol{R}$ 矩阵都要乘以数据点所在类的属性权重矩阵 $\boldsymbol{W}$。

2.2　EWAP 算法

EWAP 使用仿射传播聚类算法对数据集聚类,但在每次迭代过程中都要修改属性权重,从而使得属性在每个兴趣度子空间的权重向量逐渐精确,最终完成属性集的迭代粒化,并同时实现在每个兴趣度子空间的样本集的粒化,实现子空间聚类。

输入:待聚类数据集

输出:子空间聚类结果

Step (1)初始化:计算 n 个待聚类对象两两之间的相似度值,放在相似度矩阵 $\boldsymbol{S}$ 中;$\boldsymbol{S}$ 对角线元素赋予相同的初始值并赋给参数 p;给矩阵 $\boldsymbol{R}$ 和 $\boldsymbol{A}$ 赋初始值;给 lam(引进参数 lam 是为了控制 AP 算法的震荡)赋初始值。

Step (2)迭代:

(1) 计算 $\boldsymbol{R}$

(a) $Rold=\boldsymbol{R}$

(b)计算 $\boldsymbol{R}$

(c)$\boldsymbol{R}=\boldsymbol{WR}$

(d) 计算 $r(k,k)$,放入 $\boldsymbol{R}$ 中

(e) $\boldsymbol{R}=(1-lam)\times\boldsymbol{R}+lam\times Rold$

(2) 计算 $\boldsymbol{A}$

(a)$Aold=\boldsymbol{A}$

(b)计算 $\boldsymbol{A}$

(c)$\boldsymbol{A}=\boldsymbol{WA}$

(d)计算 $a(k,k)$，放入 $\boldsymbol{A}$ 中

(e)$\boldsymbol{A}=(1-lam)\times\boldsymbol{A}+lam\times Aold$

(3) 求出新的聚类中心

(4)依据式(13)更新隶属度矩阵 $\boldsymbol{U}$

(5)依据式(12)更新属性权重矩阵 $\boldsymbol{W}$

(6) 判断以下条件是否满足，如果满足其中之一，则迭代终止：

(a)超过最大迭代次数；

(b)信息量的改变低于某一阈值；

(c)选择的类中心保持稳定。

Step (3) 输出聚类结果

3 实验与分析

3.1 算法准确率

这里采用 FScore 标准来评价子空间聚类结果。假设标准的聚类结果是 $S=\{S_1,S_2,\cdots,S_i,\cdots S_k\}$，某算法的聚类结果为 $C=\{C_1,C_2,\cdots,C_j,\cdots,C_r\}$，其中 $S=C$。假设 $T_{ji}=C_j\cap S_i$，$n_{ji}=|T_{ji}|$，$n_j=|C_j|$，$n_i=|S_i|$，则聚类结果的查准率为

$$P_{ji}=\frac{n_{ji}}{n_j} \tag{14}$$

聚类结果的查全率为

$$R_{ji}=\frac{n_{ji}}{n_i} \tag{15}$$

聚类结果的 F 度量为

$$F_{ji}=\frac{2P_{ji}R_{ji}}{P_{ji}+R_{ji}} \tag{16}$$

聚类结果 C_j 的 FScore 为

$$\text{FScore}_j=\max F_{ji} \tag{17}$$

所以整个聚类结果的 FScore 为

$$\text{FScore}=\sum_{j=1}^{r}\frac{n_j}{N}\text{FScore}_j \tag{18}$$

3.2 实验数据集

实验采用人工生成数据集和 UCI 数据集验证算法。人工数据集 D1，D2，D3 和 D4 的簇类分布在不同的子空间。数据集 D1 有 5 个属性，其中第 4，5 个属性是噪声，D1 的簇类数为 2。数据集 D2 有 10 个属性，分为 2 个子空间，簇类数为 5。数据集 D3 有 15 个属性，分为 3 个子空间，簇类数为 6。数据集 D4 有 20 个属性，2 个子空间，簇类数为 6。UCI 数据集为 Iris，Wine，Musk 和 Ionosphere。数据集特征如表 1 所示。

表 1　实验数据集

Tab. 1　Data sets of experiments

数据集	样本数	属性个数	聚类数
D1	40	5	2
D2	100	10	5
D3	500	15	6
D4	600	20	6
Iris	150	4	3
Wine	178	13	3
Musk	476	168	2
Ionosphere	351	4	2

3.3　实验结果与分析

实验将 EWKM,EWAP 和 AP 三种聚类算法在 8 个数据集上对准确率和运行时间做了比较,结果如表 2,3 所示。

表 2　EWKM,EWAP 和 AP 算法准确率比较

Tab. 2　Accuracy of EWKM, EWAP and AP

数据集	EWKM	EWAP	AP
D1	0.93	0.96	0.71
D2	0.92	0.94	0.65
D3	0.91	0.94	0.79
D4	0.91	0.93	0.81
Iris	0.97	0.98	0.91
Wine	0.85	0.90	0.89
Musk	0.92	0.92	0.86
Ionosphere	0.86	0.89	0.83

表 3　EWKM,EWAP 和 AP 算法聚类时间比较

Tab. 3　Time of EWKM, EWAP and AP　s

数据集	EWKM	EWAP	AP
D1	0.725 6	3.206 5	0.311 6
D2	0.812 7	5.922 8	0.521 3
D3	1.997 5	11.340 1	1.137 6
D4	9.893 3	29.464 1	6.733 1
Iris	0.892 6	3.051 6	0.322 1
Wine	0.913 2	4.102 4	0.657 3
Musk	1.962 7	19.922 6	1.312 4
Ionosphere	18.091 1	18.272 6	16.697 1

由表 2 可以看出,EWAP 算法的准确率比 EWKM 算法稳定而且较高。主要原因是 EWKM 算法的初始聚类中心是随机确定的,聚类中心选择的好坏会对聚类结果的准确性产生很大影响。EWAP 算法将所有待聚类对象作为潜在的聚类中心,所以算法结果稳定,准确率高。AP 算法由于不适用于子空间聚类,所以准确率也较低。

由表 3 可以看出,由于 EWKM,EWAP 算法需要计算子空间,所以运行时间较长,效率较低,而 AP 算法效率最高。AP 算法要在所有数据点之间传递消息,所以基于 AP 聚类的 EWAP 算法效率要低于基于 K-means 的 EWKM 算法。

4　结束语

本文将粒度计算的思想引入到仿射传播聚类中,使得 AP 算法拓展到子空间领域。提出属性样本同步粒化的 AP 加权软子空间聚类算法,在仿射传播聚类算法中引入属性的权重矩阵,使得每次聚类的迭代过程都可以修改属性权重矩阵,算法结束时同时找到了兴趣度子空间和在此之上的簇类,达到属性样本同步粒化的目的。EWAP 算法虽然聚类结果有较高的精确度,也很稳定,但运行时间较长,需进一步改进算法提高效率。

参考文献：

[1] 郑馨,王勇,汪国有. EM 聚类和 SVM 自动学习的白细胞图像分割算法[J]. 数据采集与处理, 2013,28(5):614-619.
Zheng Xin, Wang Yong, Wang Guoyou. White blood cell segmentation using expectation-maximization and automatic support vector machine learning[J]. Journal of Data Acquisition and Processing, 2013,28(5):614-619.

[2] Malyszko D, Stepaniuk J. Rough entropy hierarchical agglomerative clustering in image segmentation[J]. Transactions on Rough Sets XIII, 2011, 6499: 89-103.

[3] 刘金勇,郑恩辉,陆慧娟. 基于聚类和微粒群优化的基因选择方法[J]. 数据采集与处理, 2014,29(1):83-89.
Liu Jinyong, Zheng Enhui, Lu Huijuan. Gene selection based on clustering method and particle swarm optimization[J]. Journal of Data Acquisition and Processing, 2014,29(1):83-89.

[4] 夏利,王建东,张霞,等. 聚类再回归方法在机场噪声时间序列预测中的应用[J]. 数据采集与处理,2014,29(1):152-156.
Xia Li, Wang Jiandong, Zhang Xia, et al. Application of cluster regression in time series prediction of airport noise[J]. Journal of Data Acquisition and Processing, 2014,29(1):152-156.

[5] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic subspace clustering of high dimensionating data for data mining applications[C]// Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data. [S. l.]: ACM Press, 1998: 94-105.

[6] Frey J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[7] Jia S, Qian Y T, Ji Z. Band selection for hyperspectral imagery using affinity propagation[C]// Proceedings of the 2008 Digital Image Computing: Techniques and Applications. Canberra, ACT: IEEE, 2008:137-141.

[8] Li G , Guo L, Liu T M, et al. Grouping of brain MR images via affinity propagation[C]// IEEE International Symposium on Circuits and Systems, 2009(ISCAS 2009). Taipei: IEEE, 2009: 2425-2428.

[9] Dueck D, Frey B J, Jojic N, et al. Constructing treatment portfolios using affinity propagation[C]// Proceedings of 12th Annual International Conference, RECOMB 2008. Berlin: Springer, 2008: 360-371.

[10] Kelly K. Afinity program slashes computing times[EB/OL]. http://www.news.utoronto.ca/bin6/070215-2952.asp, 2007-02-15.

[11] Huang Z, Ng M, Rong H. Automated variable weighting in K-means type clustering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005, 27(5): 657-668.

[12] Frigui H, Nasraoui O. Simultaneous clustering and attribute discrimination[C]// Proceeding of the 9th IEEE International Conference on Fuzzy Systems. San Antonio, TX: IEEE, 2000:158-163.

[13] Jing L, Michael K, Ng, Joshua Zhexue H. An entropy weigh K-means algorithm for subspace clustering of high dimensional sparse data[J]. IEEE Transactions on Knowledge and Data Engineering, 2007, 19(8):1-16.

作者简介：

朱红(1970-)，女，博士，副教授，研究方向：数据挖掘，图像处理，粒度计算和机器学习，E-mail: zhuhongwin@126.com。

丁世飞(1963-)，男，教授，博士生导师，研究方向：人工智能，智能信息处理，模式识别，机器学习和数据挖掘，E-mail: dingshifei@sina.com。