学 术 论 文

鲁棒的极大熵聚类算法 RMEC 及其例外点标识

邓赵红[1,2]，王士同[1,2,4]，吴锡生[1,2]，胡德文[3]

（1. 江南大学信息工程学院，江苏无锡 214036；2. 南京大学软件新技术国家重点实验室，南京 210016；3. 国防科技大学自动化学院，长沙 410073；
4. 南京理工大学计算机系，南京 212000）

[摘要] 针对极大熵聚类算法 MEC（maximum entropy clustering）对例外点（outliers）较敏感和不能标识例外点的缺陷，提出了一种改进的极大熵聚类算法 RMEC（robust maximum entropy clustering）。该算法的基本思想是通过引入Vapnik's ε－不敏感损失函数和权重因子重新构建目标函数，并利用优化理论推导出新的学习公式。RMEC 算法不但对例外点较之 MEC 算法有更好的鲁棒性，而且还能有效地利用学习后的权重因子标识出数据集中存在的例外点。仿真试验结果亦表明了 RMEC 算法的上述优点。

[关键词] 熵；聚类；鲁棒性；例外点；ε－不敏感损失函数；权重因子

[中图分类号] TP18 **[文献标识码]** A **[文章编号]** 1009－1742（2004）09－0038－08

1 引言

在众多的聚类算法中，一类基于熵[1]的聚类算法近年来受到广泛关注。其中极大熵聚类算法 MEC（maximum entropy clustering）已得到了较为深入的研究。该类算法与其他聚类算法[2,3]最大的不同在于：它采用熵函数作为目标函数的一个组成部分来对数据集进行分类。相关文献[4~6]发表了许多 MEC 算法及其改进版本的研究成果。

大量实验表明，MEC 算法对例外点较敏感，例外点的干扰常使学习得到的聚类中心严重偏离正确的聚类中心。现实中的数据集，例外点是普遍存在的，因此对 MEC 算法进行改进，提高算法对例外点的抗干扰能力即鲁棒性能很有必要。同时，大量事实表明，数据集中的例外点暗含一些重要的信息，这些重要信息对于科学研究和生产实践有着非凡意义。例如，在地球南极的臭氧层漏洞被人类地面的观测者发现时，地球卫星在数年前就已探测出这个漏洞的存在，只是计算机程序在处理卫星数据时未能检测出暗含漏洞信息的例外点，这使得人类对臭氧层漏洞未采取任何补救措施达 9 年之久。因此，聚类算法对数据集进行聚类的同时如能有效地对例外点进行检测，其意义重大。遗憾的是，几乎所有的 MEC 算法都没有检测和标识例外点的能力。近年来，如何利用聚类方法标识大量数据中的例外点受到了越来越多的关注，业已提出了一些检测例外点的有效方法[3,7,8]。

针对提高 MEC 聚类算法对例外点干扰鲁棒性和有效检测例外点的需要，笔者通过引入Vapnik's ε－不敏感损失函数[9]和权重因子[3]重新构建新的目标函数，利用优化理论[10]推导出新的学习公式，从而得到新的聚类算法 RMEC。该算法有如下优点：首先，ε－不敏感损失函数对例外点有很好的抗干扰能力，ε－不敏感损失函数的引入使得 RMEC 算法较 MEC 算法对例外点有更好的鲁棒性。其次，权重因子的引入，使得 RMEC 算法能有效地根据权重因子检测出数据集中的例外点，这是一般的 MEC 算法所不能实现的。

[收稿日期] 2003－09－28；修回日期 2003－11－20

[基金项目] 国家自然科学基金资助项目（60225015）；江苏省自然科学基金资助项目（BK2003017）；江苏计算机信息技术重点实验室资助。

[作者简介] 邓赵红（1981－），男，安徽省蒙城县人，江南大学硕士研究生

2　极大熵聚类算法 MEC

在多种版本的极大熵聚类算法 MEC 中，虽然描述各不相同，但只是形式上的差别。这里仅介绍文献[5]中的极大熵聚类算法 MEC。

对于样本集 $X = \{\boldsymbol{x}_i \mid \boldsymbol{x}_i \in R^d,\ i = 1, 2, \cdots, n\}$，根据某种相似性度量，它被聚类成 $c(2 \leqslant c < n)$ 个子类，各类中心用矩阵 $\boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_c]$ $\boldsymbol{v}_i \in R^d,\ i = 1, 2, \cdots, c$ 表示；划分可用矩阵 $\boldsymbol{U} = [u_{ik}] \in R^{cn}$ 表示，其中每一项满足

$$u_{ik} \in [0, 1], 1 \leqslant i \leqslant c, 1 \leqslant k \leqslant n,$$

$$\sum_{i=1}^{c} u_{ik} = 1, 1 \leqslant k \leqslant n,$$

$$0 < \sum_{k=1}^{n} u_{ik} < n \tag{1}$$

极大熵聚类算法 MEC 的目标函数定义为

$$J(\boldsymbol{U}, \boldsymbol{V}) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \| \boldsymbol{x}_k - \boldsymbol{v}_i \|^2 + \gamma \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \ln u_{ik} \tag{2}$$

这里 $\| \cdot \|$ 表示欧几里德距离，γ 是非负常数。

对于目标函数 J，如下定理成立：

定理 1　$J(\boldsymbol{U}, \boldsymbol{V})$ 取极小值的必要条件为

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik}\boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}} \quad i = 1, 2, \cdots, c \tag{3}$$

$$u_{ik} = \frac{\exp(-\| \boldsymbol{x}_k - \boldsymbol{v}_i \|^2 / \gamma)}{\sum_{h=1}^{c} \exp(-\| \boldsymbol{x}_k - \boldsymbol{v}_h \|^2 / \gamma)}, \quad i = 1, 2, \cdots, c, k = 1, 2, \cdots, n \tag{4}$$

基于定理 1，可以把极大熵聚类算法 MEC 描述如下：

Step 1　固定 $c(2 \leqslant c < n)$，置精度级为 ε_a，迭代次数 $r = 0$，初始化分割矩阵 $\boldsymbol{U}(0)$；

Step 2　利用式 (3) 计算出 $\boldsymbol{V}(r)$；

Step 3　利用式 (4) 和 $\boldsymbol{V}(r)$ 更新 $\boldsymbol{U}(r)$ 为 $\boldsymbol{U}(r+1)$；

Step 4　若 $\| \boldsymbol{U}(r+1) - \boldsymbol{U}(r) \|_F \leqslant \varepsilon_a$（亦可根据需要改置其他条件），则停止，否则置 $r = r + 1$，并返回 Step 2。

其中 $\| \cdot \|_F$ 表示 Frobenius 范数。上述极大熵聚类算法 MEC 的一个缺陷是对数据集中的例外点较敏感，这极大地影响了极大熵聚类算法 MEC 的性能，同时，该算法也不具备检测例外点的能力。针对此问题，笔者提出新的聚类算法——鲁棒极大熵聚类算法 RMEC（robust maximum entropy clustering）。

3　算法 RMEC

3.1　改进的目标函数

算法 MEC 中用二次方损失函数[11, 12]作为样本和聚类中心区分的度量尺度，选用二次方损失函数作为尺度的目的是为了算法的简便性和较低的计算负担。但是，二次损失函数也同时使 MEC 算法对于例外点较敏感。文献 [9] 中提出了许多鲁棒的损失函数，其中 Vapnik's ε - 不敏感损失函数[9, 12~14]由于它的简便性最受关注。

Vapnik's ε - 不敏感损失函数可表示为

$$|f|_\tau = \begin{cases} 0 & |f| \leqslant \varepsilon, \\ |f| - \varepsilon & |f| < \varepsilon \end{cases} \tag{6}$$

其中 $\varepsilon \geqslant 0$ 表示不敏感性参数，当 $\varepsilon = 0$ 时，Vapnik's ε - 不敏感损失函数变为绝对误差损失函数。

为了使新的算法能对数据集中的例外点进行有效的检测，新的目标函数同时引入了权重因子 $\boldsymbol{w} = (w_1, w_2, \cdots, w_n)$，$\sum_{k=1}^{n} w_k = W$，$W$ 是常数。每个数据的权重因子反映了该数据在样本集中的重要程度，可以根据它的大小确定其是否是例外点。新的目标函数将有助于如何求得每个样本点的权重因子。

利用Vapnik'sε - 不敏感损失函数代替二次损失函数，同时引入权重因子和不敏感参数惩罚项，可以重新构造目标函数为

$$J(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{w}, \boldsymbol{\varepsilon}) = \sum_{i=1}^{c}\sum_{k=1}^{n} (u_{ik} \| \boldsymbol{x}_k - \boldsymbol{v}_i \|_{\varepsilon_i} / w_k^q) + \gamma \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik} \ln u_{ik} + \alpha \sum_{i=1}^{c} \| \varepsilon_i \| \tag{7}$$

其中

$$\| \boldsymbol{x}_k - \boldsymbol{v}_i \|_{\varepsilon_i} = \sum_{j=1}^{d} |x_{ij} - v_{ij}|_{\varepsilon_{ij}},$$

$$\| \varepsilon_i \| = \sum_{j=1}^{d} |\varepsilon_{ij}|,$$

$$\sum_{k=1}^{n} w_k = W,$$

$$\gamma \geqslant 0, \alpha \geqslant 0, q \geqslant 0 \tag{8}$$

q 是一个由用户定义的指数常数。$\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \cdots,$

ε_{id}）为第 i 类的不敏感参数矢量。式（8）定义的好处在于有很好的抗例外点干扰能力，文献［9，12］已经有所论述。值得注意的是，新构造的目标函数能带来以下好处：

1）新的目标函数可以保证 J 的前两项最小的同时使得不敏感参数惩罚项尽可能的小。通过调整 γ 可以灵活地根据需要确定不敏感参数惩罚项的影响。

2）引入不敏感参数惩罚项能使推导出的新算法自动调节不敏感性参数，不再需要人为地确定不敏感性参数，这就避免了不敏感性参数选取的随意性（见定理2证明）。

3）总的来说，例外点远离数据集中的任何一类。RMEC算法赋予例外点一个较大的权重因子 w_k（一个较小的值 $1/w_k^q$），新构造的目标函数中权重因子的引入，使得RMEC算法能根据学习后得到的权重因子数值检测出数据集中的例外点 $x_k(w_k > w_0)$，w_0 为阈值，一般由专家给出。参数 q 在聚类处理中扮演重要的角色，q 足够大时，每个数据点的权重因子趋近于 W/n，也就是所有的数据有相同的权重因子，当 q 趋近于0时，权重系数的影响将达到最大。

3.2 算法 RMEC 的推导及证明

为了得到RMEC算法，首先证明下面的定理：

定理2 目标函数式（7）取极小值的必要条件是

$$u_{ik} = \exp\left[-\| x_k - v_i \|_{\varepsilon_i}/\gamma w_k^q\right] \Big/ \sum_{h=1}^{c}\exp\left[-\| x_k - v_h \|_{\varepsilon_h}/\gamma w_k^q\right] \tag{9}$$

$$w_k = \left[\sum_{i=1}^{c}(u_{ik}\| x_k - v_h \|_{\varepsilon_i}) \Big/ \sum_{h=1}^{n}\sum_{i=1}^{c}(u_{il}\| x_h - v_i \|_{\varepsilon_i})\right]^{1/(q+1)} W \tag{10}$$

$$v_{ij} = (x_{lj} + x_{mj})/2,\ \varepsilon_{ij} = (x_{mj} - x_{lj})/2,\ \forall\{l, m \mid \lambda_{ijl}^{+} \in \Gamma_{ij}^{+}, \lambda_{ijm}^{-} \in \Gamma_{ij}^{-}\},\ i = 1, 2, \cdots, c,\ j = 1, 2, \cdots, d,\ l, m = 1, 2, \cdots, n \tag{11}$$

式中：

$$\Gamma_{ij}^{+} = \{\lambda_{ijk}^{+} \mid \lambda_{ijk}^{+} \in (0, u_{ik}/w_k^q)\},\ \Gamma_{ij}^{-} = (\lambda_{ijk}^{-} \mid \lambda_{ijk}^{-} \in (0, u_{ik}/w_k^q)\},\ l = 1, 2, \cdots, n \tag{12}$$

$$\min\sum_{k=1}^{n}(\lambda_{ijk}^{+} - \lambda_{ijk}^{-})x_{xj},\ \text{s.t.}\sum_{k=1}^{n}\lambda_{ijk}^{+} = \sum_{k=1}^{n}\lambda_{ijk}^{-},\ \sum_{k=1}^{n}(\lambda_{ijk}^{+} + \lambda_{ijk}^{-}) \leqslant \alpha,\ \lambda_{ijk}^{+}, \lambda_{ijk}^{-} \in [0, u_{ik}/w_k^q] \tag{13}$$

证明 如果 w，ε，V 一定，求目标函数式（7）的极小值可以表述为求式（14）的极小值。

$$f_0(U) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}\| x_k - v_i \|_{\varepsilon_i}/w_k^q) + \gamma\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\ln u_{ik} + \alpha\sum_{i=1}^{c}\| \varepsilon_i \| \tag{14}$$

式（14）在约束条件式（1）下的拉格朗日函数为

$$F_0(U, \lambda) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}\| x_k - v_i \|_{\varepsilon_i}/w_k^q) + \gamma\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\ln u_{ik} + \alpha\sum_{i=1}^{c}\| \varepsilon_i \| - \sum_{k=1}^{n}\lambda_k\left(\sum_{i=1}^{c}u_{ik} - 1\right) \tag{15}$$

λ_k 是拉格朗日乘子。式（15）取极值时，拉格朗日函数梯度为0，得到

$$\frac{\partial F_0(U, \lambda)}{\partial \lambda_k} = 0 \tag{16}$$

$$\frac{\partial F_0(U, \lambda)}{\partial u_{ik}} = 0 \tag{17}$$

由式（16）得出

$$\sum_{i=1}^{c}u_{ik} = 1 \tag{18}$$

由式（17）得出

$$u_{ik} = \exp\left(\frac{\lambda_k}{\gamma} - 1\right)\cdot\exp\left\{\frac{\| x_k - v_i \|_{\varepsilon_i}}{\gamma w_k^q}\right\} \tag{19}$$

由式（18）和式（19）可得

$$1 = \sum_{h=1}^{c}\exp\left(\frac{\lambda_k}{\gamma} - 1\right)\exp\left\{-\frac{\| x_k - v_h \|_{\varepsilon_h}}{\gamma w_k^q}\right\} \tag{20}$$

式（19）和式（20）两边分别相除得

$$u_{ik} = \exp\left(-\| x_k - v_i \|_{\varepsilon_i}/\gamma w_k^q\right) \Big/ \sum_{h=1}^{c}\exp\left(-\| x_k - v_h \|_{\varepsilon_h}/\gamma w_k^q\right),$$

于是式（9）得证。

参照式（9）的证明方法，可得式（10）的证明结果。

如果 w，U 一定，求目标函数式（7）的极小值可以表述为求式（21）的极小值。

$$f_1(V, \varepsilon) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}\| x_k - v_i \|_{\varepsilon_i}/w_k^q) +$$

$$\gamma\sum_{i=1}^{c}\sum_{k=1}^{n}u_{ik}\ln u_{ik}+\alpha\sum_{i=1}^{c}\|\boldsymbol{\varepsilon}_i\| \tag{21}$$

式（21）取极小等价于式（22）取极小。

$$f'_1(\boldsymbol{V},\boldsymbol{\varepsilon})=\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}\|\boldsymbol{x}_k-\boldsymbol{v}_i\|_{\varepsilon_i}/w_k^q)+\alpha\sum_{i=1}^{c}\|\boldsymbol{\varepsilon}_i\|=\sum_{i=1}^{c}\sum_{j=1}^{d}\Big(\sum_{k=1}^{n}(u_{ik}|x_{kj}-v_{ij}|_{\varepsilon_{ij}}/w_k^q)+\alpha\cdot\boldsymbol{\varepsilon}_{ij}\Big) \tag{22}$$

一般来说，不是所有的样本数据分量 x_{kj}，$k=1,2,\cdots,n$，$j=1,2,\cdots,d$ 都满足不等式

$$v_{ij}-x_{kj}\leqslant\varepsilon_{ij},\quad x_{kj}-v_{ij}\geqslant\varepsilon_{ij} \tag{23}$$

如果引入松弛变量 ξ_{ijk}^{+}，$\xi_{ijk}^{-}\geqslant 0$，对于所有的样本数据分量 x_{kj} 可满足下面的不等式

$$v_{ij}-x_{kj}\leqslant\varepsilon_{ij}+\xi_{ijk}^{+},\ x_{kj}-v_{ij}\leqslant\varepsilon_{ij}+\xi_{ijk}^{-} \tag{24}$$

于是式（22）可表示为

$$f'_1(\boldsymbol{V},\boldsymbol{\varepsilon},\boldsymbol{\xi})=\sum_{i=1}^{c}\sum_{j=1}^{d}\Big(\sum_{k=1}^{n}(u_{ik}(\xi_{ijk}^{+}+\xi_{ijk}^{-})/w_k^q)+\alpha\varepsilon_{ij}\Big) \tag{25}$$

其中约束条件为式（24）和 ξ_{ijk}^{+}，$\xi_{ijk}^{-}\geqslant 0$。对含有上述约束条件的式（25）的拉格朗日函数为

$$F_1(\boldsymbol{V},\boldsymbol{\varepsilon},\boldsymbol{\xi},\boldsymbol{\lambda})=\sum_{i=1}^{c}\sum_{j=1}^{d}\Big\{\Big[\sum_{k=1}^{n}(u_{ij}(\xi_{ikj}^{+}+\xi_{ikj}^{-})/w_k^q)+\alpha\varepsilon_{ij}\Big]-\sum_{k=1}^{n}\lambda_{ijk}^{+}(\varepsilon_{ij}+\xi_{ijk}^{+}-v_{ij}+x_{kj})-\sum_{k=1}^{n}\lambda_{ijk}^{-}(\varepsilon_{ij}+\xi_{ijk}^{-}+v_{ij}-x_{kj})-\sum_{k=1}^{n}(\mu_{ijk}^{+}\xi_{ijk}^{+})-\sum_{k=1}^{n}(\mu_{ijk}^{-}\xi_{ijk}^{-})-\beta_{ij}\varepsilon_{ij}\Big\} \tag{26}$$

式中 λ_{ijk}^{+}，λ_{ijk}^{-}，μ_{ijk}^{+}，μ_{ijk}^{-}，$\beta_{ij}\geqslant 0$ 是拉格朗日乘子。问题求解的目标是拉格朗日函数式（26）相对于 v_{ij}，ξ_{ijk}^{+}，ξ_{ijk}^{-}，ε_{ij} 取最小，同时相对于拉格朗日乘子变量取最大。拉格朗日函数式（26）相对于 v_{ij}，ξ_{ijk}^{+}，ξ_{ijk}^{-}，ε_{ij} 最小时，满足条件

$$\frac{\partial F_1}{\partial v_{ij}}=\sum_{k=1}^{n}(\lambda_{ijk}^{+}-\lambda_{ijk}^{-})=0 \tag{27}$$

$$\frac{\partial F_1}{\partial \xi_{ijk}^{+}}=u_{ij}/w_k^q-\lambda_{ijk}^{+}-\mu_{ijk}^{+}=0 \tag{28}$$

$$\frac{\partial F_1}{\partial \xi_{ijk}^{-}}=u_{ij}/w_k^q-\lambda_{ijk}^{-}-\mu_{ijk}^{-}=0 \tag{29}$$

$$\frac{\partial F_1}{\partial \varepsilon_{ij}}=\alpha-\sum_{k=1}^{n}(\lambda_{ijk}^{+}+\lambda_{ijk}^{-})-\beta_{ij}=0 \tag{30}$$

$\lambda_{ijk}^{+},\lambda_{ijk}^{-},\mu_{ijk}^{+},\mu_{ijk}^{-},\beta_{ij}\geqslant 0$ 和式(27) 至式(30) 表明 $\lambda_{ijk}^{+},\lambda_{ijk}^{-}\in[0,u_{ik}/w_k^q]$，$\sum_{k=1}^{n}(\lambda_{ijk}^{+}+\lambda_{ijk}^{-})\in[0,\alpha]$。把式（27）至式（30）代入式（26）可以得到

$$\min F_1(\boldsymbol{V},\boldsymbol{\varepsilon},\boldsymbol{\xi})=\sum_{i=1}^{c}\sum_{j=1}^{d}\Big[-\sum_{k=1}^{n}(\lambda_{ijk}^{+}-\lambda_{ijk}^{-})x_{kj}\Big] \tag{31}$$

由经典的拉格朗日对偶理论知，

$$\max\sum_{i=1}^{c}\sum_{j=1}^{d}\Big(-\sum_{k=1}^{N}(\lambda_{ijk}^{+}-\lambda_{ijk}^{-})x_{kj}\Big),\quad \text{s.t.}\ \sum_{k=1}^{n}\lambda_{ijk}^{+}-\sum_{k=1}^{n}\lambda_{ijk}^{-}=0,\quad \sum_{k=1}^{n}(\lambda_{ijk}^{+}+\lambda_{ijk}^{-})\leqslant\alpha,\quad \lambda_{ijk}^{+},\lambda_{ijk}^{-}\in[0,u_{ik}/w_k^q] \tag{32}$$

即为求解问题的所谓 Wolfe 对偶表达。令

$$g_{ij}=-\sum_{k=1}^{n}(\lambda_{ijk}^{+}-\lambda_{ijk}^{-})x_{kj} \tag{33}$$

式（32）可分解为式（34）的优化问题：

$$\max\ g_{ij}=-\sum_{k=1}^{n}(\lambda_{ijk}^{+}-\lambda_{ijk}^{-})x_{kj}\quad \text{s.t.}\ \sum_{k=1}^{n}\lambda_{ijk}^{+}-\sum_{k=1}^{n}\lambda_{ijk}^{-}=0\quad \sum_{k=1}^{n}(\lambda_{ijk}^{+}+\lambda_{ijk}^{-})\leqslant\alpha,\quad \lambda_{ijk}^{+},\lambda_{ijk}^{-}\in[0,u_{ik}/w_k^q],\quad i=1,2,\cdots,c,\ j=1,2,\cdots,d \tag{34}$$

亦即

$$\min\sum_{k=1}^{n}(\lambda_{ijk}^{+}-\lambda_{ijk}^{-})x_{kj},\quad \text{s.t.}\ \sum_{k=1}^{n}\lambda_{ijk}^{+}=\sum_{k=1}^{n}\lambda_{ijk}^{-}\quad \sum_{k=1}^{n}(\lambda_{ijk}^{+}+\lambda_{ijk}^{-})\leqslant\alpha,\quad \lambda_{ijk}^{+},\lambda_{ijk}^{-}\in[0,u_{ik}/w_k^q],\quad i=1,2,\cdots,c,\ j=1,2,\cdots,d \tag{35}$$

按照优化理论的 Kuhn－Tucker定理，在鞍点，对偶变量与约束的乘积为零，即

$$\lambda_{ijk}^{+}(\varepsilon_{ij}+\xi_{ijk}^{+}-v_{ij}+x_{kj})=0,\quad \lambda_{ijk}^{-}(\varepsilon_{ij}+\xi_{ijk}^{-}+v_{ij}-x_{kj})=0,\quad \mu_{ijk}^{+}\xi_{ijk}^{+}=0,\quad \mu_{ijk}^{-}\xi_{ijk}^{-}=0,\quad \beta_{ij}\varepsilon_{ij}=0 \tag{36}$$

又由式（28）和式（29）可得

$$\mu_{ikj}^{+} = u_{ik}/w_k^q - \lambda_{ijk}^{+},$$
$$\mu_{ikj}^{-} = u_{ik}/w_k^q - \lambda_{ijk}^{-} \quad (37)$$

结合式（36）和式（37），进行如下推理：

当 $\lambda_{ijk}^{+}, \lambda_{ijk}^{-} \in (0, u_{ik}/w_k^q)$ 时，由式(37)可知 $\mu_{ijk}^{+}, \mu_{ijk}^{-} \in (0, u_{ik}/w_k^q)$，由式（36）中可知此时 $\xi_{ijk}^{+} = 0$，$\xi_{ijk}^{-} = 0$ 并可得到

$$\varepsilon_{ij} - v_{ij} + x_{kj} = 0,$$
$$\varepsilon_{ij} + v_{ij} - x_{kj} = 0 \quad (38)$$

即

$$v_{ij} = (x_{lj} + x_{mj})/2,$$
$$\varepsilon_{ij} = (x_{mj} - x_{lj})/2。 \quad (39)$$

这样，可以根据任意样本分量 x_{lj}，x_{mj}，当 λ_{ijl}^{+}，$\lambda_{ijm}^{+} \in (0, u_{ik}/w_k^q)$ 时，由式（39）确定聚类中心分量 v_{ij} 和不敏感性参数 ε_{ij} 的值。从而式（11）至式（13）式得证。

基于定理2，可以得到RMEC算法：

Step 1 固定 $c(2 \leqslant c < n)$，置精度级为 ε_a，迭代次数 $r=0$，初始化分割矩阵 $U(0)$ 和权重因子向量 $w(0)$；

Step 2 利用式（11）计算出 $\boldsymbol{V}(r)$；

Step 3 利用式（9）（10）和 $\boldsymbol{V}(r)$ 更新 $\boldsymbol{U}(r)$，$\boldsymbol{w}(r)$ 为 $\boldsymbol{U}(r+1)$ 和 $\boldsymbol{w}(r+1)$；

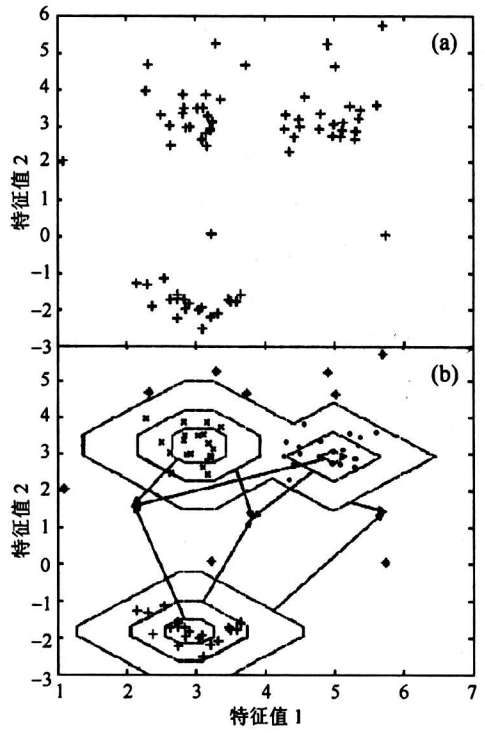
Step 4 若 $\| \boldsymbol{U}(r+1) - \boldsymbol{U}(r) \|_F \leqslant \varepsilon_a$（亦可根据需要改置其他条件），则停止，否则置 $r = r + 1$，并返回 Step 2。

对于RMEC算法中的 $\alpha \geqslant 0$，当 $\alpha = 0$ 时，有式（13）可以得出 λ_{ijk}^{+}，$\lambda_{ijk}^{-} = 0$ 此时由式（11）和式（12）可知 v_{ij}，ε_{ij} 得不到更新，故仅当 $\alpha > 0$ 时RMEC算法才能有效地执行。

4 仿真实验及实验分析

为了考察RMEC算法的性能，做了大量的仿真实验。实验从3个不同的方面对RMEC算法的性能进行考察。

实验1 此例说明可以用权重因子来标识数据集中的例外点。图1a是含有例外点的数据集，图1b显示了 $\gamma = 0.05$，$q = 0.9$，$W = 200$，$w(0) = 32$ 时，RMEC聚类算法的聚类效果。图1b中示出了数据点的等势线，等势线上的各点具有相同的权重因子。根据专家提供的权重因子阈值 w_0，就可以标识数据集中的例外点，即那些权重因子大于 w_0 的数据点。图1b中 * 点为被检测出的例外点（$w_k > w_0$）。

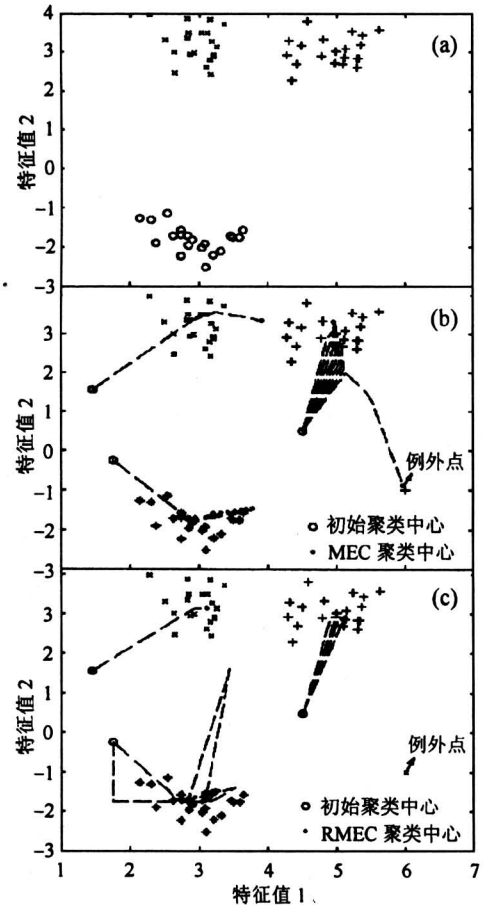


a—含有例外点的数据集，b—数据点的等势线及例外点(*)

图1 RMEC聚类算法的聚类效果和检测出的例外点

Fig.1 The clustering result of algorithm RMEC and the labeled outliers (*)

实验2 为了比较RMEC算法和MEC算法[5]对例外点的抗干扰能力，对图2a所示的数据集进行了聚类。数据集包含了3类清晰可分的样本数据，每类包含20个样本点，各类真正的类中心分别为(2.934 6，−1.798 3)，(2.959 5，−3.201 8)，(4.938 4，3.304 45)。为了试验的可比性，试验中采用相同的初始类中心：(1.75，−0.26)，(1.45，1.55) 和 (4.5，1.55)。所有的实验中，取 $\gamma = 0.05$。规定RMEC算法执行的最大迭代次数为10次，参数 $q = 0.9$，$W = 200$，MEC算法执行的最大迭代次数为100次。试验中通过在坐标 (6，−1) 处加不同数目 (1～8) 的例外点来测试2种算法的抗例外点干扰能力，图2b所示为MEC算法[5]对数据集加不同数目例外点的聚类结果。图2c所示为 $\alpha = 5$ 时，RMEC算法对数据集加不同数目例外点的聚类结果。图中虚线表示学习迭代过程中各类中心的变化轨迹。



a—数据集聚类，b—MEC 算法的聚类结果，
c—RMEC ($\alpha=5$) 算法的聚类结果

图 2　例外点个数分别为 0～8 时，MEC 算法[5]和 RMEC 算法（α = 5）聚类结果

Fig.2　The clustering results of MEC[5] and RMEC ($\alpha=5$) with a varying number (0～8) of outliers

如果真正的聚类中心用矩阵 $\boldsymbol{V}_1$ 表示，存在例外点时算法执行得到的聚类中心用矩阵 $\boldsymbol{V}_2$ 表示，则可用 Frobenius 范数 $\Delta \boldsymbol{V} = \| \boldsymbol{V}_1 - \boldsymbol{V}_2 \|_F$ 表示受例外点影响聚类中心的偏移量。$\Delta \boldsymbol{V}$ 越小，表示算法的抗例外点干扰能力越强。表 1 示出了 RMEC 算法取不同参数值 α 时类中心的偏移量和 MEC 算法[5]聚类中心的偏移量。

从表 1 可以看出，MEC 算法聚类结果受例外点影响很大，随着例外点数目的增加，偏移量 $\Delta \boldsymbol{V}$ 越来越大。从表 1 可以看出 $\alpha \geqslant 3$ 时，RMEC 算法受例外点影响则较小，$\Delta \boldsymbol{V}$ 在例外点数目为 0～8 的范围内基本稳定，并且 RMEC 算法能得到较合理的聚类中心，但 $\alpha \leqslant 2$ 时 RMEC 算法聚类中心稳定点已不合理，如图 3a 所示。试验结果表明，α 取较大值时 RMEC 算法能得到较理想的聚类效果，如图 3b 至图 3d 所示。当例外点数目很少时，MEC 算法得到的聚类中心偏移量 $\Delta \boldsymbol{V}$ 较 RMEC 算法得到的聚类中心偏移量 $\Delta \boldsymbol{V}$ 小，但随着例外点数目增加时，选取合适参数 α，RMEC 算法聚类效果明显优于 MEC 算法聚类效果。

对于 MEC 算法[5]和 RMEC 算法中相同的参数 γ，可以理解为模糊度参数。通过大量实验发现，γ 取值越大，样本点相对于各类的隶属关系越模糊，对于 c 类样本，当 γ 很大时，样本点相对于各类的隶属度趋近于 $1/c$，当 γ 太小时，2 种算法对初始类中心都特别敏感，因此 γ 取值不能太大也不能太小。如何选取最佳的参数值 γ，目前还没有很好的理论依据。

表 1　RMEC 算法和文献中的 MEC[5]算法性能比较

Table 1　The performance comparison between MEC[5] algorithm and RMEC algorithm

例外点数目	RMEC 算法 ΔV							MEC 算法 ΔV
	α = 0.1	α = 0.5	α = 1	α = 2	α = 3	α = 10	α = 100	
0	4.375 6	4.366 4	0.881 1	0.195 6	0.195 6	0.195 6	0.195 6	0.056 7
1	4.385 8	4.360 6	1.919 2	0.304 9	0.285 2	0.285 2	0.285 2	0.209 1
2	4.353 7	4.355 9	1.814 9	0.226 6	0.285 6	0.285 6	0.285 6	0.295 6
3	4.432 3	4.353 5	4.891 1	0.260 2	0.301 6	0.301 6	0.301 6	0.400 3
4	4.432 3	3.954 0	4.859 4	0.248 2	0.301 9	0.317 8	0.317 8	0.518 2
5	4.432 3	2.645 2	2.417 9	0.254 1	0.301 9	0.301 9	0.301 9	0.645 8
6	4.432 3	2.831 2	0.515 6	3.482 0	0.304 1	0.304 1	0.304 1	0.779 9
7	4.432 3	2.633 6	0.802 6	1.506 2	0.319 8	0.319 8	0.319 8	0.918 5
8	4.432 3	2.653 2	2.647 1	0.275 4	0.319 8	0.319 8	0.319 8	17.471

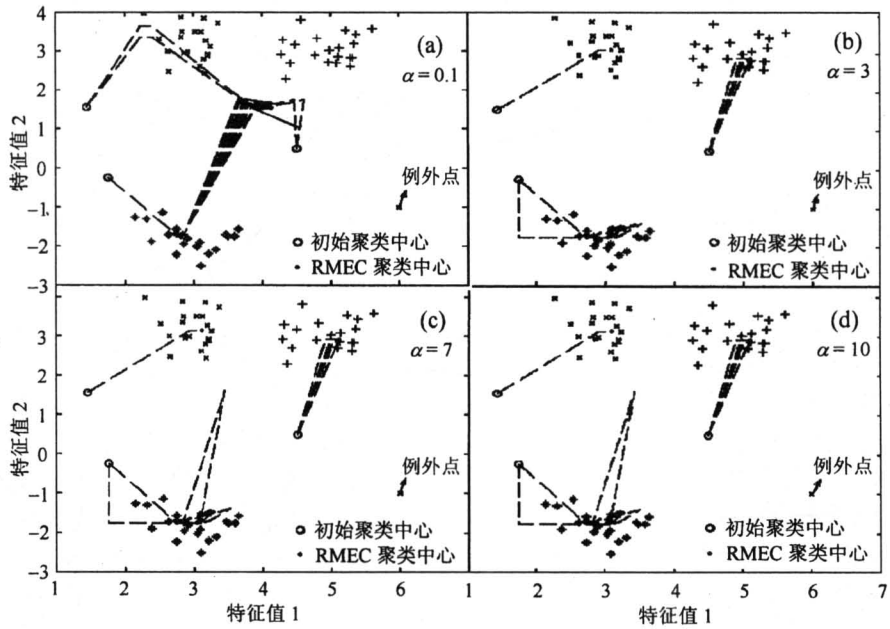


图 3 例外点个数分别为 1～8，α = 0.1，3，7，10 时 RMEC 算法的聚类效果

Fig.3 The clustering results of RMEC algorithm with different α (α = 0.1, 3, 7, 10)

实验 3 图 4a 是中国的一个城市的卫星灰度图像，通过 RMEC 聚类算法可对其进行分割并检测图像中的例外点信息，即乌云。其过程如下：第一步，以图像像素点的灰度值为特征向量构造数据集；第二步，令 $\gamma = 5$，$q = 2$，$W = 200$，$\alpha = 3$，$w(0) = 11$，用 RMEC 算法对数据集进行聚类，以聚类结果对图像进行分割的效果如图 4b 所示；第三步，利用聚类后得到的各数据的权重因子检测例外点，图 4c 示出了 RMEC 聚类算法检测出的例外点在图 4a 中的分布，这些例外点表示图像中存在的乌云信息。

从实验 3 可以看出，RMEC 算法在图像处理中也能得到很好的应用。可以预测，RMEC 算法亦可以应用于医学图像分析，数据挖掘等领域。

a. 卫星图像　　b. 分割效果　　c. 例外点信息（乌云）

图 4 卫星图像的 RMEC 算法分割结果及检测出的例外点信息（乌云）

Fig.4 The result of RMEC algorithm on the satellite image segmentation and the labeled outlier information (clouds)

5 结语

针对极大熵聚类算法 MEC 对例外点较敏感和不具备检测例外点能力的缺陷，提出了一种改进的极大熵聚类算法 RMEC。RMEC 算法不但较传统的 MEC 算法对例外点有更好的鲁棒性，而且还能

利用学习后的权重因子检测出数据集中存在的例外点，根据检测出的例外点，可以发现一些重要的信息。下一步工作，将对 RMEC 算法做更深入的研究，以期进一步有效地应用于图像处理，数据挖掘，建模等领域。

参考文献

[1] Rose K, Gurewtiz E, Fox G. A deterministic annealing approach toclustering [J]. Pattern Recognition Letters, 1990, 11: 589～594

[2] 邓赵红，陆介平，王士同. 改进的 Min-Max 模糊神经网络与函数建模 [J]. 江南大学学报，2003，2 (3)：234～239

[3] Keller A. Fuzzy clustering with outliers [A]. NAFIPS00 [M]. 2000

[4] Karayiannis N B. MECA: maximum entropy clustering algorithm [A]. Proc on IEEE Int Conf on Fuzzy Syst [C]. Orlando, F L, 1994. 630～635

[5] Li R P, Mukaidono M. A maximum entropy approach to fuzzy clustering [A]. Proc on IEEE Int Conf Fuzzy Syst[C]. Yokohama, Japan, 1995. 2227～2232

[6] 张志华，郑南宁，史　罡. 极大熵聚类算法及其全局收敛性分析 [J]. 中国科学，E 辑，2001，31(1)：59～70

[7] Las M, Kandel A. Automated perceptions in data mining [A]. Proceedings of the Eighth International Conference on Fuzzy System [C]. Seoul, Korea, 1999. 190～197

[8] Mendenhall W, Reinmuth J E, Beaver R J, Statistics for management and economics [M]. Belmont, C A: Duxbury Press, 1993

[9] Huber P J, Robust statistics [M]. New York: Wiley, 1981

[10] Gill P E, Murray W. Wright M H, Practical Optimization [M]. New York: Academic Press, 1981

[11] 王士同. 神经模糊系统及其应用 [M]. 北京：北京航空航天大学出版社，1998

[12] Steve R G, Support vector machines classification and regression [R]. University of Southampton, 1998

[13] Vapnik V, Statistical learning theory [M]. New York: Wiley, 1998

[14] Leski J, Towards a robust fuzzy clustering [J]. Fuzzy Sets and Systems, 2003, (2), 215～233

Robust Maximum Entropy Clustering Algorithm RMEC and Its Outlier Labeling

Deng Zhaohong[1,2], Wang Shitong[1,2,4], Wu Xisheng[1,2], Hu Dewen[3]

(1. *School of Information Engineering, Southern Yangtze University, Wuxi, Jiangsu* 214036, *China*; 2. *National Key Lab. Of Novel Software Technologies at Nanjing University, Nanjing* 210016, *China*; 3. *School of Automation, National Defense University of Science and Technology, Changsha* 410073, *China*; 4. *Dept. Computer, Nanjing University of Science and Technology, Nanjing* 212000, *China*)

[**Abstract**] In this paper, the novel robust maximum entropy clustering algorithm RMEC, as the improved version of the maximum entropy algorithm MEC, is presented to overcome its drawbacks: very sensitive to outliers and uneasy to label them. With the introduction of Vapnik's ε-insensitive loss function and the new weight factors, the new objective function is re-constructed, and consequently, its new update rules are derived according to the Lagrangian optimization theory. Compared with algorithm MEC, the main contributions of algorithm RMEC exist in its much better robustness for outliers and the fact that it can effectively label outliers in the dataset using the obtained weight factors. The experimental results demonstrate its superior performance in enhancing the robustness and labeling outliers in the dataset.

[**Key words**] entropy; clustering; robustness; outliers; ε-insensitive loss function; weight factors