聚类分析中竞争学习的一种新算法 [1]

魏立梅 谢维信 *

(西安电子科技大学电子工程学院 202 室 西安 710071)
*(深圳大学校长办公室 深圳 518060)

摘 要 分析指出 RPCL 算法的不足，提出一种竞争学习新算法．新算法引入数据点的密度定义，在权值的调整中考虑了数据集的几何结构对权值调整的影响，克服了 RPCL 算法的不足．理论分析与实验表明：新算法不仅可以自动确定数据集的类数，而且提高了聚类准确性和收敛速度．

关键词 聚类分析，竞争学习，密度
中图号 TP391.4

1 引 言

聚类分析作为一种无监督模式识别方式，已经广泛应用于图象处理和计算机视觉等领域．竞争学习算法以并行方式实现聚类，能够大幅度地提高聚类速度．因此，竞争学习算法的研究在聚类分析中倍受关注 [1−4]．

RPCL(Rival Penalized Competitive Learning) 算法是一种性能优良的竞争学习算法，能够自动确定数据的类数 [4]．但是，RPCL 算法也存在不足之处．本文分析指出 RPCL 算法的不足，提出竞争学习新算法．并通过实验证明新算法克服了 RPCL 算法的不足，不仅能够自动确定数据的类数，而且提高了聚类的准确性和收敛速度．

2 RPCL 算法的介绍与分析

设数据集 $X=\{x_1,x_2,\cdots,x_k,\cdots,x_N\}$, N 是 X 中元素总数．X 中的第 k 个元素 x_k 是一个 p 维矢量，$x_k=(x_{k1},x_{k2},\cdots,x_{kp})$．RPCL 算法中有 c 个节点，相应的 c 个权矢量为 w_i, $i=1,\cdots,c$．第 i 个权矢量 w_i 是一个 p 维矢量，$w_i=(w_{i1},w_{i2},\cdots,w_{ip})$．每个节点代表一个预先设置的类别，节点的权矢量表示该类的类中心．节点 i 的输出为 u_i, $u_i\in\{-1,0,1\}$．权矢量 w_i 的调整频率定义为

$$\gamma_i=m_i\Big/\sum_{j=1}^{c}m_j,\quad m_j\text{表示 } u_j=1 \text{ 的累加次数.} \tag{1}$$

当输入数据 x_k 时，RPCL 算法中各节点的输出按照 (2) 式计算．式中 w_s 被称为竞争中的赢者，w_r 被称为竞争中的次赢者．

$$u_i=1,\quad \text{当}i=s,\gamma_s\|x_k-w_s\|=\min_{j=1}^{c}\gamma_j\|x_k-w_j\|; \tag{2a}$$

$$u_i=-1,\quad \text{当}i=r,\gamma_r\|x_k-w_r\|=\min_{j=1,j\neq s}^{c}\gamma_j\|x_k-w_j\|; \tag{2b}$$

$$u_i=0,\quad \text{其它.} \tag{2c}$$

[1] 1998-03-16 收到，1999-02-18 定稿

各节点权矢量的调整公式为 (3) 式，式中 α_i 是权矢量 w_i 的学习率， β_i 是 w_i 的遗忘率。 $\alpha_i > 0, \beta_i > 0$ 。

$$\text{当 } u_i = 1 \text{ 时}, \quad \Delta w_i = \alpha_i(x_k - w_i); \tag{3a}$$

$$\text{当 } u_i = -1 \text{ 时}, \quad \Delta w_i = -\beta_i(x_k - w_i); \tag{3b}$$

$$\text{当 } u_i = 0 \text{ 时}, \quad \Delta w_i = 0. \tag{3c}$$

从 (2a) 式和 (3a) 式可以看到，数据对竞争中的赢者施加吸引力，使调整后的赢者更靠近它。从 (2b) 式和 (3b) 式可以看到，数据对竞争中的次赢者施加斥力，使调整后的次赢者远离它。从 (2c) 式和 (3c) 式可以看到，数据对其它节点的权矢量没有作用力。

RPCL 中数据将赢者吸引过来的同时，将次赢者推开。从宏观上看，每个类别只将一个权矢量吸引向它的类中心，而以斥力阻止较近的权矢量再向它靠近。所以， RPCL 算法能够自动地确定数据集的类数。

RPCL 算法存在一个缺陷：权值的调整中，只考虑了输入数据和权矢量之间的相对位置 $(x_k - w_i)$ 对权值调整的作用，却没有考虑数据集的几何结构对权值调整的影响。

从本质上讲，权值的调整过程就是数据点对赢者和次赢者产生作用力，使它们发生位移的过程。赢者和次赢者的位移，不仅应该与数据点和它们之间的相对位置有关，也应该与数据点的几何位置有关。当数据点位于某个类中心附近时，赢者在数据点吸引力作用下的位移要大，这样赢者才能尽快地向类中心收敛；次赢者在数据点斥力作用下的位移也要大，这样次赢者才很难靠近赢者对应的类中心。当数据点位于某个类边缘时，赢者在数据点吸引力作用下的位移要小，以免赢者的运动方向偏离类中心；次赢者在数据点斥力作用下的位移也要小。

赢者和次赢者的位移只有满足上述条件，赢者在向某个类收敛时，才能不受边缘数据点的干扰，迅速向类中心收敛。而一旦已经有一个权矢量收敛于某个类中心时，该类中心就会以很强的斥力阻止第二个权靠近。因此，考虑数据点的几何位置对权值调整的作用，能够提高算法的收敛速度和聚类的准确性。据此，本文提出新算法。

3 新算法

新算法的关键是如何确定数据点相对于类中心的位置。类中心附近的数据密度大，类边缘区域的数据密度小。因此，新算法用数据点的密度来近似衡量数据点相对于类中心的位置。数据点 x_k 的密度定义如下：

$$d_k = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{1+\|x_j - x_k\|/r}, \qquad r\text{为半径}, \tag{4}$$

参数 r 是反映类的致密度的量。如果数据构成的类很致密， r 就取得小一些；如果类很疏松， r 就取得大一些。

也可以按下面方法简单计算数据点的密度：

$$\text{当 } \|x_j - x_k\| \le r \text{ 时}, \text{ 令 } \|x_j - x_k\| = 0; \tag{5a}$$

$$\text{当 } \|x_j - x_k\| > r \text{ 时}, \text{ 令 } \|x_j - x_k\| = +\infty; \tag{5b}$$

$$\text{则 } d_k = \frac{1}{N}\sum_{j=1}^{N}\frac{1}{1+\|x_j - x_k\|/r} = \text{距离 } x_k \text{ 不大于 } r \text{ 的数据点的数目除以 } N \text{ 。} \tag{5c}$$

新算法的步骤如下：

第 1 步　给定半径 r，计算各数据点的密度 d_k.

第 2 步　初始化各节点的权矢量 w_i.

第 3 步　依次输入数据，按 (2) 式确定各节点的输出，按照 (6) 式确定各节点的新权值.

$$\text{当 } u_i = 1 \text{ 时}, \quad \Delta w_i = \alpha_i d_k (x_k - w_i); \tag{6a}$$

$$\text{当 } u_i = -1 \text{ 时}, \quad \Delta w_i = -\beta_i d_k (x_k - w_i); \tag{6b}$$

$$\text{当 } u_i = 0 \text{ 时}, \quad \Delta w_i = 0. \tag{6c}$$

第 4 步　判断是否满足迭代结束条件. 不满足就返回第三步.

通常，我们可以比较连续两次迭代得到的分类结果，当两次的分类结果一样时，就停止迭代；否则，就继续迭代. 也可以预先设置迭代的最大次数，或者让学习率和遗忘率随时间衰减到 0，使算法自动收敛.

4 实验结果

文献 [4] 通过实验验证 RPCL 算法能够自动确定数据集类数的能力. 实验中，数据集由四个高斯分布的类别构成，数据的维数 $p = 2$. 本文也用类似的实验验证新算法不仅能够自动确定数据集的类数，而且提高了聚类准确性和收敛速度.

本文的数据集由五个高斯分布的类别构成，数据的维数 $p = 2$. 五个类的类中心分别是 (5, 5)、(5, 15)、(15, 15)、(15, 5) 和 (10, 10). 五个类的方差分别为 2.5、2.5、2.5、2.5 和 1.2. 数据总数为 500，每类数据量为 100.

实验中取学习率 $\alpha = 0.01$、遗忘率 $\beta = 0.001$、密度半径 $r = 0.1$. 按 (5) 式计算数据点的密度.

我们分别取数据集的类数 $c = 5$、$c = 7$、$c = 8$. 对于每个 c 值作 20 次实验. 每次实验中，随机初始化 c 个权矢量 $w_i, i = 1, \cdots, c$. 数据输入的顺序也是随机的. 记录每次实验中 RPCL 算法和新算法的聚类结果和收敛所需的迭代次数 (全部数据输入一次按迭代一次计算). 实验结果如下：

$c = 5$ 时，RPCL 和新算法在 20 次实验中都得到正确的聚类结果. 收敛时，新算法的平均迭代次数为 3.5 次，RPCL 算法的平均迭代次数为 6 次.

$c = 7$ 时，RPCL 算法只在 18 次实验中准确地确定了类数，得到正确的聚类结果. 新算法在 20 次实验中都准确地确定了数据的类数，得到正确的聚类结果. 在两种算法同时得到正确结果的 18 次实验中，新算法收敛速度快，见图 1(a).

$c = 8$ 时，RPCL 算法的性能有较大的降低. 20 次实验中，RPCL 正确确定类数的只有 10 次. 而新算法又都准确确定了数据类数，得到正确结果. 在同样得到正确结果的 10次实验中，新算法的速度要大大快于 RPCL 算法，见图 1(b).

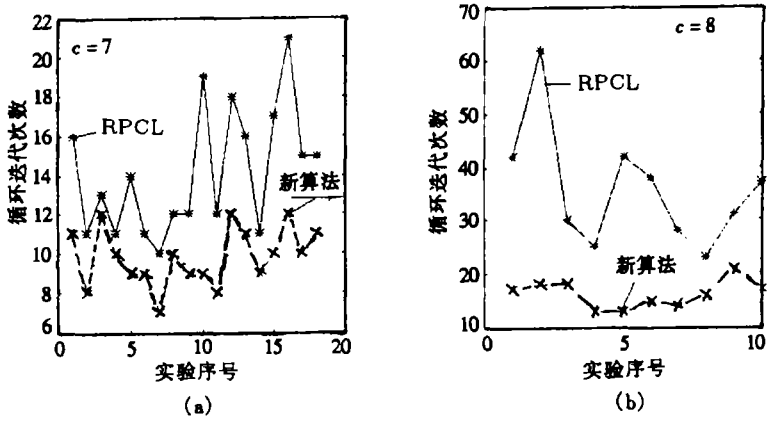

图 1 新算法和 RPCL 算法收敛速度的比较

本文还给出 $c = 8$ 时，RPCL 得到错误的聚类结果，而新算法得到正确结果的例子，见图 2 至图 4 .

实验中数据的维数 $p = 2$，因此每个数据都与二维平面中的一个点对应. 我们在该点处画个小圆圈，用这个小圆圈表示数据在二维平面中的位置. 图 2 给出了二维平面中数据的分布. 各节点的权矢量也是二维矢量，本例中它们在二维平面中的初始化位置就用图 2 中的大圆环表示. 从图 2 可以看到：数据构成五个类别.

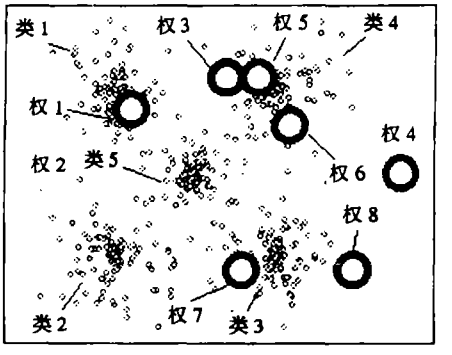

图 2 数据的分布和各节点权矢量的初始化位置

对于图 2 给出的数据集和各节点权矢量的初始值，RPCL 算法得到的聚类结果见图 3，新算法的聚类结果见图 4 .

图 3 中的大圆环仍旧表示各节点权矢量的初始化位置，从圆环中心出发的曲线表示权矢量的迭代轨迹. 图 3(a) 是 RPCL 算法的聚类结果，不同类别的数据用不同符号表示，权矢量的迭代轨迹叠加在聚类结果上. 图 3(b) 中则只标明了 RPCL 算法中八个权矢量的初始化位置和迭代轨迹，图中的五个小圆圈表示五个类别的类中心.

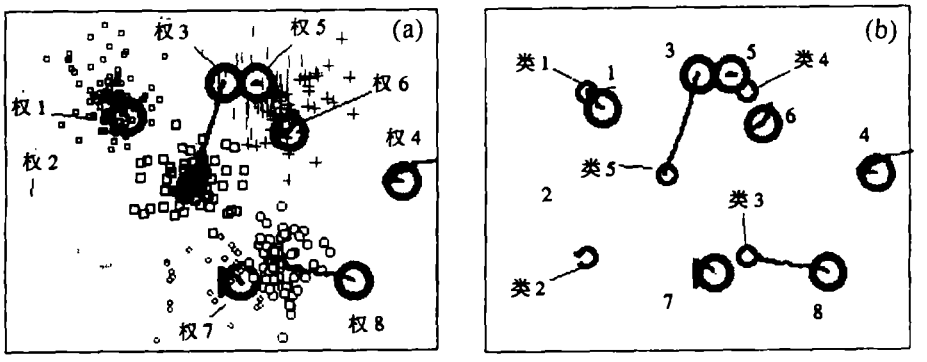

图 3　$c = 8$ 时，RPCL 算法的聚类结果
(a) 错误的分类结果，　　(b) 权矢量的迭代轨迹

对照图 3(a) 和图 3(b) 可以得到：聚类结束时，权 1 、权 2 和权 3 分别收敛于类 1 、类 2 和类 5 的类中心；权 4 被推出数据区域；而权 7 和权 8 则同时收敛于类 3 ，权 5 和权 6 也同时收敛于类 4 。所以，聚类结束时，RPCL 算法将数据分成了七类，而实际上数据集只包含五类数据。

图 4 给出了新算法的正确聚类结果。对照图 4(a) 和图 4(b) 可以得到：聚类结束时，权 4 、权 6 和权 7 被推出数据区域；权 1 、权 2 、权 3 、权 5 和权 8 分别收敛于各类别的类中心；数据被正确地分成五类。因此，新算法不仅得到了正确的聚类结果，而且正确地确定了数据集的类数。

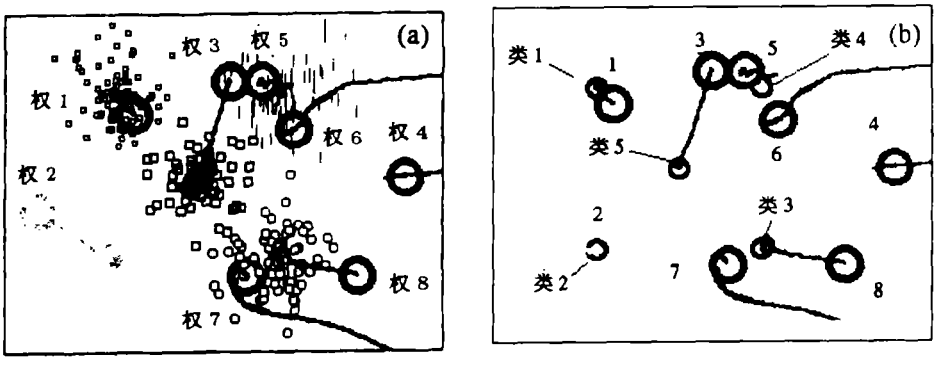

图 4　$c = 8$ 时，新算法的聚类结果
(a) 正确的分类结果，　　(b) 权矢量的迭代轨迹

5 结束语

本文在分析 RPCL 算法不足的基础上，提出一种竞争学习新算法。理论分析和实验表明：新算法能够自动确定数据集的类数，提高了收敛速度和聚类的准确性。

需要指出：本文算法对超球形分布的数据集是有效的。当数据集中的数据并不是呈超球形分布时，如：数据呈超椭球形分布，或数据集中存在两类咬合的情况 (见图 5) ，如何解决这样数据集的聚类问题是我们今后研究的任务。

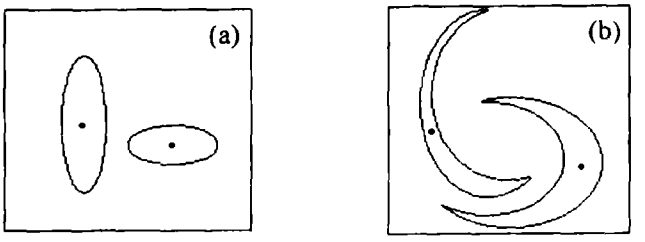

图 5 非球形分布数据集示意图
(a) 超椭球形情况, (b) 两类咬合的情况

参 考 文 献

[1] Rumelhart D E, Zipser D. Feature discovery by competitive learning. Cognitive Science, 1985, 9(1): 75–112.

[2] Grossberg S. Competitive learning: from iterative activation to adaptive resonance. Cognitive Science, 1987, 11(1): 23–63.

[3] Ahalt S C, Krishnamurty A K, Chen P, Meltion D E. Competitive learning algorithms for vector quantization. Neural Networks, 1990, 3(2): 277–291.

[4] Xu L, Krzyzak A, Oja E. Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. IEEE Trans. on Neural Networks, 1993, 4(4): 636-649.

A NEW COMPETITIVE LEARNING ALGORITHM FOR CLUSTERING ANALYSIS

Wei Limei Xie Weixin*

(*Lab 202, School of Electronic Engineering, Xidian University, Xi'an* 710071)

*(*President Office, Shenzhen University, Shenzhen* 518060)

Abstract Based on the analysis of the defect of the RPCL, a new competitive learning algorithm is proposed. In the new algorithm the data density is introduced, and the modification of the weights is taken into account to surmount the defect of the RPCL. It is shown by the theoretical analysis and experimental results that the new algorithm can automatically select the appropriate number of the clusters in a data set, and improve the clustering accuracy and convergence speed.

Key words Clustering analysis, Competitive learning, Density

魏立梅: 女, 1970 年生, 博士, 主要研究方向为: 神经网络、模糊聚类、计算机视觉、图像处理、模糊控制.

谢维信: 男, 1941 年生, 博士生导师、教授. 现任深圳大学校长, 主要研究方向包括: 神经网络、模糊聚类、计算机视觉、模式识别、图像处理、模糊控制.