

基于样本空间分布密度的改进次胜者受罚竞争学习算法

谢娟英^{1,2*}, 郭文娟¹, 谢维信^{2,3}, 高新波²

(1. 陕西师范大学 计算机科学学院, 西安 710062;

2. 西安电子科技大学 电子工程学院, 西安 710071; 3. 深圳大学 信息工程学院, 广东 深圳 518060)

(* 通信作者电子邮箱 xiejuan@snnu.edu.cn; juanyingxie@gmail.com)

摘要:针对传统次胜者受罚竞争学习(RPCL)算法忽略数据集几何结构对节点权值调整的影响,以及魏立梅等提出的新RPCL算法(魏立梅, 谢维信. 聚类分析中竞争学习的一种新算法. 电子科学学刊, 2000, 22(1): 13-18)引入密度来对节点的权值进行调整时,密度定义的主观性,提出基于样本空间分布密度的改进RPCL算法。该算法根据数据集样本自然分布定义样本密度,将此密度引入RPCL节点权值调整;使用UCI机器学习数据库数据集以及随机生成的带有噪声点的人工模拟数据集对算法进行实验测试,对算法确定数据集类簇数目的准确率、运行时间、聚类误差平方和、聚类结果的Rand指数、Jaccard系数以及Adjust Rand index参数进行分析比较。各项实验结果显示:所提算法优于原始RPCL算法和魏立梅算法,具有更好的聚类效果,对噪声数据有很强的抗干扰性能。所提算法不仅能根据样本的自然分布确定数据集的合理类簇数目,而且能确定合适的类簇中心,提高聚类的准确性,使聚类结果尽可能快地收敛到全局最优解。

关键词:聚类;次胜者受罚竞争学习算法;样本密度;聚类数目;聚类中心

中图分类号: TP18; TP301.6 **文献标志码:** A

Improvement rival penalized competitive learning algorithm based on pattern distribution of samples

XIE Juan-ying^{1,2*}, GUO Wen-juan¹, XIE Wei-xin^{2,3}, GAO Xin-bo²

(1. School of Computer Science, Shaanxi Normal University, Xi'an Shaanxi 710062, China;

2. School of Electronic Engineering, Xi'an University, Xi'an Shaanxi 710071, China;

3. School of Information Engineering, Shenzhen University, Shenzhen Guangdong 518060, China)

Abstract: The original Rival Penalized Competitive Learning (RPCL) algorithm ignores the influence of the geometry structure of a dataset on the weight variation of its nodes. A new RPCL algorithm proposed by Wei Limei *et al.* (WEI LIMEI, XIE WEIXIN. A new competitive learning algorithm for clustering analysis. Journal of Electronics, 2000, 22(1): 13-18) overcame the drawback of the original RPCL by introducing the density of samples to adjust the weights of nodes, while the density was not much objective. This paper defined a new density for a sample according to the pattern distribution of samples in a dataset, and introduced the density into the adjusting for the weights of nodes in RPCL to overcome the disadvantages of the available RPCL algorithms. The authors' improved RPCL algorithm was tested on some well-known datasets from UCI machine learning repository and on some synthetic data sets with noisy samples. The accuracy of determining the number of clusters of a dataset and the run time and the clustering error of the algorithms were compared. The Rand index, the Jaccard coefficient and the Adjust Rand index were used to analyze the performance of the algorithms. The experimental results show that the improved RPCL algorithm outperforms the original RPCL and the new RPCL proposed by WEI LIMEI *et al.* greatly, and achieves much better clustering results and has a stronger anti-interference performance for noisy data than that of the other two RPCL algorithms. All the analyses demonstrate that the improved RPCL algorithm can not only determine the right number of clusters for a dataset according to its sample distribution, but also uncover the suitable centers of clusters and advance the clustering accuracy as well as approximate the global optimal clustering result as fast as possible.

Key words: clustering; Rival Penalized Competitive Learning (RPCL) algorithm; sample density; cluster number; cluster center

0 引言

聚类分析作为无指导的学习方法是模式识别、机器学习、数据挖掘中的重要研究内容^[1-2]。聚类将物理或抽象对象按照一定的相似性度量准则划分为若干类簇,使得同一个类簇

中的对象之间具有较高的相似度,而不同类簇的对象间相似度很小^[3]。

次胜者受罚竞争学习(Rival Penalized Competitive Learning, RPCL)算法由Xu等^[4]于1993年提出,是一种性能优良的竞争学习算法,能够自动确定数据集的类簇数^[5-6],实

收稿日期: 2011-09-11; 修回日期: 2011-11-24。

基金项目: 中央高校基本科研业务费专项资金资助项目(GK200901006, GK201001003); 陕西省自然科学基金基础研究计划项目(2010JM3004)。

作者简介: 谢娟英(1971-),女,陕西西安人,副教授,CCF会员,主要研究方向:智能信息处理、模式识别、机器学习、数据挖掘; 郭文娟(1986-),女,甘肃武威人,硕士研究生,主要研究方向:智能信息处理、模式识别; 谢维信(1941-),男,广东广州人,教授,博士生导师,主要研究方向:智能信息处理、目标识别、智能人机交互、图像处理、模式识别; 高新波(1972-),男,山东莱芜人,教授,博士生导师,主要研究方向:机器学习、计算智能、视觉信息、无线通信。

现无监督学习,即聚类。然而原始 RPCL 算法在节点权值调整中,没有考虑数据集几何结构对节点权值调整的影响。魏立梅等^[6]引入数据密度来调整权值,而该密度的定义有一定主观性。本文根据数据集的自然分布信息定义了样本密度,并将该密度引入 RPCL 算法的节点权值调整,提出一种基于样本空间分布密度的改进 RPCL 算法,以克服现有 RPCL 算法的不足。经过 UCI 机器学习数据库数据集以及随机生成的带有噪声点的人工模拟数据集的实验测试,证明本文算法最优,同时对噪声数据有很强的抗干扰性能。

1 传统 RPCL 算法及其缺陷分析

RPCL 算法的基本思想如下:

设数据集 $X = \{x_1, x_2, \dots, x_j, \dots, x_n\}$, n 是 X 中数据对象总数。 X 中的第 j 个数据对象 x_j 是一个 p 维矢量, $x_j = (x_{j1}, x_{j2}, \dots, x_{jp})$ 。RPCL 算法中有 q 个节点,相应的有 q 个权矢量 ω_i ($i = 1, 2, \dots, q$)。第 i 个权矢量 ω_i 是一个 p 维矢量, $\omega_i = (\omega_{i1}, \omega_{i2}, \dots, \omega_{ip})$ 。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 ω_i 的调整频率定义为:

$$\gamma_i = \frac{m_i}{\sum_{j=1}^q m_j} \quad (1)$$

其中: m_j 表示 $u_j = 1$ 的累加次数。

当输入数据 x_j 时, RPCL 算法中各节点的输出按照式(2)计算:

$$u_i = \begin{cases} 1, & i = s, \gamma_s \|x_j - \omega_s\| = \min_{c=1}^q \gamma_c \|x_j - \omega_c\| \\ -1, & i = r, \gamma_r \|x_j - \omega_r\| = \min_{c=1, c \neq s}^q \gamma_c \|x_j - \omega_c\| \\ 0, & \text{其他} \end{cases} \quad (2)$$

其中: ω_s 被称为获胜单元, ω_r 被称为次胜单元。

各节点权矢量的调整公式为:

$$\Delta \omega_i = \begin{cases} \alpha_i (x_j - \omega_i), & u_i = 1 \\ -\beta_i (x_j - \omega_i), & u_i = -1 \\ 0, & u_i = 0 \end{cases} \quad (3)$$

其中: α_i 是权矢量 ω_i 的学习率, β_i 是 ω_i 的遗忘率, $\alpha_i > 0$, $\beta_i > 0$ 。

式(2)和(3)说明, RPCL 算法中输入数据将获胜单元吸引过来的同时, 将次胜单元推开。从宏观上看, 每个类别只将一个权矢量吸引向它的类中心, 且阻止较近的权矢量再向它靠近。所以, RPCL 算法能够自动确定数据集的类簇数目。

RPCL 算法的缺陷在于: 其权值调整中, 从数据集中随机选取样本, 没有考虑数据集几何结构对权值调整的影响。当数据对象位于某个类中心附近时, 获胜单元在该数据对象吸引力作用下的位移要大, 获胜单元能尽快地向该类中心收敛; 次胜单元在该数据对象斥力作用下的位移也要大, 次胜单元很难靠近获胜单元对应的类中心。当数据对象位于某个类边缘时, 获胜单元在数据对象吸引力作用下的位移要小, 以免获胜单元偏离类中心; 次胜单元在数据对象斥力作用下的位移也要小。获胜单元和次胜单元的位移只有满足上述条件, 获胜单元在向某个类收敛时, 才能不受边缘数据的干扰, 迅速向该类中心收敛。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矢量靠近。因

此, 考虑样本数据在整个数据集中的几何位置对权值调整的作用, 能够加快算法的收敛并且提高聚类的准确性。基于此观点, 魏立梅等^[6]提出了一种能够竞争学习的 RPCL 新算法(下面称魏立梅算法), 该算法能够自动确定数据集的类数, 提高了算法收敛速度和聚类准确性。但是该算法定义数据密度时需要主观选择一些参数, 缺少客观性。

2 本文改进的 RPCL 算法

为克服传统 RPCL 算法以及魏立梅算法的不足, 本文根据数据集样本的自然分布, 定义样本的密度 $d(x_j)$, 将该密度引入到节点权值调节公式, 对各节点权矢量进行调节, 得到改进的 RPCL 算法。下面是本文改进的 RPCL 算法中相关概念的定义。

定义1 数据对象 $x = (x^1, x^2, \dots, x^p)$ 和 $y = (y^1, y^2, \dots, y^p)$ 之间的距离为:

$$d(x, y) = \sqrt{(x^1 - y^1)^2 + \dots + (x^p - y^p)^2} \quad (4)$$

定义2 数据对象 x_j 的密度 $d(x_j)$ 定义为:

$$v(x_j) = \frac{\sum_{i=1}^n d(x_i, x_j)}{\sum_{i=1}^n d(x_i, x_i)}; j = 1, 2, \dots, n \quad (5)$$

$$d(x_j) = \exp(-v(x_j)); j = 1, 2, \dots, n \quad (6)$$

定义3 本文 RPCL 算法节点权矢量调整公式修正原始 RPCL 节点权值调整式(3)为式(7)。

$$\Delta \omega_i = \begin{cases} \alpha_i d(x_j) (x_j - \omega_i), & u_i = 1 \\ -\beta_i d(x_j) (x_j - \omega_i), & u_i = -1 \\ 0, & u_i = 0 \end{cases} \quad (7)$$

式(6)定义的样本密度 $d(x_j)$, 综合考虑了整个数据集样本的原始几何分布结构。若某数据对象周围样本分布密集, 则该数据对象的密度就大。因此, 由式(7)可知本文算法能使获胜单元更快地向相应类簇的中心移动, 使次胜单元远离类簇中心。以获胜单元为中心形成的类簇中包含较多样本, 而以非获胜单元为中心形成的类簇中只包含较少数目的样本。包含样本数目较少的类称为“冗余类”, 删除冗余类可以提高聚类准确性。

本文算法描述如下:

第1步 初始化节点数目 q , 最大迭代次数 T , 删除冗余类的阈值; 根据式(6) 计算每一个样本的密度 $d(x_j)$; 初始化权矢量 ω_i ; 初始化迭代次数统计量。

第2步 依次输入各样本数据, 按式(2) 确定各节点输出; 根据式(7) 调整各节点权值。

第3步 迭代次数增1, 判断是否满足收敛条件(即是否到达预定的迭代次数), 满足转第4步; 否则转第2步。

第4步 对每一个数据对象, 找出离它最近的节点单元, 然后将它分配到相应类簇中。

第5步 测试形成的类簇中样本数目, 如果某类的样本数目与数据集样本总数之比低于阈值, 则删除该冗余类。

第6步 输出满足条件的节点数目和相应节点的权值。

第7步 分配每一个样本到最近的节点, 得到最终的聚类结果。

3 实验结果与分析

本文实验在 UCI 数据集和随机生成的人工模拟数据集两大类数据集上进行, 实验环境为 Intel 酷睿 2 6320 @ 1.86 GHz CPU, 1 GB 内存, 160 GB 硬盘, Windows XP 操作系统,

Matlab 应用软件。

3.1 UCI 机器学习数据库数据集实验

实验选用 UCI 机器学习数据库中的 9 个聚类算法测试常用的数据集对本文算法进行测试。为了测试本文算法对于大数据集的处理能力,对 Segmentation 数据集不仅选用了常用的包含 210 个样本的数据集,还选用了包括 2 310 个样本的数据全集进行模拟实验。为了区分不同大小的 Segmentation 数据集,在 Segmentation 数据集全集后标上“-test”。实验所用 UCI 数据集描述见表 1。表 1 中的数据集代码是为了后面更方便地展示本文实验结果而对各个数据集的简称。

表 1 数据集描述

数据集	数据集代码	样本数	属性数	类数
Soybean-small	D1	47	35	4
Iris	D2	150	4	3
Wine	D3	178	13	3
Segmentation	D4	210	19	7
Ionosphere	D5	351	34	2
WDBC	D6	569	30	2
Pima Indians Diabetes	D7	768	8	2
Yeast	D8	1 484	8	10
Segmentation-test	D9	2 310	19	7

最终确定的类簇数目是一种衡量聚类算法性能的指标,算法确定的类簇数目越接近数据集的真实类簇数,证明该算法越有效^[7-10]。因此首先以此来测试本文算法的性能。实验中相关参数设置为:迭代次数 T 为 80,学习率 α 为 0.01,遗忘率 β 为 0.001。对 9 个数据集分别运行原始 RPCL 算法、魏立梅算法和本文算法。实验中的初始节点数目 q 取 5 个不同的值,对每个 q 值各算法均执行 20 次,因此每种算法共执行 100 次。统计 100 次实验中正确确定聚类数目的次数,记为 $Cnumber$,用 $rate$ 表示确定聚类数目 k 的准确率, $rate$ 定义为: $rate = Cnumber/100$ 。图 1 是原始 RPCL 算法、魏立梅算法和本文算法在 9 个不同 UCI 数据集上,确定 k 值准确率的比较。

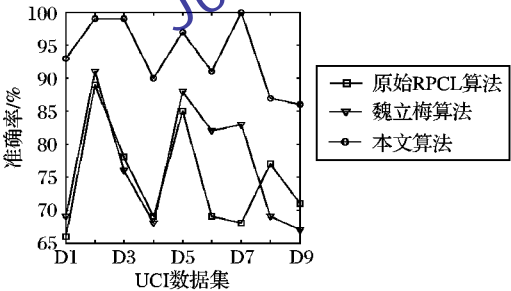


图 1 UCI 数据集上确定 k 值准确率比较

为了进一步评价本文算法的性能,下面采用常用的聚类误差平方和以及聚类时间对 3 种算法进行评价,同时还采用 Rand 指数、Jaccard 系数^[9-13],以及 Adjusted Rand index 参数^[14]对 3 种算法的聚类结果进行分析,其中后 3 个聚类评价指标都是在已知正确分类信息的前提下对聚类算法的聚类结果进行评价的有效指标。后 3 个评价指标的定义如下:设 U 和 V 分别是关于数据集的两种划分,其中 U 是已知的正确划分,而 V 是通过某种聚类算法得到的划分结果。定义 a, b, c, d 4 个参数。 a 为在 U 和 V 都在同一类的样本对数目; b 表示在 U 中为同一类,而在 V 中却不在同一类的样本对数目; c 表示在 V 中为同一类,而在 U 中却不在同一类的样本对数目; d 为在

U 和 V 都不在同一类簇的样本对数目。则 $a + b + c + d = n(n - 1)/2$, n 为数据集所含样本数,也即数据集的规模。定义 M 是所有可能的样本对,则 $M = a + b + c + d$ 。Rand 指数、Jaccard 系数和 Adjusted Rand index 参数分别定义如下。用 R 表示 Rand 指数, J 表示 Jaccard 系数, RI 表示 Adjusted Rand index 参数。

Rand 指数:

$$R = (a + d)/M$$

Jaccard 系数:

$$J = a/(a + b + c)$$

Adjusted Rand index 参数:

$$RI = \frac{2(ad - bc)}{(a + b)(b + d) + (a + c)(c + d)}$$

从定义可知,Rand 指数表示聚类结果与原始数据集样本分布的一致性;Jaccard 系数表示实现正确聚类的样本对占聚类前或后在同一类簇的样本对的比率; RI 值越大表示实现正确聚类的样本对越多,聚类效果越好(其上界为 1,则聚类结果与原始数据集的样本分布完全一致;下界为 -1,则聚类结果与原始数据集的样本分布完全不一致)。文献[15]研究指出 Adjusted Rand index 参数是最好的聚类有效性评价准则。

表 2~3 分别是 3 种算法取 5 个不同的 q 值,分别运行 20 次,每种算法共运行 100 次的平均时间比较和聚类误差平方和平均值比较。图 2~4 分别是 3 种算法运行 100 次所得聚类结果的 Adjusted Rand index 参数、Rand 指数和 Jaccard 系数平均值比较。

从图 1 可见,本文算法确定 k 值的准确率在每个数据集上都明显高于原始 RPCL 算法和魏立梅算法。

表 2 UCI 数据集上 RPCL 算法运行时间比较 s

数据集	原始 RPCL 算法	魏立梅算法	本文算法
Soybean-small	1.841	1.910	1.685
Iris	2.928	2.947	2.502
Wine	3.607	3.413	2.879
Segmentation	3.481	3.658	3.392
Ionosphere	4.780	4.880	4.615
WDBC	7.508	7.130	6.717
Pima Indians Diabetes	8.739	8.922	8.447
Yeast	17.082	17.736	17.050
Segmentation-test	24.638	25.749	25.573

表 3 UCI 数据集上 RPCL 算法聚类误差平方和比较

数据集	原始 RPCL 算法	魏立梅算法	本文算法
Soybean-small	341.43	1 875.04	263.53
Iris	143.55	86.99	82.30
Wine	4.45E+06	8.51E+07	3.78E+06
Segmentation	3.26E+06	1.06E+07	3.18E+06
Ionosphere	2.67E+03	2.68E+03	2.48E+03
WDBC	2.30E+08	7.27E+08	1.57E+08
Pima Indians Diabetes	9.79E+06	1.68E+07	5.84E+06
Yeast	56.02	54.87	54.83
Segmentation-test	2.11E+07	8.00E+07	2.12E+07

由表 2 可知:本文算法的时间性能在前 8 个数据集上优于魏立梅算法和原始 RPCL 算法,但在 Segmentation 数据集上的时间性能介于原始 RPCL 算法和魏立梅算法之间,分析原因是:本文算法因为引入合理的样本密度到节点的权值调整,从而加快了算法的收敛,所以优于魏立梅算法;但是本文

算法要为计算每个样本的密度付出时间代价,因此其收敛速度在大数据集上稍稍落后于原始 RPCL 算法。由此可见本文算法具有很好的收敛速度。

由表3可知:本文算法在前8个较小数据集上优于原始 RPCL 算法和魏立梅算法;在 Segmentation 数据全集这样的大数据集上,聚类效果明显优于魏立梅算法,但是略次于原始 RPCL 算法。由此可见,本文算法具有良好的聚类效果。

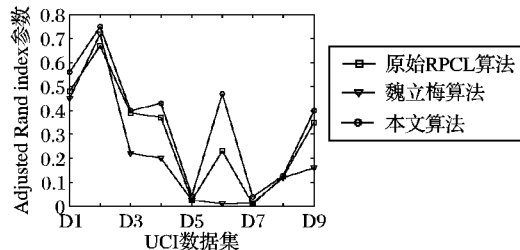


图2 UCI数据集上 Adjusted Rand index 参数的比较

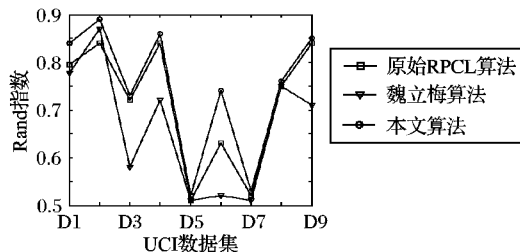


图3 UCI数据集上 Rand 指数的比较

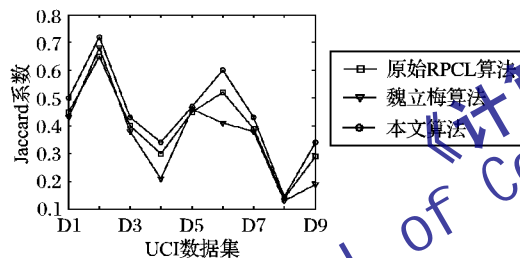


图4 UCI数据集上 Jaccard 系数的比较

通过图2~4的比较可看出:在所有数据集上,本文算法的这3个非常有效的聚类结果评价参数均最高,聚类结果更准确。因此本文算法具有更好的聚类效果。

3.2 人工模拟数据集实验

为进一步测试本文算法对噪声数据的抗干扰性能,随机生成了分别含有0%、5%、10%、15%、20%、25%、30%、35%、40%不同比例噪声的人工模拟数据集来对算法进行测试。模拟数据集包含3个类簇,每一类中含有120个二维样本,这些样本符合正态分布,其中第 i 类的横坐标 X 的均值为 μ_x^i ,纵坐标 Y 的均值为 μ_y^i ,第 i 类的标准差为 σ^i 。在第二类加入噪声点,噪声点的标准差为 σ^l 。随机生成的人工模拟数据集样本的生成参数如表4所示。实验具体参数为:迭代次数 T 为80,学习率 α 为0.01,遗忘率 β 为0.001。

表4 随机生成的带有噪声点数据集的各项参数

类簇	均值	标准差
第一类	$\mu_x^1 = 0, \mu_y^1 = 0$	$\sigma^1 = 1.5$
第二类	$\mu_x^2 = 6, \mu_y^2 = 2$	$\sigma^2 = 0.5, \sigma^l = 2$
第三类	$\mu_x^3 = 6, \mu_y^3 = -1$	$\sigma^3 = 0.5$

在随机生成的分别含有0%、5%、10%、15%、20%、25%、30%、35%、40%不同比例噪声的9个人工模拟数据集上分别运行3种算法。实验中,令 q 依次取4,5,6,7,8,每个 q 值各算

法分别执行20次,每种算法共执行100次。统计100次实验中正确确定聚类数目 k 的次数 C_{number} ,计算确定聚类数目 k 的准确率 $rate$ 。 $rate$ 定义同前所述。确定 k 值准确率的比较如图5所示。

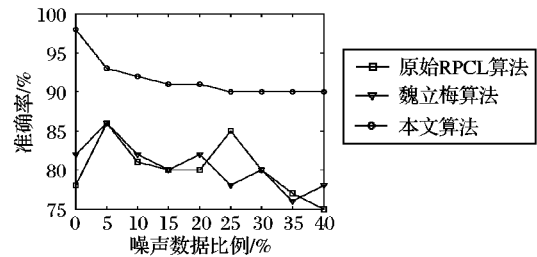


图5 人工模拟数据集上确定 k 值准确率比较

图6~7分别表示3种算法的聚类时间平均值比较和聚类误差平方和平均值比较,其中每种算法分别运行100次。图8~10分别是3种算法在各个含有不同比例噪声的人工模拟数据集上分别运行100次的 Adjusted Rand index 参数、Rand 指数和 Jaccard 系数的平均值比较。

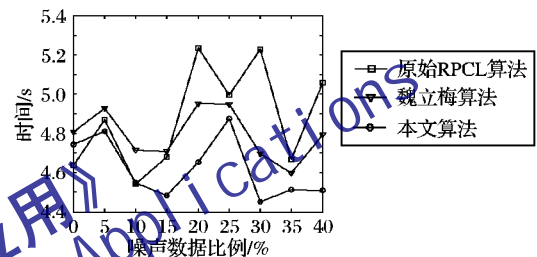


图6 人工模拟数据集上运行时间的比较

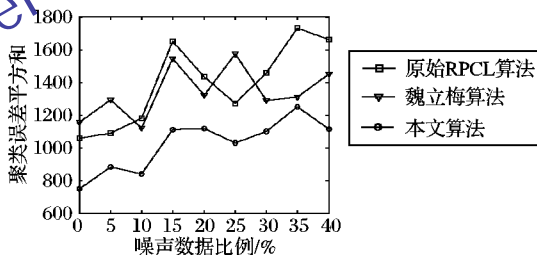


图7 人工模拟数据集上聚类误差平方和的比较

图5关于确定数据集类簇数的准确率比较显示,本文算法确定 k 值的准确率明显高于原始 RPCL 算法和魏立梅算法,具有最高的聚类准确率。

图6各个 RPCL 算法运行时间比较结果显示,本文算法在不含噪声的人工模拟数据集上的运行时间介于其他两种算法之间;但在含有噪声的人工模拟数据集上所需要的平均运行时间最短。这说明本文引入的样本密度比魏立梅算法定义的样本密度更客观、合理,加快了算法的收敛;同时为计算样本的密度需要耗费时间。因此,本文算法在含有噪声的数据集上具有最快的收敛速度,而在不含噪声的数据集上收敛速度介于其他两种算法之间。因此可以说,本文算法的时间性能优于原始 RPCL 算法和魏立梅算法。

图7聚类误差平方和比较显示,本文算法在含有不同比例噪声的人工模拟数据集上的聚类误差平方和明显优于原始 RPCL 算法和魏立梅算法。

比较图8~10可看出,本文算法的聚类结果绝对地优于原始 RPCL 算法和魏立梅算法。同时,原始 RPCL 算法的抗噪声性能优于魏立梅算法。但是魏立梅算法在不含噪声的人工模拟数据集上的聚类效果优于原始 RPCL 算法。

以上人工模拟数据集实验结果的分析显示:本文算法具

有最好的聚类效果,聚类性能绝对地优于原始 RPCL 算法和魏立梅算法。对含有噪声的人工模拟数据集可以实现最好的聚类,而不受噪声点的影响,因此该算法具有很强的抗噪声能力。

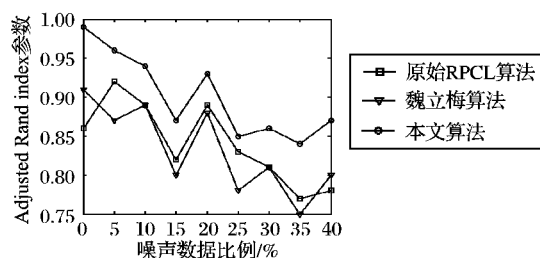


图8 人工模拟数据集上 Adjusted Rand index 参数的比较

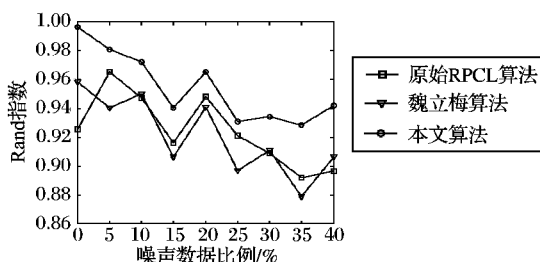


图9 人工模拟数据集上 Rand 指数的比较

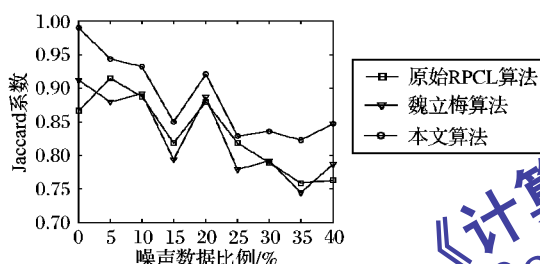


图10 人工模拟数据集上 Jaccard 系数的比较

4 结语

在分析现有 RPCL 算法不足基础上,提出一种基于样本空间分布密度的改进 RPCL 算法,引入数据集的几何结构,利用数据集样本的自然分布信息定义样本密度,将该密度引入到 RPCL 算法的节点权值调整,解决现有 RPCL 算法没有考虑数据集几何结构对节点权值调整的影响,或者考虑不足的问题。UCI 机器学习数据库数据集和随机生成的带有不同比例噪声的人工模拟数据集上的实验共同表明:本文算法能够

有效确定数据集的合适类簇数目和初始类簇中心。聚类时间、聚类误差平方和,以及 Rand 指数、Jaccard 系数和 Adjusted Rand index 参数 3 个聚类有效性指标参数的比较分析显示,本文算法收敛速度快,聚类效果好,对噪声数据有很强的抗干扰性能。不足之处是:本文算法依然是对球形数据进行分析的,关于非球形数据的分析有待进一步研究。

参考文献:

- [1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报, 2008, 19(1): 48-61.
- [2] HAN J W, KAMBER M. 数据挖掘概念与技术[M]. 范明, 孟小峰, 译. 北京: 机械工业出版社, 2000.
- [3] JAIN A K, DUBES R C. Algorithms for clustering data[M]. Upper Saddle River, NJ: Prentice Hall, 1988: 1-334.
- [4] XU L, KRZYSAK A, OJA E. Rival penalized competitive learning for clustering analysis[J]. IEEE Transactions on Neural Networks, 1993, 4(4): 636-649.
- [5] 李听,郑宇,江芳泽. 用改进的 RPCL 算法提取聚类的最佳数目[J]. 上海大学学报, 1999, 40(8): 120-122.
- [6] 魏立梅, 谢维信. 聚类分析中竞争学习的一种新算法[J]. 电子科学学刊, 2000, 22(1): 13-18.
- [7] 张忠平,王爱杰,柴旭光. 简单有效的确定聚类数目算法[J]. 计算机工程与应用, 2009, 45(15): 166-168.
- [8] 张惟皎,刘春煌,李芳玉. 聚类质量的评价方法[J]. 计算机工程, 2005, 31(20): 10-12.
- [9] 程乾生. 模糊聚类方法中的最佳聚类数的搜索范围[J]. 中国科学, 2002, 32(2): 274-280.
- [10] 王开军,李健,张军英,等. 聚类分析中类数估计方法的实验比较[J]. 计算机工程, 2008, 34(9): 198-199.
- [11] 杨善林,李永森,胡笑旋,等. K-means 算法中的 k 值优化问题研究[J]. 系统工程理论与实践, 2006(2): 97-101.
- [12] PARK H S, JUN C H. A simple and fast algorithm for K-medoids clustering[J]. Expert Systems with Applications, 2009, 36(2): 3336-3341.
- [13] 杨燕,靳蕃, KAMEL M. 聚类有效性评价综述[J]. 计算机应用研究, 2008, 25(6): 1631-1632.
- [14] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193-218.
- [15] VINH N X, EPPS J, NAILEY J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? [C]// Proceedings of the 26th International Conference on Machine Learning. New York: ACM Press, 2009: 1073-1080.

(上接第 628 页)

- [4] DERAKHSHAN R, ORLOWSKA M, LI X. RFID data management: challenges and opportunities [C]// IEEE International Conference on RFID. Piscataway, NJ: IEEE Press, 2007: 175-182.
- [5] FAZZINGA B, FLESCA S, MASCIARI E, et al. Efficient and effective RFID data warehousing [C]// IDEAS'09: Proceedings of the 2009 International Database Engineering & Applications Symposium. New York: ACM Press, 2009: 251-258.
- [6] GONZALEZ H, HAN J W, LI X L. Mining compressed commodity workflows from massive RFID data sets [C]// Proceedings of the 15th ACM International Conference on Information and Knowledge Management. New York: ACM Press, 2006: 162-171.
- [7] BAI Y, WANG F S, LIU P Y, et al. RFID data processing with a data stream query language [C]// ICDE 2007: IEEE the 23rd International Conference on Data Engineering. Piscataway, NJ: IEEE Press, 2007: 1184-1193.
- [8] GONZALEZ H, HAN J W, LI X L, et al. Warehousing and analy-

zing massive RFID data sets [C]// ICDE'06: Proceedings of the 22nd International Conference on Data Engineering. Piscataway, NJ: IEEE Press, 2006: 83-92.

- [9] 王霞. RFID 数据存储和管理技术综述[J]. 计算机应用与软件, 2008, 24(12): 175-176.
- [10] 陈竹西,孙艳,胡孔法,等. 基于路径编码的 RFID 数据压缩技术研究[J]. 扬州大学学报: 自然科学版, 2008, 11(2): 53-56.
- [11] CHAWATHE S, KRISHNAMURTHY V, RAMACHANDRAN S, et al. Managing RFID data [C]// Proceedings of the 30th Very Large Data Bases Conference. Piscataway, NJ: IEEE Press, 2004: 1189-1195.
- [12] WANG F S, LIU P Y. Temporal management of RFID data [C]// Proceedings of the 31st International Conference on Very Large Data Bases. New York: ACM Press, 2005: 1128-1139.
- [13] GONZALEZ H, HAN J W, LI X L. FlowCube: Constructing RFID FlowCubes for multi-dimensional analysis of commodity flows [C]// VLDB'06: Proceedings of the 32nd International Conference on Very Large Data Bases. New York: ACM Press, 2006: 834-845.