Journal of Computer Applications
计算机应用,2013,33(9):2482 -2485,2657
ISSN 1001-9081
CODEN JYIIDU
2013-09-01
http://www.joca.cn

文章编号:1001-9081(2013)09-2482-04
doi:10.11772/j.issn.1001-9081.2013.09.2482

基于改进流形距离 K-medoids 算法

邱兴兴*,程　霄
(九江学院 信息科学与技术学院, 江西 九江 332005)
(*通信作者电子邮箱 qiuxingxing@gmail.com)

摘　要:针对空间分布复杂的数据以及空间分布未知的现实数据聚类问题,设计了一种改进流形距离作为不相似测度。该不相似测度可有效利用所有数据点之间的全局一致性,挖掘无类属数据集的空间分布信息。通过使用该不相似测度,提出了基于改进流形距离 K-medoids 算法。将新算法与基于已有的流形距离和基于欧氏距离的 K-medoids 算法进行性能比较,对八个人工数据集以及 USPS 手写体数字识别问题的实验结果表明:新算法针对不同结构的测试数据集,在聚类性能上均优于或接近于另外两种 K-medoids 算法,并且对于各种分布的,无论简单或复杂,凸或者非凸的数据都可以进行聚类。

关键词:不相似测度;K-medoids 算法;聚类;流形距离;模式识别
中图分类号:TP181;TP391.4　　**文献标志码**:A

K-medoids algorithm based on improved manifold distance

QIU Xingxing*, CHENG Xiao
(*School of Information Science and Technology, Jiujiang University, Jiujiang Jiangxi 332005, China*)

Abstract: In this paper, an improved manifold distance based dissimilarity measure was designed to identify clusters in complex distribution and unknown reality data sets. This dissimilarity measure can mine the space distribution information of the data sets with no class labels by utilizing the global consistency between all data points. A K-medoids algorithm based on the improved manifold distance was proposed using the dissimilarity measure. The experimental results on eight artificial data sets with different structure and the USPS handwritten digit data sets indicate that the new algorithm outperforms or performs similarly to the other two K-medoids algorithms based on the existing manifold distance and Euclid distance and has the ability to identify clusters with simple or complex, convex or non-convex distribution.

Key words: dissimilarity measure; K-medoids algorithm; clustering; manifold distance; pattern recognition

0　引言

聚类就是将物理或抽象的数据对象,按照对象间的相似性进行分组或分类的过程[1]。由聚类所生成的类或簇是一组数据对象的集合,且同类中的对象彼此相似,不同类中的对象彼此相异。聚类作为一种重要的数据分析方法已经被广泛应用于模式识别、数据挖掘、信息检索和图像处理等领域。

在现有的聚类算法中,基于划分的方法以数据点与类原型之间的距离为基础,通常将聚类结果的评判标准定义为一个目标函数。类原型、数据点与类原型间距离以及目标函数定义的不同就产生了多种多样的划分聚类算法。典型的划分聚类算法有 K-means 算法[2]、K-medoids 算法[3]。不相似测度对这类算法的性能有重要影响。最简单的不相似测度是欧氏距离,但以欧氏距离作为不相似测度存在一个缺点,即它只对空间分布为球形或超球体的数据具有较好的性能,而对空间分布复杂的数据性能很差[4]。

Zhou 等[5]发现数据聚类表现出两类不同的一致性:

1)局部一致性。在空间位置上的距离接近的数据点相似度高。

2)全局一致性。在同一流形结构中的数据点相似度高。

欧氏距离可以很好地表现数据的局部一致性,却无法表现数据的全局一致性。图 1 中,数据点 a 和数据点 b 之间的欧氏距离大于数据点 a 和数据点 c 之间的欧氏距离,用欧氏距离作为不相似测度,数据点 a 和数据点 c 被划分为同一类的概率要大于数据点 a 和数据点 b 被划分为同一类的概率,而实际期望的是数据点 a 和数据点 b 被划分为同一类。因此,采用欧氏距离作为不相似测度,无法反映图 1 中所示数据的全局一致性,导致了算法对现实世界复杂分布的数据聚类性能不能令人满意。

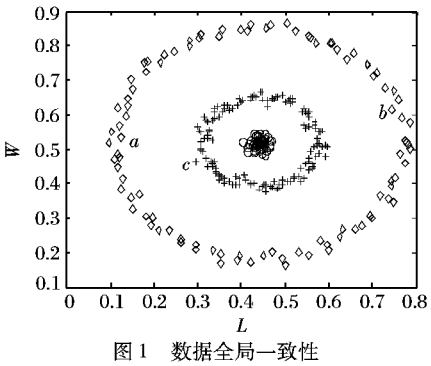


图 1　数据全局一致性

针对这一问题,Chapelle 等[6]给出一种流形距离,定义如下:

定义 1　将数据点看作是图 $G=(V,E)$ 的顶点,V 是顶点集,E 是边集。顶点序列 $p=(p_0,p_1,\cdots,p_l)$ 表示图上一条

收稿日期:2013-04-01;**修回日期**:2013-04-28。

作者简介:邱兴兴(1979 -),男,江西九江人,讲师,硕士,CCF 会员,主要研究方向:数据挖掘、进化计算;　程霄(1978 -),男,江西九江人,讲师,硕士,主要研究方向:程序设计方法学、数据挖掘、进化计算。

连接 p_0 与 p_l 的路径,其中,$p_k \in V(0 \leqslant k \leqslant l)$;$(p_k,p_{k+1}) \in E(0 \leqslant k < l)$。令 P_{ij} 表示图上连接数据点 x_i,x_j 的所有路径集,则 x_i,x_j 之间的流形距离为:

$$MD(x_i,x_j) = \min_{p \in P_{ij}} \sum_{k=0}^{l-1} (\rho^{dist(p_k,p_{k+1})} - 1) \tag{1}$$

其中:$dist(p_k,p_{k+1})$ 表示 p_k,p_{k+1} 之间的欧氏距离,$\rho > 1$ 是可调参数。

从定义可以得出,流形距离可以度量沿着流形上的最短路径,放大位于不同流形上的数据点间的距离,而缩短位于同一流形上的数据点间的距离,可以体现聚类数据的空间分布特性,表现数据的全局一致性。Wang 等[7]将该流形距离作为不相似测度成功应用于 K-means 算法,并在文献[7]指出聚类结果对可调参数 ρ 在区间 $(1,e^{18}]$ 取值不敏感。近年来,一些研究者将该测度陆续应用于特定的算法[8-15]对一些分布复杂的流形结构进行聚类分析,效果良好,但是,该测度对可调参数取值不敏感导致其在表现数据的全局一致性方面仍有缺陷。

为了在更广泛的情况下能更好地表现数据的全局一致性,本文设计一种改进流形距离作为不相似测度,并将其应用于 K-medoids,提出基于改进流形距离 K-medoids 算法(K-medoids algorithm based on improved manifold distance, IMDK-medoids)。

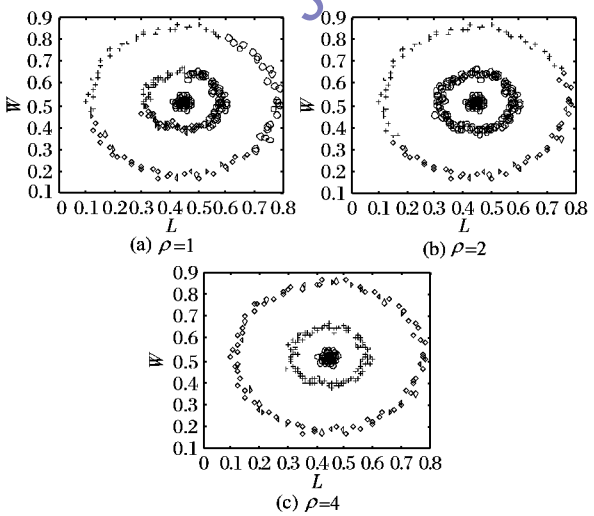
1　改进流形距离测度

改进流形距离定义如下:

定义 2　将数据点看作是图 $G = (V,E)$ 的顶点,V 是顶点集,E 是边集。顶点序列 $p = (p_0,p_1,\cdots,p_l)$ 表示图上一条连接 p_0 与 p_l 的路径,其中,$p_k \in V(0 \leqslant k \leqslant l)$;$(p_k,p_{k+1}) \in E(0 \leqslant k < l)$。令 P_{ij} 表示图上连接数据点 x_i,x_j 的所有路径集,则 x_i,x_j 之间的改进流形距离为:

$$IMD(x_i,x_j) = \min_{p \in P_{ij}} \sum_{k=0}^{l-1} dist(p_k,p_{k+1})^{\rho} \tag{2}$$

其中:$dist(p_k,p_{k+1})$ 表示 p_k、p_{k+1} 之间的欧氏距离,$\rho \geqslant 1$ 是可调参数。式(2)和式(1)的最大不同是参数 ρ 的变化对聚类结果有明显影响,因此可以通过对因子进行调整得到更好的聚类结果(如图 2 所示)。



(a) ρ=1　(b) ρ=2　(c) ρ=4

图 2　参数 ρ 的不同取值对聚类结果的影响

改进流形距离满足测度的四个条件:

1)对称性:$IMD(x_i,x_j) = IMD(x_j,x_i)$;

2)非负性:$IMD(x_i,x_j) \geqslant 0$;

3)自反性:$IMD(x_i,x_j) = 0$,当且仅当 $x_i = x_j$;

4)三角不等式:对任意 x_i,x_j,x_k,有 $IMD(x_i,x_j) \leqslant IMD(x_i,x_k) + IMD(x_k,x_j)$。

2　基于改进流形距离的 K-medoids 算法

K-means 算法和 K-medoids 算法是基于划分的聚类算法代表。K-means 算法以类中所有数据点的均值点作为每个类别的类原型,当数据为球形或超球体分布时,算法可以取得较好性能,而当数据呈现复杂分布的流形结构时,均值点可能偏离类别,用均值点作为类别的类原型,对性能会有较大影响。K-medoids 算法从当前类中选取一个到类中的其他所有数据点的距离之和最小的数据点作为类原型,可以确保类原型不会偏离类别。当数据呈现复杂分布,尤其是不符合球形或超球体的分布时,K-medoids 算法比 K-means 算法更有优势。基于以上分析,本文将改进流形距离作为不相似测度用于 K-medoids 算法,得到 IMDK-medoids 算法。

目标函数定义如下:

定义 3　给定一个数据集 $\boldsymbol{X} = \{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,将其划分为子集类 $C = \{C_1,C_2,\cdots,C_k\}(k \leqslant n)$,其中 $C_i \subseteq \boldsymbol{X}, C_i \neq \varnothing (i = 1,2,\cdots,k)$,$\bigcup_{i=1}^{k} C_k = \boldsymbol{X}$ 且 $C_i \cap C_j = \varnothing (i,j = 1,2,\cdots,k \wedge i \neq j)$,$k$ 个类原型为 $\boldsymbol{M} = \{\boldsymbol{m}_1,\boldsymbol{m}_2,\cdots,\boldsymbol{m}_k\}$。划分的目标函数为:

$$J(C,\boldsymbol{M}) = \sum_{j=1}^{k} \sum_{x_i \in C_j} D(\boldsymbol{x}_i,\boldsymbol{m}_j) \tag{3}$$

其中:$D(\boldsymbol{x}_i,\boldsymbol{m}_j)$ 表示 $\boldsymbol{x}_i$、$\boldsymbol{m}_j$ 之间的距离。

算法描述如下:

输入　数据集 $\boldsymbol{X}$;聚类数 k。

输出　数据集 $\boldsymbol{X}$ 的一个 k 划分 C。

1)初始化。对任意两个数据点 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$,计算距离 $D(\boldsymbol{x}_i,\boldsymbol{x}_j)$,随机地选择 k 个不同的数据点作为 k 个类原型。

2)将每个数据点归类到和它距离最近的类原型所代表的类,计算目标函数值 J。

3)对每个类原型 $\boldsymbol{m}$:

3.1)对每个不是类原型的数据点 $\boldsymbol{o}$:

3.1.1)将 $\boldsymbol{o}$ 替换 $\boldsymbol{m}$,重新将每个数据点归类到和它距离最近的类原型所代表的类,计算目标函数值 J'。

3.1.2)如果 $J' < J$,则以 $\boldsymbol{o}$ 替代 $\boldsymbol{m}$ 作为新的类原型,$J = J'$。

4)重复 3),直到 k 个类原型不发生变化。

其中:式(3)中的 $D(\boldsymbol{x}_i,\boldsymbol{m}_j)$ 取 $IMD(\boldsymbol{x}_i,\boldsymbol{m}_j)$,则算法是 IMDK-medoids 算法;$D(\boldsymbol{x}_i,\boldsymbol{m}_j)$ 取 $MD(\boldsymbol{x}_i,\boldsymbol{m}_j)$,则算法是 MDK-medoids 算法;$D(\boldsymbol{x}_i,\boldsymbol{m}_j)$ 取欧氏距离 $dist(\boldsymbol{x}_i,\boldsymbol{x}_j)$,则算法是 K-medoids 算法。

3　实验分析

将 IMDK-medoids 算法、MDK-medoids 算法和 K-medoids 算法进行性能比较,采用的测试问题包括 8 个人工数据集。以及现实世界的 USPS 数据集。表 1 给出了这些数据集的性质。

对于算法的聚类性能,采用错误分类比率(Error Rate,ER)

和 ARI(Adjusted Rand Index)[17]来衡量。*ARI* 计算公式如下：

$$ARI = \frac{\sum_{lk}\binom{n_{lk}}{2} - \left[\sum_{l}\binom{n_{l\cdot}}{2} \times \sum_{k}\binom{n_{\cdot k}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_{l}\binom{n_{l\cdot}}{2} + \sum_{k}\binom{n_{\cdot k}}{2}\right] - \left[\sum_{l}\binom{n_{l\cdot}}{2} \times \sum_{k}\binom{n_{\cdot k}}{2}\right]/\binom{n}{2}} \tag{4}$$

其中：n_{lk} 表示被划分到类属 l 和类属 k 的数据点的个数。$ARI \in [-1,1]$，其数值越大，说明聚类划分的正确率越高。

表 1 实验中使用的数据集

数据集	样本数	维数	类数	数据集	样本数	维数	类数
Size5	1 000	2	4	3Circles	299	2	3
Square1	1 000	2	4	2Moons	400	2	2
Square4	1 000	2	4	{0,8}	2 261	257	2
LineBlobs	266	2	3	{3,7}	1 616	257	2
Spiral	1 000	2	2	{3,5,8}	2 248	257	3
Sticks	512	2	2	{0,2,4,8}	4 042	257	4

3.1 人工数据集聚类实验

为了能直观观察 3 个算法的性能，首先将 3 个算法应用于 8 个人工数据集的聚类问题。这 8 个数据集为 Size5 数据集、Square1 数据集、Square4 数据集、LineBlobs 数据集、Spiral 数据集、Sticks 数据集、3Circles 数据集、2Moons 数据集。IMDK-medoids 算法中收缩因子 ρ 设置为 4，MDK-medoids 算法中收缩因子 ρ 设置为 e^2。

3.1.1 人工数据集 A

为了说明对于分布特点不同的数据集的聚类性能，将人工数据分为两组。人工数据集 A 为 3 个具有球形分布的数据集，分别是 Size5 数据集、Square1 数据集和 Square4 数据集。下面分别展示三个算法对于人工数据集 A 的典型聚类结果，以使数据分布情况及聚类效果更直观显现。

表 2 三个算法在求解人工数据集 A 时的性能比较

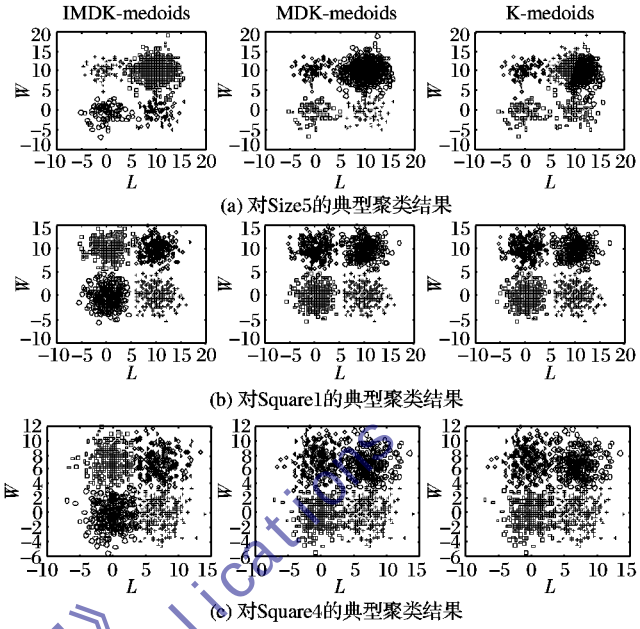
数据集	IMDK-medoids		MDK-medoids		K-medoids	
	ARI	*ER*	*ARI*	*ER*	*ARI*	*ER*
Size5	0.955 3	0.015 0	0.977 5	0.008 0	0.404 2	0.462 0
Square1	0.973 5	0.010 0	0.970 9	0.011 0	0.970 8	0.011 0
Square4	0.786 6	0.085 0	0.822 9	0.070 0	0.837 2	0.064 0

对于具有球形分布的三个数据集的聚类问题，三个算法都没有得到完全正确的结果，对比典型结果，在 Size5 数据集上，MDK-medoids 算法的结果最优，IMDK-medoids 算法和 MDK-medoids 算法的结果相差不大，而 K-medoids 算法的结果较差，原因在于 Size5 数据集虽然是球形分布，但类和类之间规模相差悬殊，结构相对比较复杂。在 Square1 数据集上，三个算法表现相当，IMDK-medoids 算法的结果最优。在 Square4 数据集上，K-medoids 算法的结果最优，IMDK-medoids 算法的结果稍差，三个算法得到的结果均是满意解(见表 2 和图 3)。

3.1.2 人工数据集 B

人工数据集 B 为五个具有复杂流形分布的数据集，分别是 LineBlobs 数据集、Spiral 数据集、Sticks 数据集、3Circles 数据集、2Moons 数据集。下面分别展示两个算法对于人工数据集 B 的典型聚类结果。

IMDK-medoids 算法在五个具有流形分布的数据集上取得了完全正确的聚类结果，K-medoids 算法的结果非常不理想，MDK-medoids 算法的结果介于两者之间，但仍然不理想(见表 3 和图 4)。这组实验说明在流形结构的数据聚类中，采用式(2)的改进流形距离可以很好地表现数据的全局一致性，而式(1)的流形距离在表现数据全局一致性上仍有缺陷，欧氏距离则无法表现数据全局一致性。



(a) 对Size5的典型聚类结果

(b) 对Square1的典型聚类结果

(c) 对Square4的典型聚类结果

图 3 三个算法对人工数据集 A 的典型聚类结果

表 3 三个算法在求解人工数据集 B 时的性能比较

数据集	IMDK-medoids		MDK-medoids		K-medoids	
	ARI	*ER*	*ARI*	*ER*	*ARI*	*ER*
LineBlobs	1	0	0.440 1	0.233 1	0.434 8	0.236 8
Spiral	1	0	0.290 9	0.230 0	0.042 3	0.396 0
Sticks	1	0	0.703 6	0.146 5	0.506 9	0.275 4
3Circles	1	0	0.076 3	0.521 7	0.071 9	0.525 1
2Moons	1	0	0.178 6	0.287 5	0.218 9	0.265 0

以上两组实验说明给出的改进流形距离可以很好地表现数据的全局一致性，将其应用于 K-medoids 算法得到的 IMDK-medoids 算法在处理球形分布和流形分布数据集的聚类问题非常有效。

3.2 真实数据集聚类实验

选择 USPS 数据集作为测试数据，将三个算法应用于手写体数字识别中。USPS 数据集取自美国 Buffalo 日常信件信封上的手写体数字，USPS 数据集由 9 298 个样本构成，其中每个样本是一个 16×16 维灰度图像。实验取全部测试样本作为聚类数据集。从中挑选{0,8}、{3,7}、{3,5,8}和{0,2,4,8}四组数据进行识别。IMDK-medoids 算法对参数取值敏感，因此收缩因子 ρ 设置为 1 ~ 10，取最佳结果。MDK-medoids 算法收缩因子 ρ 设置为 e^2。

IMDK-medoids 算法在四个真实数据集上的聚类结果最优，而 MDK-medoids 算法和 K-medoids 算法在真实数据集上的聚类结果不能令人满意(见表 4)。

表 4 三个算法在求解真实数据集是的性能比较

数据集	IMDK-medoids		MDK-medoids		K-medoids	
	ARI	*ER*	*ARI*	*ER*	*ARI*	*ER*
{0,8}	0.930 5	0.017 2	0.001 1	0.312 7	0.120 2	0.321 5
{3,7}	0.955 9	0.011 1	0.000 0	0.490 7	0.822 9	0.046 4
{3,5,8}	0.761 3	0.086 3	0.000 0	0.633 0	0.488 9	0.198 0
{0,2,4,8}	0.737 8	0.131 1	0.000 3	0.615 0	0.357 2	0.461 7

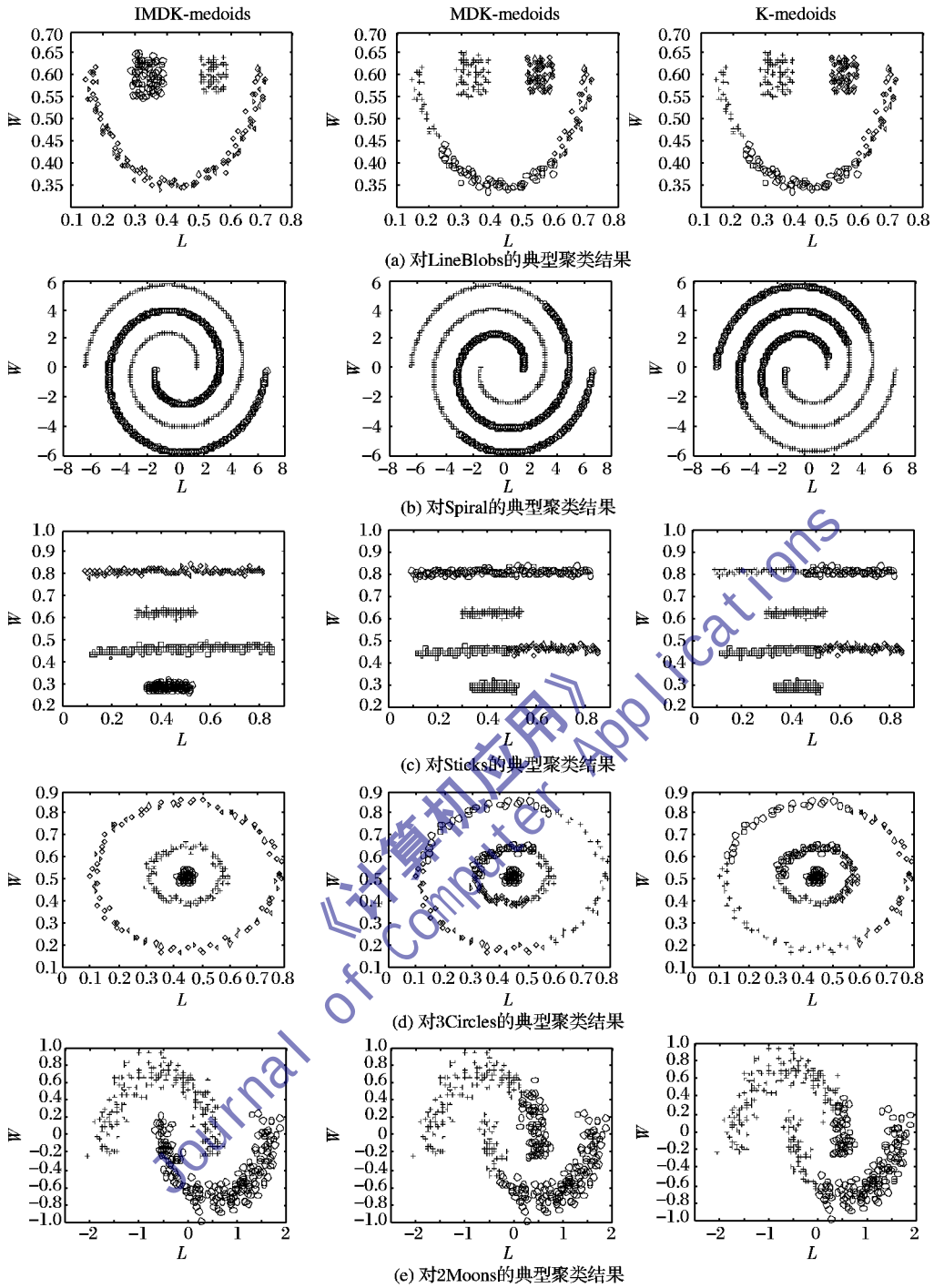


图 4　三个算法对人工数据集 B 的典型聚类结果

以上实验说明式(2)中改进流形距离在真实数据集上表现数据的全局一致性最好,IMDK-medoids 算法在实际问题中同样是可行的。

4　结语

本文设计了一种改进流形距离不相似测度,应用于 K-medoids 算法,提出 IMDK-medoids 算法具有较好的聚类性能。在不同数据集上,将 IMDK-medoids 算法和 MDK-medoids 算法、K-medoids 算法进行了对比测试,证明其对复杂分布数据聚类问题和真实世界模式识别问题的有效性。在人工数据集的测试中,IMDK-medoids 对球形分布数据的聚类结果与最佳结果接近,而对于流形分布数据的聚类效果,其优势十分明显,在真实世界的 USPS 数据集模式识别问题中,准确率同样最高,明显优于其他两个算法。

参考文献:

[1] XU R, WUNSCH D. Survey of clustering algorithms[J]. IEEE Transactions on Neural Networks, 2005, 16(3): 645 – 678.

[2] HARTIGAN J A, WONG M A. A k-means clustering algorithm[J]. Applied Statistics, 1979, 28(1): 100 – 108.

[3] KAUFMAN L, ROUSSEUW P J. Finding groups in data: an introduction to cluster analysis[M]. New York: John Wiley & Sons, 1990: 108 – 110.

[4] SU M C, CHOU C H. A modified version of the k-means algorithm with a distance based on cluster symmetry [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(6): 674 – 680.

(下转第 2657 页)

4.2　原型系统的应用

本文提出的健康度评价方法可用于事故模拟训练系统中。目前的矿山灾难性事故模拟训练系统一般由矿难环境模拟系统和智能控制系统组成[12-13],矿难环境模拟系统包括产烟控制系统、升温控制系统、火焰控制毒气释放及控制系统等,智能控制系统包括人员定位系统、视频监控系统等[14]。从矿难模拟系统和智能控制系统中可得到各个影响因子的实时值,因此可以利用健康度计算原型系统,对受训的逃生人员进行健康度评价。受训人员选择不同的逃生路径时,其健康度损失值将会根据不同路径上的影响因子的不同情况而变化,因而其健康度也会随之变化。长期训练将有利于加强受训人员对各不利因素的判断,帮助其在实际灾难中作出正确的逃生决策。

此外,在实际矿难中,引入健康度评价体系也能为救援人员的决策提供一个量值的依据。根据各个区域中的逃生人员的健康度情况,可估算各中段的人员伤亡情况,确定对哪些区域进行重点救援,同时也有助于救援人员利用通信设施指导逃生人员选择正确的逃生路线。

5　结语

本文研究了井下火灾中影响人员逃生的各个影响因子及其危害,并对逃生人员在逃生过程中的人员健康度评价方法进行了研究。该评价方法具有可扩展性,依照该方法,针对任意类型的井下灾害事故,其人员健康度评价可分为以下四步:1)分析该事故中影响人员逃生的主要影响因子,尤其是致死因素;2)建立影响因子指标体系,分析各个指标的取值范围、不同取值对人员健康的影响以及人员的临界暴露时间;3)利用模糊综合评价方法计算单个时间步长内的健康度影响值;4)根据各时间步长的健康度影响值计算出人员健康度水平随时间的变化情况。

在矿山灾难性事故模拟训练系统中,该方法可用于评价受训人员选择不同逃生路径时的健康度变化情况,从而训练作业人员的逃生意识,提高其选择正确逃生路径的能力;而救援人员也可以根据人员健康度的变化选择合适的救援措施,比如估算各中段的人员伤亡情况、选择通风策略、确定重点救援区域等。进一步的研究工作主要包括两个方面:一是结合具体的矿山实例,将该评价方法应用到具体的矿山模拟仿真系统中;二是对井下其他事故类型中的逃生人员健康度评价进行分析。

参考文献:

[1] 赵杰, 郭红卫, 施玮, 等. 过高和过低摄入必需微量元素的危险度评价[J]. 中华预防医学杂志, 2002, 36(6): 414-416.

[2] 黎强, 刘清辉, 张慧, 等. 火灾烟气中有毒气体的体积分数分布与危害[J]. 自然灾害学报, 2003, 12(3): 69-74.

[3] 赵杰. 火灾烟雾中的有毒气体[J]. 疾病控制杂志, 2003, 7(4): 338-340.

[4] ALARIE Y. Toxicity of fire smoke[J]. CRC Critical Reviews in Toxicology, 2002, 32(4): 259-289.

[5] 杨立中, 方伟峰, 邓志华, 等. 火灾中的烟气毒性研究[J]. 火灾科学, 2001, 10(1): 29-33.

[6] 孙继平. 煤矿井下紧急避险系统研究[J]. 煤炭科学技术, 2011, 39(1): 69-71.

[7] YANG L Z, FANG T Y, ZHOU X D, *et al*. Establishment of a new dynamic RRC model in smoke toxicity evaluation and engineering application[J]. Progress in Natural Science, 2005, 15(3): 262-264.

[8] 高福启. 高温热烟气在火场中的危害[J]. 吉林劳动保护, 1999, 39(4): 39.

[9] 王新民, 姚建, 彭欣. 火灾时期致命因素危害时间的研究[J]. 消防理论研究, 2005, 24(1): 28-30.

[10] 邢海涛, 王欣. 原油储罐火灾环境风险评价[J]. 城市环境与城市生态, 2004, 17(2): 18-19.

[11] 马绥华. 火灾烟雾颗粒粒径分布的测量与计算模拟[D]. 合肥: 中国科技大学, 2006.

[12] 左敏, 杜军平. 矿山灾难性事故模拟训练系统的设计与实现[J]. 金属矿山, 2006(6): 65-67.

[13] 芦江文. 矿山应急救援综合模拟演习训练系统研究与应用[D]. 西安: 西安科技大学, 2009: 1-53.

[14] ORR-T J, MALLETT-L G, MARGOLIS-K A. Enhanced fire escape training for mine workers using virtual reality simulation[J]. Mining Engineering, 2009, 61(11): 41-44.

(上接第 2485 页)

[5] ZHOU D Y, BOUSQUET O, LAL T N, *et al*. Learning with local and global consistency[C]// Proceedings of Advances in Neural Information Processing Systems 16. Cambridge: MIT Press, 2004: 321-328.

[6] CHAPELLE O, ZIEN A. Semi-supervised classification by low density separation[EB/OL]. [2012-10-10]. http://eprints.pascal-network.org/archive/00000388/01/pdf2899.pdf.

[7] WANG L, BO L F, JIAO L C. A modified k-means clustering with a density-sensitive distance metric[C]// Proceedings of the First International Conference on Rough Sets and Knowledge Technology. Berlin: Springer-Verlag, 2006: 544-551.

[8] 王玲, 薄列峰, 焦李成. 密度敏感的谱聚类[J]. 电子学报, 2007, 35(8): 1577-1581.

[9] GONG M G, JIAO L C, WANG L, *et al*. Density-sensitive evolutionary clustering[C]// Proceedings of PAKDD 2007, LNAI 4426. Berlin: Springer-Verlag, 2007: 507-514.

[10] 公茂果, 焦李成, 马文萍, 等. 基于流形距离的人工免疫无监督分类与识别算法[J]. 自动化学报, 2008, 34(3): 367-375.

[11] 潘晓英, 刘芳, 焦李成. 密度敏感的多智能体进化聚类算法[J]. 软件学报, 2010, 21(10): 2420-2431.

[12] 何飞, 梁治国, 王晓晨, 等. 基于流形距离的生产状态聚类分析[J]. 计算机应用研究, 2011, 28(9): 3242-3244.

[13] 李阳阳, 石洪竺, 焦李成, 等. 基于流形距离的量子进化聚类算法[J]. 电子学报, 2011, 39(10): 2343-2347.

[14] 公茂果, 王爽, 马萌, 等. 复杂分布数据的二阶段聚类算法[J]. 软件学报, 2011, 22(11): 2760-2772.

[15] 李岩波, 宋琼, 郭新辰. 基于流形距离的人工免疫半监督聚类算法[J]. 计算机科学, 2012, 39(11): 204-207.

[16] USPS dataset[EB/OL]. [2012-10-10]. http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html.

[17] HUBERT L, ARABIE P. Comparing partitions[J]. Journal of Classification, 1985, 2(1): 193-218.