一种基于流形距离的迭代优化聚类算法

王娜[1]，杜海峰[2]，王孙安[1]
(1. 西安交通大学机械工程学院，710049，西安；2. 西安交通大学公共管理与复杂性科学研究中心，710049，西安)

摘要：针对传统欧氏距离测度描述复杂结构的数据分布会失效的问题，引入能有效反映样本集固有的全局一致性信息的流形距离作为样本间相似度度量测度，并设计了反映类内相似度大、类间相似度小的聚类目标的准则函数，把数据聚类转化成准则函数优化问题，提出了一种迭代优化的聚类算法.通过4个人工数据集的仿真试验结果表明，新方法的参数很少且实现简单，由于实现过程中没有引入随机操作，因此结果比较确定.与标准k均值算法相比，新方法能够自动确定聚类数目，对于样本空间分布复杂的聚类问题具有良好的分类效果.

关键词：流形距离；准则函数；聚类
中图分类号：TP181 **文献标志码**：A **文章编号**：0253-987X(2009)05-0076-04

Iterative Optimization Clustering Algorithm Based on Manifold Distance

WANG Na[1], DU Haifeng[2], WANG Sun′an[1]
(1. School of Mechanical Engineering, Xi′an Jiaotong University, Xi′an 710049, China; 2. Center for Administration and Complexity Science, Xi′an Jiaotong University, Xi′an 710049, China)

Abstract: Aiming at the problem that classical Euclidean distance metric may be invalid when it is used to measure the complicated data structures, a manifold distance based on similarity metric and being able to measure the geodesic distance along the manifold is introduced, and a criterion function used to express the clustering target is designed, where the samples in the same cluster are somehow more similar than samples in different one. Accordingly, the clustering problem is converted to function optimization problem, and an iterative optimization clustering algorithm is proposed. The steps of the algorithm are discussed in detail. Simulation results on four artificial datasets with different manifold structures show that the new algorithm is more straightforward due to the less pre-defined parameters and it is a deterministic algorithm due to the lack of random operations. A comparison with k-means clustering algorithms indicates the ability to determine the cluster number automatically and identify complex non-convex clusters.

Keywords: manifold distance; criterion function; clustering

聚类，即无监督分类，是一种重要的数据分析方法，已经被广泛应用于信息检索、数据挖掘和模式识别等领域.在现有的聚类方法中，基于目标函数的聚类算法把聚类问题归结为一个优化问题，具有深厚的泛函基础，是聚类算法研究的重要分支之一，而样本之间的相似度度量以及待优化的准则函数设计就成为此类算法研究的核心问题[1].通常，样本之间的相似度度量就是样本之间的距离，最简单的相似度度量是欧氏距离，它对空间分布为球形或超球体的数据具有很好的性能，但对于空间分布复杂的流形结构的数据效果很差[2]，因此为此类数据设计更加合理的相似度度量是非常必要的工作.准则函数的设计力图反映聚类目标，即把样本分为多个类，同类中的样本具有较高的相似度，不同类中的样本差别

收稿日期：2008-10-06. 作者简介：王娜(1979—)，女，讲师. 基金项目：国家自然科学基金资助项目(50505034)；教育部博士点新教师基金资助项目(20070698022).

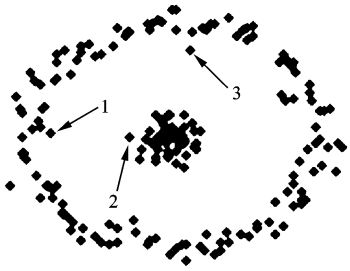
较大，简单且应用广泛的准则函数是误差平方和准则、相关的最小方差准则和散布准则[1]. 虽然这些准则在很多问题中都体现出很强的实用性，但对于复杂的数据结构（密集类被稀疏类包围或互相绞缠在一起的线条式的几个类）依然无法正确聚类[1].

针对复杂结构的数据聚类问题，本文引入基于流形距离的相似度度量并设计了新的准则函数，提出一种迭代优化准则函数的聚类算法，并将其用于 4 个具有复杂结构空间分布的人工数据聚类问题的仿真试验中，通过和标准 k 均值算法的结果进行比较，考察了新算法的性能.

1　相似度度量和准则函数

1.1　基于流形距离的相似度度量

聚类问题对应 2 个一致性假设[3]：①局部一致性，即邻近的数据点具有较高的相似度；②全局一致性，即同一流形上的数据点具有较高的相似度. 一般情况下，相似度是数据点间距离的函数，对于图 1 所示的例子，按照欧氏距离进行相似度度量时，数据点 1 和 2 的欧氏距离明显小于数据点 1 和 3 的欧氏距离，即数据点 1 和 2 的相似度高于数据点 1 和 3 的相似度，因此不能反映图 1 所示数据的全局一致性. 这也是对于复杂结构的数据分布，采用简单的采用欧氏距离作为相似度度量得不到正确聚类的原因.



1～3：数据点

图 1　无法反映样本全局一致性的欧氏距离

为了解决正确反映全局一致性的问题，新的相似度度量应该重新计算数据点之间的距离，使得数据点 1 和 2 的距离大于数据点 1 和 3 的距离，即两点间直接相连的路径长度不一定最短. 采用流形距离测度[4]可以实现这一需求. 由 n 个样本组成的集合 $X=\{x_1, x_2, \cdots, x_n\}$ 中，$l(x_i, x_j)\in \boldsymbol{L}$ 为 2 个样本 x_i 和 x_j 间的线段长度，即

$$l(x_i, x_j)=\begin{cases} d(x_i, x_j) & x_i, x_j \text{ 是邻近点} \\ \text{无穷大} & \text{否则} \end{cases} \tag{1}$$

式中：$d(x_i, x_j)$ 是样本 x_i 和 x_j 间的欧氏距离. 对每个样本与其他所有样本之间的欧氏距离进行递增排序，距离较小的前 K 个点为该样本的邻近点，K 为邻域数据点数目.

从 x_i 到 x_j 的所有路径构成集合 P_{ij}，则 x_i 与 x_j 之间的流形距离为

$$d_{\mathrm{D}}(x_i, x_j)=\min\{P_{ij}\} \tag{2}$$

新的距离测度满足测度的 4 个条件[1]：①非负性，$d_{\mathrm{D}}(x_i, x_j)\geqslant 0$；②自反性，$d_{\mathrm{D}}(x_i, x_j)=0$，当且仅当 $x_j=x_i$；③对称性，$d_{\mathrm{D}}(x_i, x_j)=d_{\mathrm{D}}(x_j, x_i)$；④三角不等式，对于任意数据点 x_k，$d_{\mathrm{D}}(x_i, x_j)\leqslant d_{\mathrm{D}}(x_i, x_k)+d_{\mathrm{D}}(x_k, x_j)$.

利用流形距离测度度量任意 2 点的距离，同一流形上的 2 点可以用许多较短的线段相连接，而不同流形上的 2 点要用较长的线段相连接，即实现放大不同流形上的数据点间的距离，而缩短同一流形上的数据点间的距离，从而同时满足 2 个一致性假设.

2 个数据点间的距离越小，它们的相似度越高. 因此，定义 x_i 和 x_j 间的相似度为

$$a_{ij}=1/(1+d_{\mathrm{D}}(x_i, x_j)) \tag{3}$$

需要说明的是，本文定义样本自身的相似度为 0，即当 $i=j$ 时，$a_{ij}=0$.

1.2　准则函数

聚类的目的在于把对象分成满足“同类对象间相似度高、不同类对象间相似度低”的若干个类，首先设计有效反映类内类间关系的相似度测度. 令 X_p 和 X_q 为聚类得到的 m 个集合中的 2 个子集，定义 2 个子集间的相似度为

$$e_{pq}=\sum_{x_i\in X_p}\sum_{x_j\in X_q}a_{ij}\Big/\sum_{x_i\in X}\sum_{x_j\in X}a_{ij};\ p, q=1, 2, \cdots, m \tag{4}$$

当 $p=q$ 时，e_{pp} 是 X_p 内对象之间的相似度测度，当 $p\neq q$ 时，e_{pq} 是 X_p 和 X_q 之间对象的相似度测度. 有效的数据聚类结果需要同时满足 $\max(\sum_p e_{pp})$ 和 $\min(\sum_{p,q} e_{pq})$；$p\neq q$，$p, q=1, 2, \cdots, m$. 因此，定义聚类的评价准则为

$$J=\sum_{p=1}^{m}\Big[e_{pp}-\Big(\sum_{q=1}^{m}e_{pq}\Big)^2\Big] \tag{5}$$

式中：$\sum_{p=1}^{m}e_{pp}$ 反映子集内部对象的连接情况，是类内相似度；$\sum_{p=1}^{m}\big(\sum_{q=1}^{m}e_{pq}\big)^2$ 体现了子集之间节点的连接情况，是类间相似度. 显然，类内相似度越大、类间相似度越小，则 J 越大，这正是聚类的目标.

2 基于流形距离的迭代优化聚类算法

2.1 算法基本思想

确定准则函数后，聚类问题转化为找到最优划分，使准则函数取极值的优化问题. 考虑到计算的复杂度和实用性，采用迭代最优化的解决途径，具体实现过程是把每个样本作为一类，根据准则函数两两合并类，直到找到最优划分. 假定在寻找最优划分的过程中，已经将对象集合 X 合并为 k 类，准则函数值为 J_k. 此时，如果要将 X_f 和 X_g 合并，J_k 将变化为 J_{k-1}，即

$$J_{k-1}=\sum_{p=1}^{k-1}e_{pp}-\sum_{p=1}^{k-1}\Big(\sum_{q=1}^{k-1}e_{pq}\Big)^2=\sum_{p=1}^{k}e_{pp}+e_{fg}+e_{gf}-\sum_{p=1}^{k}\Big(\sum_{q=1}^{k}e_{pq}\Big)^2-2\sum_{p=1}^{k}e_{fp}\sum_{p=1}^{k}e_{gp}=J_k+2e_{fg}-2\sum_{p=1}^{k}e_{fp}\sum_{p=1}^{k}e_{gp} \tag{6}$$

令 $\delta_{fg}=J_{k-1}-J_k$，则有

$$\delta_{fg}=2e_{fg}-2\sum_{p=1}^{k}e_{fp}\sum_{p=1}^{k}e_{gp} \tag{7}$$

式中：$e_{fg}=e_{gf}$ 为子集 X_f 和 X_g 间的相似度；e_{fp} 为子集 X_f 和 X_p 间的相似度；e_{gp} 为子集 X_g 和 X_p 间的相似度；δ_{fg} 为合并子集 X_f 和 X_g 引起的 J 的取值的变化量.

可见，合并 X_f 和 X_g 会使 J 发生式(7)的变化. 考虑到 J 值越大聚类效果越好，因此只要满足 $\delta_{fg}>0$ 就说明将 X_f 和 X_g 合并是有利的，反之则不能合并. 进一步，如果 $\max\limits_{f,g}(\delta_{fg})>0$，就能保证此次合并是最优且有效的.

2.2 算法描述

引入流形距离度量样本相似度，综合考虑类内类间关系设计准则函数，采用迭代最优化的方法解决聚类准则函数优化问题，形成基于流形距离的迭代优化聚类算法. 算法流程如下.

步骤 1 利用 K 最近邻算法[5]的思想，由式(1)构建线段长度矩阵 $\boldsymbol{L}$，利用 Floyd 算法[6]计算 2 个数据点间的最短路径，得到流形距离，并根据式(3)计算 2 个数据点间的相似度.

步骤 2 把每一个对象视为一类，根据式(7)计算初始合并时所有的准则变化量.

步骤 3 找出变化量中的最大值 δ_{fg}（如果存在多个最大值，选取第 1 个即可），如果 $\delta_{fg}>0$，继续. 否则，转到步骤 5.

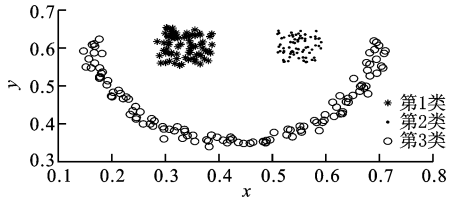
步骤 4 合并 X_f 和 X_g，重新计算所有的准则变化量，返回步骤 3.

步骤 5 算法停止并输出最终聚类结果.

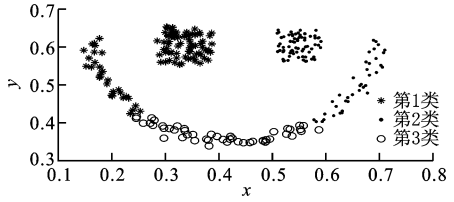
算法的计算量集中在 Floyd 算法以及计算 $\max(\delta_{fg})>0$ 上，Floyd 算法计算复杂度为 $O(n^3)$，每次只合并 2 个子集，只需重新计算少数准则变化量的值，计算复杂度最大为 $O(n^2)$. 由于样本规模 n 是有限的，每次迭代合并 2 个子集，决定了最大迭代次数是 $n-1$，即最多经过 $n-1$ 次合并后，所有样本被聚为一类，算法停止.

3 试验与讨论

为了能够直观地考察本文算法发现聚类结构的能力，将算法应用于 4 个人工数据集的聚类问题. 这 4 个数据集具有不同的流形结构，能够用于考查算法对不同结构数据的聚类性能，并且把本文算法的聚类结果与标准的 k 均值算法[7]的聚类结果进行比较，进一步验证了算法性能，如图2～图5所示.

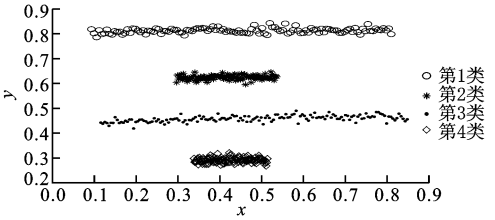


(a)本文算法的聚类结果

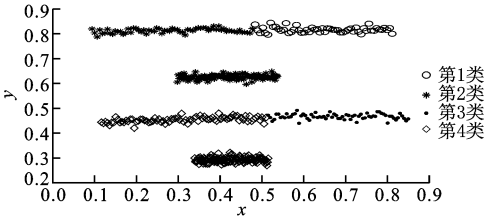


(b)k 均值算法的聚类结果

图 2 2 种算法对数据集 Line-blobs 的聚类结果

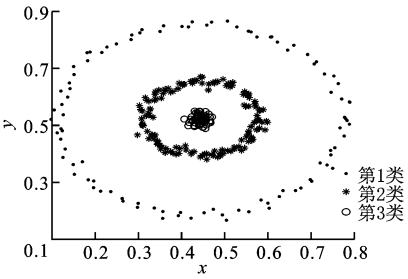


(a)本文算法的聚类结果

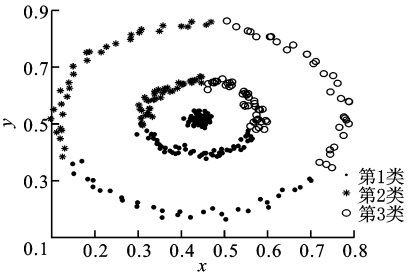


(b)k 均值算法的聚类结果

图 3 2 种算法对数据集 Sticks 的聚类结果

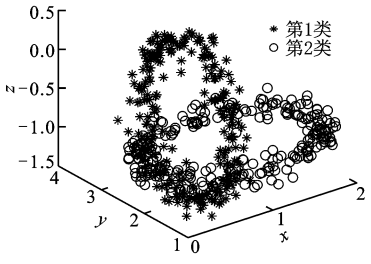


(a)本文算法的聚类结果

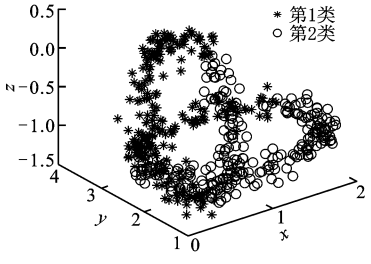


(b)k 均值算法的聚类结果

图 4　2 种算法对数据集 Three-circles 的聚类结果



(a)本文算法的聚类结果



(b)k 均值算法的聚类结果

图 5　2 种算法对数据集 Two-donut 的聚类结果

试验中本文算法邻近样本数目 $K=7$.

试验结果表明，本文算法对于不同的流形结构和非球形分布数据，能够取得较好的聚类效果，测试数据集聚类正确率均为百分之百，说明新的基于流形距离的相似度度量和准则函数对于复杂结构的数据聚类问题是非常有效的. 此外，本文算法实现简单，参数很少，与 k 均值算法相比，本文算法能够自动确定聚类数目，由于算法实现过程中没有引入随机操作，所以其结果是确定的.

算法中 K 是惟一的参数且为正整数，大量的参数敏感试验表明：当 K 很小时，会把同一流形分成不同流形，增加了类数目；当 K 很大时，又会将不同流形合并，从而减少类数目；当 K 取值在 5～7 之间时，聚类效果较好.

4　结　论

聚类研究的 2 个重要问题是样本间相似度测度和聚类准则函数的设计. 针对传统的欧氏距离测度无法正确描述数据的全局一致性的问题，本文引入流形距离作为相似度度量，不仅满足了数据的局部一致性假设，而且满足了全局一致性的假设，因此对于复杂结构聚类问题，更能准确地描述其数据结构. 设计同时满足类内相似度大、类间相似度小的聚类准则函数，更好地反映了聚类目标. 本文通过迭代优化准则函数的方法来实现聚类，过程比较简单，结果具有确定性，并能取得较好的聚类效果.

参考文献：

[1] DUDA R O, HART P E, STORK D G. Pattern classification [M]. New York, USA: Wiley, 2001: 538-548.

[2] SU Muchu, 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.

[3] ZHOU Dengyong, BOUSQUET O, LAL T N, et al. Learning with local and global consistency [C]// Advances in Neural Information Processing Systems: 16. Cambridge, USA: MIT Press, 2004: 321-328.

[4] TENENBAUM J B, DE SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290(550): 2319-2323.

[5] SHAKHNAROVICH G, DARRELL T, INDYK P. Nearest-neighbor methods in learning and vision [M]. Cambridge, USA: MIT Press, 2005.

[6] FLOYD R W. Algorithm 97: shortest path [J]. Communications of the ACM, 1962, 5(6): 345.

[7] MACQUEEN J B. Some methods for classification and analysis of multivariate observations [C] // The 5th Berkeley Symposium on Mathematical Statistics and Probability. Berkeley, USA: Univ of Calif Press, 1967: 281-297.

（编辑　管咏梅）