基于相对密度和流形上k近邻的聚类算法

古凌岚[1] 彭利民[2]
(广东轻工职业技术学院计算机工程系 广州 510300)[1]
(华南理工大学自动化科学与工程学院 广州 510006)[2]

摘 要 针对传统的基于欧氏距离的相似性度量不能完全反映复杂结构的数据分布特性的问题,提出了一种基于相对密度和流形上k近邻的聚类算法。基于能描述全局一致性信息的流形距离,及可体现局部相似性和紧密度的k近邻概念,通过流形上k近邻相似度度量数据对象间的相似性,采用k近邻的相对紧密度发现不同密度下的类簇,设计近邻点对约束规则搜寻k近邻点对构成的近邻链,归类数据对象及识别离群点。与标准k-means算法、流形距离改进的k-means算法进行了性能比较,在人工数据集和UCI数据集上的仿真实验结果均表明,该算法能有效地处理复杂结构的数据聚类问题,且聚类效果更好。

关键词 流形距离,流形上k近邻,k近邻相似度,相对密度
中图法分类号 TP391.9 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.039

Clustering Algorithm Based on Relative Density and k-nearest Neighbors over Manifolds

GU Ling-lan[1] PENG Li-min[2]
(Department of Computer Engineering, Guangdong Industry Technical College, Guangzhou 510300, China)[1]
(School of Computer Science and Engineering, South China University of Technology, Guangzhou 510006, China)[2]

Abstract For the problem that traditional Euclidean distance similarity measure cannot fully reflect the distribution characteristics of the complicated data structure, a clustering algorithm based on relative density and k-nearest neighbors over manifolds was proposed. The manifold distance which describes the global consistency and the k-nearest neighbors concept that shows local similarity and affinity were introduced. Based on above descriptions, firstly, the similarity between two objects is measured through the k-nearest neighbors similarity over manifolds. Secondly, the cluster under different densities is found by adapting the relative uniformity of the k-nearest neighbors. Lastly, the k-nearest neighbor pair constraint rule is designed to search the nearest neighbor chain which is composed of the k-nearest data points, in order to classify data objects and identify outliers. Experimental results show that compared with traditional k-means clustering algorithm and the improved k-means clustering algorithm by manifold distance, the algorithm can effectively deal with the clustering problem for complicated data structure and achieve better clustering effect on artificial data sets and UCI public data sets.

Keywords Manifold distance, k-nearest neighbors over manifolds, k-nearest neighbors similarity, Relative density

1 引言

聚类分析是通过确定数据对象在某些属性上的相似性,将其划分成群组或类簇的过程。聚类的目标是最大程度上实现类中的数据对象相似度最大、类间的数据对象相似度最小。聚类分析主要应用于模式识别、机器学习和数据挖掘等多个领域,具有广泛的应用前景。

基于划分的方法是现有聚类分析算法中较为常用的一类算法,它是以欧氏距离作为定义数据对象间相似度的基础,形成了由不同相似度定义方法而产生的多种聚类算法,较典型的算法有k-means、近邻传播聚类(Affinity Propagation clustering, AP)、谱聚类等[1],在密度和尺寸相近的球形体数据集上聚类效果较好,但k-means对于初始中心选择及噪声数据很敏感,且易陷入局部最优[2];AP在数据结构松散的情况下易产生局部聚类,聚类效果不佳[3];谱聚类能识别非球形体数据集,但其计算时空复杂度非常高[4]。

聚类分析的目的在于发现数据集中隐藏的聚类结构,实际应用中的数据集可能存在不同密度的流形结构,局部空间上的相近性不再是测度相似性的决定性因素,因而,基于欧氏距离的方法无法反映数据潜在的复杂结构。流形距离测度[5]可以测度沿着流形上的最短路径,能有效地描述数据集聚类的全局一致性[6],已有研究者将其用于聚类算法方面的研究,其对于复杂结构分布的数据聚类有良好的效果[5,7,8],但直接采用流形距离计算复杂度较高,且密度不均或带有噪声点的数据集上的聚类效果仍较差。王立敏等[9]提出基于k-近邻结构相似度的相对密度,对基于密度的聚类算法进行改进,一

到稿日期:2015-11-07 返修日期:2016-04-28 本文受国家档案局科技项目(2015-X-54),广东省自然科学基金资助项目(S2012040007599),广东省档案局科技项目(YDK-95-2014)资助。

古凌岚(1965—),女,硕士,副教授,主要研究方向为分布式计算、数据挖掘,E-mail:Li_Lace@126.com;彭利民(1976—),男,博士后,副教授,主要研究方向为分布式计算等。

定程度上解决了密度不均匀时聚类结果出错的问题，为本文提供了有用的借鉴和参考。

针对现有算法在处理复杂结构数据集时存在的不足，本文基于流形上的线段长度[6]和 k 近邻[10]提出了一种新的聚类算法。1)通过流形上的线段长度测度数据对象在局部空间上的相似度，利用 k 近邻汇聚流形上的同类点，无需反复比较计算流形上的最短路径；2)测度每个数据对象的 k 近邻点的疏密度和分布均匀度，同时搜索非流形上的 k 近邻，以发现密度不同的类簇中心；3)设计近邻点对约束规则，利用已知信息，沿流形上查找每个类簇中心对象的 k 近邻，以实现数据对象的归类及离群点的识别；4)在人工数据集和真实数据集上对本文算法进行了验证，证实了所提算法能有效地发现密度不均匀、任意形状的类簇。

2 相似度测度与相对密度

聚类问题符合以下假设[6]：1)邻近的数据点很可能具有相似性(局部一致性)；2)同一流形上的数据点很可能具有相似性(全局一致性)。进一步可假设：包含高密度区域的流形结构中邻近点很可能是同一类簇，这里的近邻点可以是流形上两个互邻的相近点，也可以是由多个互邻的相近点对所形成路径的两个端点。由此，同一类簇内的点分布趋于相对均匀且密集，不同类簇间存在着分布趋于稀疏的区域。

图 1 同一流形上近邻点和不同流形上的非近邻点

图 1 中，x,y 就是同一流区域中相隔多个互邻相近点的两点，应属同类，x,y 与 z 之间找不到一条由互邻相近点组成的路径，则应分属两个类，但若使用欧氏距离测度，显然 x 与 z 较 x 与 y 距离要小，会被错误地划为同一类，而采用标准的流形距离虽能克服欧氏距离的缺陷，但需要计算比较 x 到 y 以及 x 到 z 的所有路径，得到最短路径(流形距离)来确定类的划分，计算复杂度相当高。为此，本文基于 k 近邻和流形上的线段长度，定义一种新的流形距离——流形上 k 近邻相似度。

定义 1 对于数据集 D 的任意点 p 和正整数 k，$R_{p,k}$ 表示距离(采用流形线段长度测度) p 最近的 k 个数据点集合，且第 k 个点 r 距离最远，p 到 $R_{p,k}$ 上任意点 q(除 r 外)的距离称为 p 与 q 的近邻相似度(记作 $s(p,q)$)，p 到 r 的距离称为 p 的 k 近邻相似度(记作 $s_k(p)$)并将 k 个数据点称为 p 的 k 近邻。

$$s(p,q)=\begin{cases}\rho^{d(p,q)}-1, & q\in R_{p,k}\\ \infty, & \text{else}\end{cases} \tag{1}$$

$s_k(p)$ 与 $s(p,q)$ 的计算方法相同，式(1)中 $d(p,q)$ 是欧氏距离，$\rho^{d(p,q)}-1$ 是 p 与 q 间的流形线段长度，能够缩小短线段的长度，放大长线段的长度，ρ 为调节因子，取值为 e^2[6]。

定义 2 对于数据集 D，其上任意流形区域 f 的任意两点 p_i 和 p_j 之间，如果存在由多个 k 近邻对连成的路径，即 p_{i+1} 为 p_i 的 k 近邻，p_{i+2} 为 p_{i+1} 的 k 近邻，…，p_{j-1} 为 p_j 的 k 近邻，且 $s(p_h,p_{h+1})\leqslant s_k(p_i)$，其中 $i\leqslant h<j$，则点 p_i 和 p_j 间的流形上 k 近邻相似度为：

$$s_f(p_i,p_j)=\begin{cases}\min\limits_{p\in R_{ij}}\sum\limits_{i\leqslant h<j}s(p_h,p_{h+1}), & R_{ij}\neq\varnothing\\ \infty, & \text{else}\end{cases}$$

其中 R_{ij} 为路径集合，称 $p_{i+1},p_{i+2},\cdots,p_j$ 为 p_i 的流形上 k 近邻。

利用流形上 k 近邻相似度刻画任意两点间距离，同一流形上的两点可以通过多个线段较短的 k 近邻点相连接，而不同流之间不存在任何由 k 近邻点组成的路径，以缩短同一流形上数据点间的距离，使同一流形上的点更具相似性，放大不同流形之间数据点的距离，令不同流形上的点相似度较小，从而满足了聚类问题的局部一致性和全局一致性假设。由定义 1 和定义 2 可知如下性质。

性质 1 p 与 r_1 同属类 c，且 r_1 为 p 的 k 近邻，若 r_2 为 r_1 的 k 近邻，则 r_2 也属于类 c，称为近邻点对约束，即有以下传递关系：

$$(p,r_1)\in c\&\&(p,r_1)\in neighbour\&\&$$
$$(r_1,r_2)\in neighbour\Rightarrow(p,r_2)\in c$$

由定义 1 可知，k 一定的情况下，$s_k(p)$ 越小，p 周围数据点越密集，反之，越稀疏，这一定程度上反映了局部范围内数据的疏密度，但直接作为类簇的判断条件还存在局限性。因 $s_k(p)$ 是基于距离刻画密度的，通过比较 p 的各 k 近邻点的 $s_k(p)$，可评估数据点分布的均匀程度，对于发现不同密度的类会更为有利。如图 2 中，假设 $k=4$，六边形点、圆形点和菱形点分别为区域 1,2 和 3 上 p 的 k 近邻，当 $s_k(p)$ 较大时，有两种可能：1)区域 3，p 的 k-邻域中存在噪声点，会出现 $s_k(p)$ 与其 k-邻域内某些点(如 a,b)的 k 近邻相似度的差值较大；2)区域 2，介于高密集区域和噪声之间，相对均匀且较密集，应被认为是一个类簇，这里 p 的 $s_k(p)$ 及其各 k 近邻点的 k 近邻相似度都相近。同时观察图 2 可知，数据点密集的区域 1 的 $s_k(p)$ 与区域 2 有类似特点，因而以 $s_k(p)$ 比值为依据，能有效地发现密度均匀区域，无论是数据点密集的区域 1，还是相对均匀且较密集的区域 2。

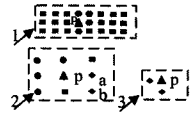

图 2 密度不均匀的数据集

根据上述分析，本文基于 k 近邻相似度比值给出如下定义。

定义 3 数据集 D 上任意点 p 的 k 近邻相似度，称为 p 的 k 近邻密度(记作 $dens(p)$)，即 $dens(p)=s_k(p)$，则 p 的 k 近邻的平均 k 近邻密度为 $dens_{avg}(p)=\frac{1}{k}\sum\limits_{1\leqslant i\leqslant k}s_k(q_i)$，$p$ 的 k 近邻密度与其 k 近邻的平均 k 近邻密度之比称为 p 的 k 近邻均匀度，记为 w：

$$w=|(dens(p)/dens_{avg}(p))-1| \tag{2}$$

定义 4 数据集 D 上任意点 p 的 k 近邻相对密度，记为 $rd_k(p)$：

$$rd_k(p)=w*dens(p) \tag{3}$$

当 p 的 k 近邻点分布均匀时，各点的 k 近邻密度必然相近，p 的 k 近邻密度与其 k 近邻点的平均 k 近邻密度之比将接近 1，即 k 近邻均匀度 w 值会较小。w 越小，表明 p 的各 k 近邻点的一致性越高，而 $dens(p)$ 是对 p 的 k 近邻点疏密度的刻画，k 近邻均匀度系数加权后得到的度量指标可体现数据点间的相对紧密程度。$rd_k(p)$ 越小的点，越有可能成为类中心，反之，可能是离群点。直接使用相对密度识别离群点时，其个数的确定通常需要依赖基于用户经验的离群度(相对密度)阈值[11-13]，这种方法较难准确地确定数据集中的离群点。由于离群点可以描述为不属于任何类簇的点或是归属于类簇内元素数很小的点，因此可对数据集先进行聚类处理，划分为

若干个类簇，并将未划入任何类簇的点确定为离群点。即根据定义2、定义4，如果数据集 D 上存在非聚类中心点 q，对于任意类簇中心点 p，q 都不是 p 的流形上 k 近邻，则 q 为离群点。

3 基于相对密度和流形上 k 近邻的聚类算法

3.1 算法描述

如上所述，基于 k 近邻和流形上线段长度，依据 k 近邻相对密度测度类簇的疏密度和均匀度，找到较合理的簇划分，采用流形上 k 近邻相似度刻画同一流形上点间的相似性，以发现类簇内的空间分布信息，可以反映聚类的局部和全局一致性，将会得到更为理想的聚类结果。本文提出一种基于相对密度和流形上 k 近邻的聚类算法，首先采用流形线段长度构建流形度量矩阵，由此形成 k 近邻相似度矩阵，计算各数据点的 k 近邻相对密度；然后，找寻 k 近邻相对密度值 $rd_k(p)$ 小（数据点相对密度高），且互不为流形上 k 近邻的点作为类簇中心，再对各类簇中心的流形上 k 近邻点进行归类；最后将未能归类的点划入离群点集中。下面给出算法的具体描述。

输入：数据集 $X=\{x_1,x_2,\cdots,x_n\}$，近邻个数 k，聚类个数 v

输出：聚类划分 $C=\{c_1,c_2,\cdots,c_v\}$

1. 对于 X 任意两点 x_i,x_j，计算两点间的欧氏距离 $d(x_i,x_j)=\|x_i-x_j\|=\sqrt{\sum_{h=1}^{s}(x_{ih}-x_{jh})^2}$，以及流形线段长度 $l(x_i,x_j)=\rho^{d(x_i,x_j)}-1$，构造流形度量矩阵 $[M(i,j)]_{n\times n}$，其中 $m_{ij}=l(x_i,x_j)$；
2. 根据流形度量矩阵 M 及 k，对各数据点到其他点的距离进行升序排序，并选取各数据点及其前 k 个距离最小的点，构成近邻相似度矩阵 $[R(i,j)]_{n\times(k+1)}$，$j=0$ 时，r_{ij} 表示数据点 x_i 本身，$j>0$ 时，r_{ij} 是数据点 x_i 的 k 近邻点 x_j 编号；
3. 根据流形度量矩阵 M，由式(3)计算 x_i 的 k 近邻相对密度 $rd_k(p)$，并进行升序排序；
4. 选 $rd_k(p)$ 前 $k*v$ 个点作为候选类簇中心集，并将其中最小的点作为首个类簇中心 c_1，放入集合 C；
5. 从候选集中搜寻流形距离最远，且不是集合 C 中任何中心点的流形上 k 近邻点，作为下一个类簇中心，并放入集合 C，重复此步骤直到找到 v 个类簇中心；
6. 对于每个类簇中心 c_i，依据近邻点对约束规则，发现其成员数据点，若成员数据点可同时归属多个类簇中心，则将其归于流形上 k 近邻相似度最小的类簇；
7. 将未能归入任何类簇的点划入离群点集；
8. 输出聚类结果。

对于算法步骤 2 和步骤 3 的排序处理，考虑到待排数据是随机分布的，且无稳定性方面要求，因而这里均采用经典快速排序算法[14]。

3.2 算法分析

所提算法的复杂度主要取决于流形度量矩阵 M 构建、近邻相似度矩阵 R 构建，以及类簇中心和成员数据点发现。下面分别进行具体分析，首先假设数据点个数为 n。

(1)流形度量矩阵的构建，需计算各数据点间的欧氏距离，再利用 ρ 参数调节得到流形距离，时间复杂度为 $O(n^2)$。

(2)近邻相似度矩阵 R 的构建，是先对 M 的各行进行快速排序，再选取其前 k 个点构成 k 近邻，时间复杂度为 $O(n\times n\log n)$。

(3)类簇中心的发现，包括候选类簇中心的筛选和类簇中心的确定。作为筛选依据的 k 近邻相对密度，需要计算和排序（快速排序），时间复杂度为 $O(n+n\log n)$；确定环节是寻找 $rd_k(p)$ 小且互为非流形上 k 近邻的点的过程，需要针对每个候选类簇中心遍历数据集所有点，搜索是否存在 k 近邻点对序列组成的路径，时间复杂度为 $O(v\times s\times n)\approx O(n^2)$，其中，$k$ 为近邻个数，v 为类簇个数，s 为候选类簇中心集规模，且 $v\ll n$，$s\ll n$，这与标准流形距离的 $O(n^3)$[6] 相比计算复杂度明显减少。

(4)成员数据点的发现是以类簇中心为初始点，利用 R 不断发现类簇内各点的 k 近邻的过程，无需通过计算流形上 k 近邻相似度来判断归类，时间复杂度为 $O(nk)$。

综上所述，所提算法的总体时间复杂度为 $O(n^2\log n)$。

参数影响方面，所提算法有两个参数 k 值和 v 值。近邻个数 k 的选择对算法的结果会产生重大影响，实际应用中一般不大于 20[11]，但过小容易出现过拟合，经过多次实验，k 值为[7,20]时，聚类效果较好，当数据规模较大时，k 值可取大些。v 为给定的聚类个数。

4 实验与分析

为了验证所提聚类算法的性能，在 Matlab 2012b 上实现了所提算法，采用 7 个人工数据集和 4 个 UCI 数据库中的常用数据集作为实验数据，通过 k-means、采用流形距离改进的 k-means(Fk_means)和本文算法进行了实验。

4.1 人工数据集上实验及算法比较

(1)采用人工数据集 smile，two spirals，three circles，three lines 和 two moons，其流形结构各不相同，能够考量在不同数据结构下算法的性能。

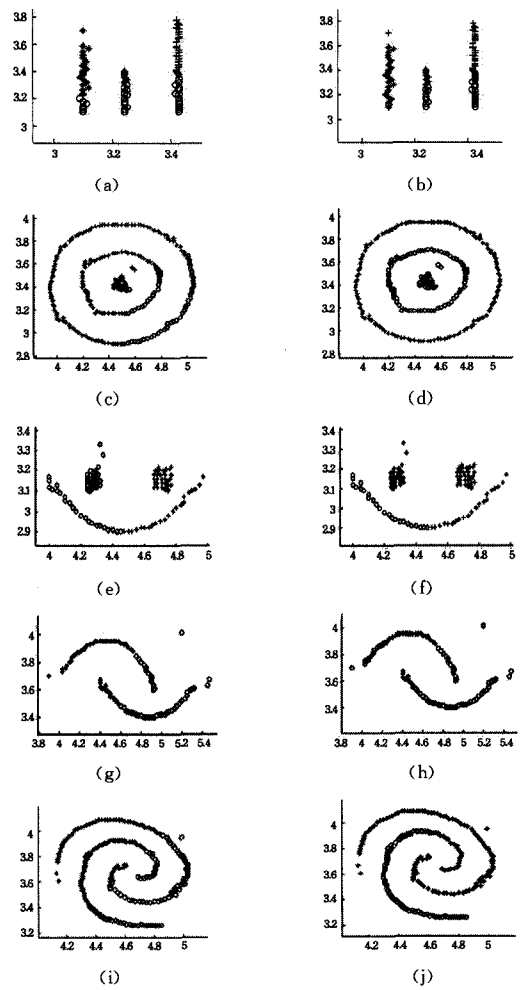

图 3 其他两种算法对 smile 等 5 个数据集的聚类结果

对于上述数据集，分别运行 k-means，Fk_means 和本文算法各 10 次，本文算法参数 $k=[9,10]$，均选最佳效果予以展示，k-means 算法的实验结果如图 3 中(a)、(c)、(e)、(g)和

(i)所示。由于采用欧氏距离无法有效地描述流形结构中数据点的相似性,导致聚类结果错误,图 3 中其他小图为 Fk_means 算法的实验结果,聚类效果有很明显的改善,尤其在 two moons 和 smile 上,但仍不够理想;另外,由于对初始值敏感,聚类效果也不稳定。而所提算法对于 5 种不同结构的数据集,10 次运行均有正确的聚类结果,并能识别出离群点(图中三角形状的点),实验结果如图 4 所示,说明采用流形上 k 近邻相似度归类数据点,能够更好地弥补欧氏距离无法反映聚类全局一致性的不足,结合 k 近邻相对密度发现类簇中心,聚类效果较为稳定。

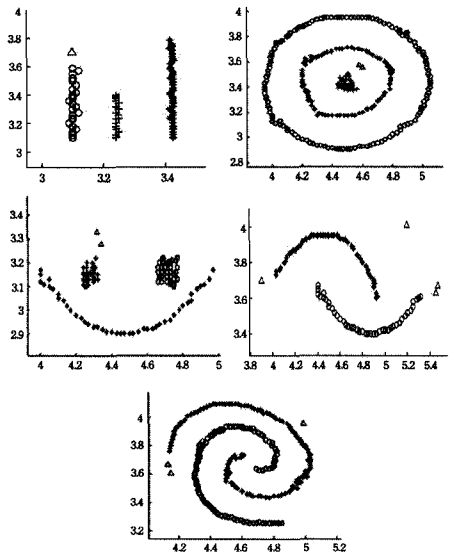

图 4　所提算法对 smile 等 5 个数据集的聚类结果

(2)采用人工数据集 two squares 和 mult squares(数据集中包含不均匀密度区域),以验证 3 种算法对于不均匀密度下的数据集的聚类效果。

对于上述数据集,分别运行 k-means,Fk_means 和所提算法各 10 次,所提算法参数 $k=[9,10]$,同样选取最佳效果予以展示,图 5(a)和图 5(b)显示出 k-means 和 Fk-means 算法对于包含两个不同密度区域下的数据集 two squares 聚类效果都比较好;从图 5(c)和图 5(d)可知,对于有多个不同密度区域的数据集 mult squares,前两种算法均出现一个区域边缘点被划入其他区域的情况。图 6 所示的实验结果表明,所提算法采用 k 近邻均匀度系数加权的密度指标,对于两个数据集都能够发现不均匀密度下的类簇及离群点。

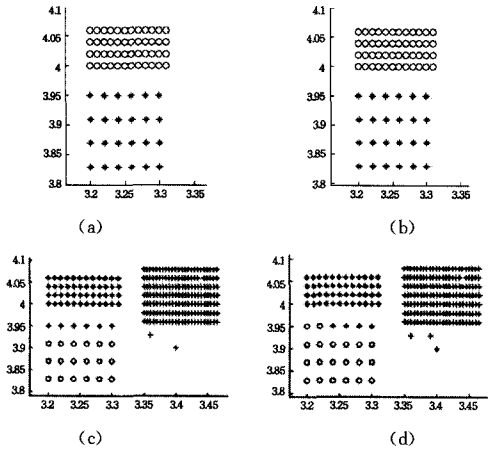

图 5　其他算法对 two squares 等两个数据集的聚类结果

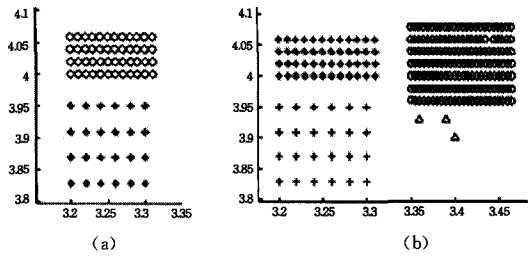

图 6　所提算法对 two squares 等两个数据集的聚类结果

(3)根据 k-means,Fk_means 和本文算法在上述 7 个人工数据集上的处理时间,分析算法的执行效率。表 1 为各算法运行 10 次的平均时间(单位为秒),以及前两种算法的平均迭代次数(单位为次)。

表 1　3 种算法在人工数据集上的运行情况

数据集	k-means (运行时间/迭代次数)	Fk_means (运行时间/迭代次数)	所提算法
smile	2.07/2	2.216/2	1.137
two spirals	2.762/2	4.047/3	1.490
three circles	5.200/3	8.024/4	3.301
three lines	2.562/3	3.336/9	0.473
two moons	1.754/3	0.686/2	0.874
two squares	1.645/1	0.400/1	0.634
mult squares	3.026/2	4.437/1	3.018

由表 1 可知,k-means 和 Fk_means 的迭代次数相近,但在运行时间上,除规模很小的两个数据集(two moons 和 two squares 数据点个数均小于 100)外,后者明显比前者的运行时间长,这表明流形距离计算具有较高的时间复杂度,影响了算法的执行效率;而所提算法采用基于 k 近邻和流形上线段长度的流形距离,降低了计算复杂度,且无需多次迭代,在运行效率方面具有明显优势。

4.2　UCI 数据集上实验及算法比较

为了验证算法在真实数据集上的聚类效果和运行效率,选用了 UCI 数据库的 4 个数据集,数据集详细信息如表 2 所列。

表 2　UCI 数据集信息

数据集	样本数	属性数	分类数
iris	150	4	3
glass	214	9	6
bupa	345	6	2
vehicle	846	18	4

实验评价聚类结果采用常用的 *F-score* 指标[15]。*F-score* 指标通过算法的准确率和查全率计算得到。类 t 和聚类 C_k 的准确率和查全率分别为 $Prec(t,C_k)=N_{tk}/N_k$ 和 $Rec(t,C_k)=N_{tk}/N_t$。其中,N_k 表示聚类算法识别出的类簇 k 的样本数,N_{tk} 表示类簇 k 中 t 的样本数,N_t 表示类簇 t 原有的样本数。

相应的 *F-score* 计算公式为:

$$F\text{-}score=\frac{2*Prec(t,C_k)*Rec(t,C_k)}{Prec(t,C_k)+Rec(t,C_k)} \tag{4}$$

整个聚类划分的 *F-score* 值为:

$$F(C)=\sum_{t\in T}\frac{N_t}{N}\max_{C_k\in C}(F\text{-}measure(t,C_k)) \tag{5}$$

其中,N 表示数据集内数据点的总数,C 表示算法运行得到的类的集合,T 表示数据集的真实类簇的集合。*F-score* 指标的取值范围是[0,1],其取值越大表示算法越准确。

所提算法分别计算 k 值为 7－20 的结果，再从结果集中选取最佳者计算 *F-score* 指标。3 种算法在 4 个真实数据集上聚类结果的 *F-score* 指标和运行时间如表 3 所列(表中的时间表示平均运行时间，单位为秒；次数表示平均迭代次数，单位为次)，所提算法在 4 个数据集上的 *F-score* 指标均优于其他两种算法，且在 iris，glass 和 bupa 上表现较为突出，这表明所提算法所采用的聚类方法更为有效。另外，4 个数据集上的运行情况也表明，所提算法在运行效率方面表现最佳。

表 3　3 种算法在 UCI 数据集上的 *F-score* 指标和运行情况

数据集	k-means		Fk_means		所提算法	
	F-score	时间/次数	F-score	时间/次数	F-score	时间
iris	0.9077	2.387/5	0.9088	8.007/8	0.9640	0.767
glass	0.4887	4.480/9	0.4914	22.701/6	0.5428	2.229
bupa	0.4675	3.924/10	0.4675	27.183/8	0.5537	3.587
vehicle	0.4291	26.78/20	0.4291	227.613/19	0.4332	24.66

结束语　提出了一种基于相对密度和流形上 k 近邻的聚类算法。该算法通过流形上 k 近邻相似度描述数据点间的相似性，克服了基于欧氏距离的测度方法无法反映全局一致性的局限性，采用 k 近邻相似度度量局部疏密度，同时还考虑了数据点分布均匀度所带来的影响，从而有效地发现不均匀密度下的类簇。实验表明，与 k-means、流形距离改进的 k-means 算法相比，该算法不仅能够识别不同数据结构、密度不均匀的类簇和离群点，而且在 UCI 真实数据集上也能够有较好的聚类精准度。如何将该算法应用于图像分割和分类，将是下一步的研究重点。

参 考 文 献

[1] Jin Jian-guo. Review of clustering method [J]. Computer Science, 2014, 41(11A): 288-293(in Chinese)
金建国. 聚类方法综述[J]. 计算机科学, 2014, 41(11A): 288-293

[2] Huang X, Su W. An improved K-means clustering algorithm [J]. Journal of Networks, 2014, 9(1): 161-167

[3] Chen Y, Bruzzone L, Sun F, et al. A fuzzy-statistics-based affinity propagation technique for clustering in multispectral images [J]. IEEE Transactions on Geoscience and Remote Sensing, 2010, 48(6): 2647-2659

[4] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[J]. Advances in Neural Information Processing Systems, 2002(2): 849-856

[5] Chapelle O, Zien A. Semi-supervised classification by low density separation[C]//Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics. Savannah, Barbados: Pascal Press, 2005: 57-64

[6] Gong Mao-guo, Jiao Li-cheng, Ma Wen-ping, et al. Unsupervised classification and recognition using an artificial immune system based on manifold distance[J]. Acta Automatica Sinica, 2008, 34(3): 367-375(in Chinese)
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

[7] Li Yang-yang, Shi Hong-zhu, Jiao Li-cheng, et al. Quantum-inspired evolutionary clustering algorithm based on manifold distance[J]. Acta Automatic Sinica, 2011, 39(10): 2343-2347(in Chinese)
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

[8] Yang Rui-rui, Niu Jian-qiang, Meng Hong-fei. Pavement crack extraction using iterative clustering algorithm based on manifold distance[J]. Computer Engineering, 2011, 37(12): 212-214(in Chinese)
杨瑞瑞，牛建强，孟红飞. 基于流形距离的迭代聚类算法路面裂缝提取[J]. 计算机工程, 2011, 37(12): 212-214

[9] Wang Li-min, Gao Xue-dong, Gong Yu, et al. Community structure detection algorithm based on relative density[J]. Computer Engineering, 2009, 35(1): 117-119(in Chinese)
王立敏，高学东，宫雨，等. 基于相对密度的社团结构探测算法[J]. 计算机工程, 2009, 35(1): 117-119

[10] Li Wen-jie, Li Wen-ming. Design and simulation of positioning method based on k-nearest neighbors algorithm[J]. Computer Simulation, 2009, 26(4): 194-196(in Chinese)
李文杰，李文明. 基于 k-近邻算法的定位方法设计和仿真[J]. 计算机仿真, 2009, 26(4): 194-196

[11] Breunig M M, Kriegel H P, Ng R T, et al. LOF: identifying density-based local outliers[J]. Acm Sigmod Record, 2000, 29(2): 93-104

[12] Jin W, Tung A K H, Han J, et al. Ranking outliers using symmetric neighborhood relationship[M]//Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2006: 577-593

[13] Tang J, Chen Z, Fu W C, et al. Enhancing effectiveness of outlier detections for low density patterns[C]//Proceedings of the 6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining. Springer-Verlag, 2002: 535-548

[14] 严蔚敏，吴伟民. 数据结构(C 语言版)(第三版)[M]. 北京：清华大学出版社, 2009: 274-277, 289

[15] Witten I H, Frank E, Hall M A. Data mining: practical machine learning tools and techniques(3rd Ed)[M]. San Fransisco: Morgan Kaufmann Publishers, 2011: 175

[16] Li Yan-bo, Song Qiong, Guo Xin-chen. Artifical immune clustering semi-supervised algorithm based on manifold distance[J]. Computer Science, 2012, 39(11): 204-207(in Chinese)
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

[17] Garcia S, Derrac J, Cano J R, et al. Prototype selection for nearest neighbor classification: Taxonomy and empirical study[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012, 34(3): 417-435