基于秩约束密度敏感距离的自适应聚类算法

任永功 刘 洋 赵 月
(辽宁师范大学计算机与信息技术学院 大连 116029)

摘 要 传统的聚类算法一般使用欧氏距离获得数据的相似矩阵,在处理一些较复杂的数据时,欧氏距离由于不能反映全局一致性,因此无法有效地描述出数据点的实际分布。提出了一种基于秩约束密度敏感距离(Rank Constraints Density Sensitive Distance,RCDSD)的自适应聚类算法。该方法首先引入密度敏感距离的相似性度量得到相似矩阵,有效地扩大了不同类数据点之间的距离,缩小了同类数据点间的距离,从而解决了传统聚类算法使用欧氏距离作为相似性度量导致聚类结果出现偏差的弊端;其次,在相似矩阵的拉普拉斯矩阵上施加秩约束,使相似矩阵的连通区域数等于聚类数,直接将数据点划分到正确的类中,得到最终的聚类结果,而不需要执行k-means或其它离散化程序。在人工仿真数据集和真实数据集上进行了大量实验,结果表明,所提算法得到了准确的聚类结果,并提高了聚类性能。

关键词 密度敏感,相似矩阵,秩约束,聚类
中图法分类号 TP39 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.05.050

Adaptive Clustering Algorithm Based on Rank Constraint Density Sensitive Distance

REN Yong-gong LIU Yang ZHAO Yue
(School of Computer and Information Technology, Liaoning Normal University, Dalian 116029, China)

Abstract The traditional clustering algorithms generally use Euclidean distance to acquire the similar matrix. In some more complex data processing, Euclidean distance doesn't have the ability of describing the characters of data because it can't reflect the global consistency. An adaptive clustering algorithm based on rank constraint density sensitive distance (RCDSD) was proposed in this paper. First, a density sensitive distance similarity measure is introduced to acquire the similar matrix which enlarges the distance between the different classes and reduces the distance between the same classes effectively, so as to solve the disadvantages of clustering results deviation of the traditional clustering algorithm based on Euclidean distance. Second, the rank constraint is imposed to the Laplacian matrix of the similarity matrix, thus the number of connected area of the similar matrix is equal to the number of clustering, and the data can be directly divided into the right class and the algorithm can take the final clustering result, while the algorithm does not need to perform k-means or other discrete procedure. Experimental results show that the approach can obtain accurate clustering results and improve the clustering performance on both artificial simulation data sets and real data sets.

Keywords Density sensitive, Similarity matrix, Rank constraints, Clustering

1 引言

随着信息技术的飞速发展,海量的数据信息涌现,人们希望能够以更高的效率挖掘出海量数据中隐含的有价值的信息,聚类分析作为探讨性的数据分析技术,已经成为数据挖掘领域的研究热点[1-3]。所谓聚类就是按照特定的要求,将数据划分到不同的类,使同一类内的对象具有很大的相似性,不同类之间的对象具有很大的相异性。在过去的几十年中,许多聚类算法被提出,如k-means聚类、分层聚类、谱聚类、支持向量机聚类等。其中,k-means算法因其简单实用的特点成为应用和研究最广泛的聚类算法之一。k-means算法[4-6],即指定数据到某一个类,使得该数据与这个类的聚类中心的距离比它到其他类的聚类中心的距离要近,该算法以欧氏距离作相似性度量,当聚类是密集的,并且类与类之间的区别明显时,k-means算法的聚类效果较好。谱聚类算法[7-10]通常比k-means算法呈现出更好的聚类性能,该算法以谱图划分理论为基础,把输入的所有数据样本看成是一个无向图的顶点,把样本之间的相似度看成是顶点之间的边,接下来优化一个目标函数,实现对图中顶点的分割,分割在同一子图中的数据样本为同一类,从而实现聚类划分,这是聚类算法中一个较新的研究方向,极大地丰富了聚类算法的研究内容。尽管谱聚类算法具有明显的优势,不同形式的相似矩阵对算法的影响却

到稿日期:2015-12-07 返修日期:2016-01-24 本文受国家自然科学基金项目(F020806),辽宁省高等学校优秀人才支持计划项目(LR2015033),辽宁省科技计划项目(2013405003),大连市科技计划项目(2013A16GX116)资助。
任永功(1972-),男,博士,教授,主要研究方向为数据库技术、数据挖掘、智能信息计算等,E-mail:renyonggong@gmail.com;刘 洋(1991-),女,硕士生,主要研究方向为数据挖掘;赵 月(1990-),女,硕士生,主要研究方向为数据挖掘。

很大，在处理很多复杂的数据集时，谱聚类算法由于使用欧氏距离作相似性度量无法有效地反映出数据的实际聚类分布。针对上述问题，本文提出了一种基于秩约束密度敏感距离的自适应聚类算法，实验证明所提算法可行有效，并且具有更优的聚类性能。

2　RCDSD 算法

2.1　基于密度敏感的距离

如图 1 所示，其据点可分为两类，点 B 属于其中一个类，点 A,C,D,E 属于另外一个类。同一类中数据的相似性一定比不是同一类中数据的相似性高，即点 A 与点 C,D,E 的相似性要比点 A 与点 B 的相似性高。但是在图 1 这种情况下，使用欧氏距离的测量结果为 A 与 D 之间的相似性小于 A 与 B 之间的相似性，这就与理想结果不一致。基于以上问题，本文采用基于密度敏感距离[11-13]的测量方法代替传统的基于欧氏距离的测量方法。

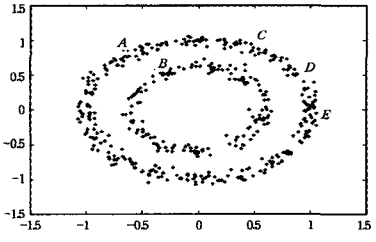


图 1　不同类中的数据点

定义 1　密度可调节的线段长度：

$$L(x_i,x_j)=\rho^{dist(x_i,x_j)}-1 \tag{1}$$

其中，$dist(x_i,x_j)$为数据点 x_i 和 x_j 的欧氏距离，ρ 为伸缩因子，且 $\rho>1$，对于复杂且非凸的数据集，ρ 可以取较小的值。$L(x_i,x_j)$不仅可以用来调节两点间的长度，而且可以用来描述聚类的全局一致性。

定义 2(密度敏感距离)　把所有的数据点对应到一个无向加权图 $G=(V,E)$，其中 $V=\{x_1,x_2,\cdots,x_n\}$，$E=\{W_{ij}\}$ 为两个数据点之间的相似度，$P=\{p_1,p_2,\cdots,p_l\}\in V$ 表示图中顶点数为 l 的点 p_1 和 p_l 之间的路径，$l=|P|$，$(p_k,p_{k+1})\in E$，$1\leqslant k\leqslant l$，p_{ij} 表示连接数据点 x_i,x_j 的所有路径集合，$i\geqslant 1$，$j\leqslant n$。得到数据点 x_i,x_j 的密度敏感距离如式(2)所示：

$$D_{ij}=\min_{p\in p_{ij}}\sum_{}^{|p|-1}L(p_k,p_{k+1}) \tag{2}$$

其中，$L(p_k,p_{k+1})$为 p_k,p_{k+1} 两点间的密度可调节长度。

由于欧氏距离矩阵满足非负性、自反性和对称性，密度敏感的距离矩阵继承欧氏距离矩阵的属性。对 $\forall x_i,x_j,x_k\in V$，$1\leqslant i,j,k\leqslant n$，密度敏感的距离具有以下 4 个性质：

(1) $D_{ij}\geqslant 0$；

(2) $D_{ij}=0$ 当且仅当 $x_i=x_j$；

(3) $D_{ij}=D_{ji}$；

(4) $D_{ij}\leqslant D_{ik}+D_{kj}$。

其中，$D_{ik}+D_{kj}$为点 x_i 经过点 x_k 到点 x_j 的最短距离，D_{ij}为点 x_i 到 x_j 之间的最短距离，根据三角形两边之和大于第三边，可以得到 $D_{ij}\leqslant D_{ik}+D_{kj}$。

因此，密度敏感距离可以有效地计算不同流行上的最短路径，使得在同一密度区域中的两个点可以用较短的边连接，而位于不同密度区域中的两个点要用较长的边连接，从而实现放大位于不同密度区域上数据点之间的距离，而缩短位于同一密度区域上数据点之间的距离。

2.2　自适应聚类算法

2.2.1　构造相似矩阵

对于数据集$\{x_1,x_2,\cdots,x_n\}$，$x\in R^{n\times d}$为数据矩阵，考虑距离因素对聚类结果的影响，对于每一个数据点 i，所有的数据点连接到 x_i 的距离对应一个概率 s_{ij}，在这里采用密度敏感的距离测量方法测量数据点之间的距离，可以很容易地判断出一个较小的距离 D_{ij}^2 被分配到同一密度区域的概率 s_{ij} 较大；同理，一个较大的距离 D_{ij}^2 被分配到同一密度区域的概率 s_{ij} 较小。该过程可描述为如式(3)所示的问题并得到概率矩阵 $s_{ij}\in R^{n\times n}$：

$$\begin{aligned}&\min_{S}\sum_{i,j=1}^{n}(D_{ij}^2s_{ij}+\lambda s_{ij}^2)\\&\text{s. t. }\ \forall i,s_i^{\mathrm{T}}1=1,0\leqslant s_i\leqslant 1\end{aligned} \tag{3}$$

通过式(3)获得一个密度敏感距离分配的概率矩阵 S 为块对角结构，其中 λ 为正则化参数，一个足够大的参数 λ 可以使位于同一区域的概率 s_{ij} 相同，即 $s_{ij}=1/n_c$，本文后续将给出参数 λ 的确定方法。

由式(3)获得的密度敏感距离概率矩阵 $S\in R^{n\times n}$可看成由 n 个数据点作为节点的图的相似矩阵。

2.2.2　引入秩约束

聚类的目标是把数据划分为 c 类，一个理想的密度敏感距离分配是将数据连接到准确的 c 个连通分支，通常情况下，在式(3)中，对于任意一个 λ 值，密度敏感距离分配无法得到一个理想的情况，所有数据点将被连接到一个连通区域。本节将秩约束引入到式(3)中，实现 RCDSD 算法将数据点划分到正确的 c 类，而不需要执行 k-means 或其他的离散化程序。

假设每一个节点 i 分配一个函数值 $y_i\in R^{1\times c}$，可推出式(4)：

$$\sum_{i,j}^{n}L^2(y_i,y_j)s_{ij}=2Tr(Y^{\mathrm{T}}L_SY) \tag{4}$$

在图理论中，$L_s=D_s-\frac{1}{2}(S^{\mathrm{T}}+S)$为拉普拉斯矩阵，以相似矩阵 S 为基础生成，度矩阵 D_S 是全部度值为对角元素的一个对角矩阵，第 i 个对角元素为$\sum_j(s_{ij}+s_{ji})/2$，数据点的度可以有效地反映其周边其它数据的分布情况。如果相似矩阵 S 非负，那么拉普拉斯矩阵有一个重要的性质[14,15]：在矩阵 S 的连接图中，拉普拉斯矩阵的 0 特征值的出现次数 c 等于图的连通区域个数。根据该性质推出：如果拉普拉斯矩阵的秩 $r(L_S)=n-c$，则根据块对角结构，图能够准确地划分数据点到精确的 c 类。结合式(3)得到聚类模型如式(5)所示：

$$\begin{aligned}&J=\min_{S}\sum_{i,j=1}^{n}(D_{ij}^2s_{ij}+\lambda s_{ij}^2)\\&\text{s. t. }\ \forall i,s_i^{\mathrm{T}}1=1,0\leqslant s_i\leqslant 1,r(L_S)=n-c\end{aligned} \tag{5}$$

式(5)即是本文要解决的核心问题，通过引入秩约束 $r(L_S)=n-c$，实现同时获得数据样本的相似矩阵和聚类结构，接下来，提出一个有效的方法解决式(5)。

设 δ_i 为拉普拉斯矩阵 L_S 的第 i 个最小特征值，L_S 是半正定的矩阵，因此 $\delta_i\geqslant 0$。对于一个足够大的 η，式(5)等同于式(6)：

$$\sum_{i,j=1}^{n}(D_{ij}^2 s_{ij}+\lambda s_{ij}^2)+2\eta\sum_{i=1}^{c}\delta_i \tag{6}$$

s. t. $\forall i, s_i^{\mathrm{T}}1=1, 0\leqslant s_i\leqslant 1, r(L_S)=n-c$

一个足够大的 η 将保证拉普拉斯矩阵的秩为 $n-c$。根据 ky Fan 的理论[16]，得到式(7)：

$$\sum_{i=1}^{c}\delta_i=\min_{Y}Tr(Y^{\mathrm{T}}L_S Y) \tag{7}$$

将式(7)代入式(6)，得到式(8)：

$$\min_{S,Y}\sum_{i,j=1}^{n}(D_{ij}{}^2 s_{ij}+\lambda s_{ij}^2)+2\eta Tr(Y^{\mathrm{T}}L_S Y) \tag{8}$$

s. t. $\forall i, s_i^{\mathrm{T}}1=1, 0\leqslant s_i\leqslant 1$

接下来，通过选择优化的方法解决式(8)。

首先，固定相似矩阵 S，处理矩阵 Y，则式(8)变成：

$$\min_{Y}Tr(Y^{\mathrm{T}}L_S Y) \tag{9}$$

式(9)中的最优矩阵 Y 由 c 个特征向量组成，对应拉普拉斯矩阵 L_S 的 c 个最小特征值。

然后，固定矩阵 Y，处理相似矩阵 S，则式(8)变成：

$$\min_{S}\sum_{i,j=1}^{n}(D_{ij}{}^2 s_{ij}+\lambda s_{ij}^2)+2\eta Tr(Y^{\mathrm{T}}L_S Y) \tag{10}$$

s. t. $\forall i, s_i^{\mathrm{T}}1=1, 0\leqslant s_i\leqslant 1$

将式(4)代入式(10)，则式(10)等同于如下问题：

$$\min_{S}\sum_{i,j=1}^{n}(D_{ij}{}^2 s_{ij}+\lambda s_{ij}^2+\eta L^2(y_i,y_j)s_{ij}) \tag{11}$$

s. t. $\forall i, s_i^{\mathrm{T}}1=1, 0\leqslant s_i\leqslant 1$

设 $f_{ij}^x=L^2(x_i,x_j)$，$f_{ij}^y=L^2(y_i,y_j)$，并且 $f_i\in R^{n\times 1}$ 为一个向量，向量的第 j 个元素 $f_{ij}=f_{ij}^x+\eta f_{ij}^y$，则式(11)写成向量形式为：

$$\min_{s_i}\| \frac{f_i}{2\lambda}+s_i \|_2^2 \tag{12}$$

在算法中，可以迭代更新选择优化步骤中的矩阵 Y 和 S，得到最终理想的聚类结果。

3 参数 λ 的取值分析

在聚类模型(5)中，有两个参数，分别是 λ 和 c。聚类数 c 可由数据的先验知识直接固定，或使用一些已有的有效方法讨论确定。因此，在新算法中，唯一需要优化的参数为 λ。本节提出了一个有效的方法来确定正则参数 λ 的值。

对于每一个数据点 i，式(5)中的 λ 与式(12)中的 λ 值相等，式(12)的拉格朗日函数为：

$$\delta(s_i,\alpha,\beta)=\frac{1}{2}\| \frac{f_i^x}{2_i\lambda_i}+s_i \|^2-\alpha(s_i^{\mathrm{T}}1-1)-\beta_i^{\mathrm{T}}s_i \tag{13}$$

其中，α,β 为拉格朗日因子，根据 KKT 条件[17]，得到最优值：

$$s_{ij}=-f_{ij}^x/2\lambda_i+\alpha \tag{14}$$

在实验中，要实现较好的性能通常需要专注于数据的位置信息，因此首选获得一个稀疏的 s_i，即只有 x_i 的 k 个最近距离的点能够连接到 x_i。计算稀疏相似矩阵 S 的另一个优点是可以很大程度地减轻后续的计算负担。

设 $f_{i1}^x, f_{i2}^x, \cdots, f_{in}^x$ 是从小到大有序排列的。如果最优 s_i 只有 k 个非零元素，即 $s_{ik}>0, s_{i,k+1}=0$，根据式(14)，可得：

$$\begin{cases}-f_{ik'}^x/2\lambda_i+\alpha>0, & k'=1,\cdots,k\\ -f_{ik'}^x/2\lambda_i+\alpha\leqslant 0, & k'=k+1,\cdots,n\end{cases} \tag{15}$$

根据不等式(15)和约束条件 $s_i^{\mathrm{T}}1=1$，得出：

$$\alpha=\frac{1}{k}(1+\frac{1}{2\lambda_i}\sum_{j=1}^{k}f_{ij}^x) \tag{16}$$

根据式(15)和式(16)，得到不等式(17)如下：

$$\begin{cases}\lambda_i>(kf_{ik}^x-\sum_{j=1}^{k}f_{ij}^x)/2\\ \lambda_i\leqslant(kf_{ik+1}^x-\sum_{j=1}^{k}f_{ij}^x)/2\end{cases} \tag{17}$$

为了得到最优 s_i，设置 $\lambda=\frac{1}{2n}\sum_{i=1}^{n}(kf_{i,k+1}^x-\sum_{j=1}^{k}f_{ij}^x)$，其中 k 是一个正数，并且很容易得到。

4 算法描述

算法 RCDSD 算法

输入：数据集 $X\in R^{n\times d}$，聚类数 k，参数 λ，一个足够大的 η

输出：具有正确连通分支数 c 的概率矩阵 $S\in R^{n\times n}$

过程：

Step1 对于 $\forall x_i, x_j\in X$，计算两点之间的欧氏距离 $\mathrm{dist}(x_i,x_j)$；

Step2 对于 $\forall x_i, x_j\in X$，计算两点之间的密度敏感距离 $D_{ij}=\min_{p\in P_{i,j}}\sum_{k=1}^{|p|-1}(p^{\mathrm{dist}(x_i,x_j)}-1)$；

Step3 初始化矩阵 S 的第 i 列 s_i 为 $s_i^{\mathrm{T}}\min_{1=1,0\leqslant s_i\leqslant 1}\sum_{i,j=1}^{n}(D_{ij}{}^2 s_{ij}+\lambda s_{ij}^2)$；

Step4 更新由 L_S 的 c 个特征向量对应的 c 个最小特征值组成的矩阵 Y，直到算法收敛；

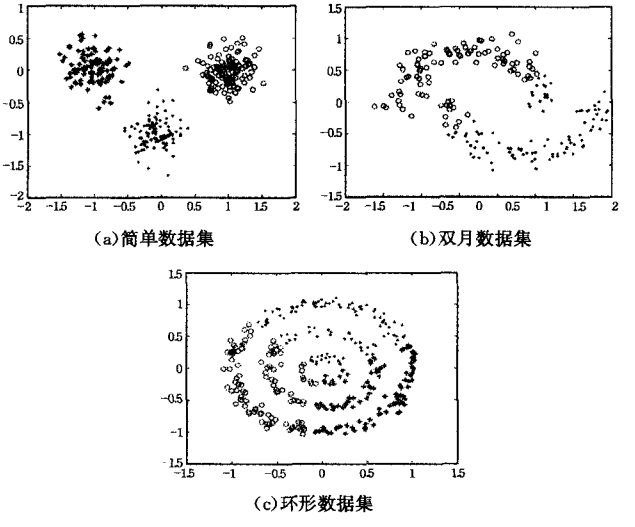
Step5 对于每个数据点 i，更新概率矩阵 S 的第 i 列为 $\min_{s_i}\| \frac{f_i}{2\lambda}+s_i \|_2^2$，直到算法收敛。

5 实验与结果分析

实验环境：Matlab2009，电脑配置：HP Pro 3381 MT PC，内存：4G，Win7 64 位操作系统。

5.1 人工仿真数据集

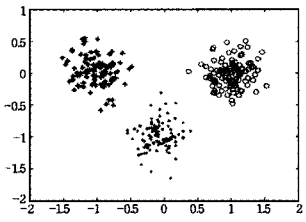
为了验证算法的有效性，首先在人工仿真数据集上进行实验，实验中选取的 3 个人工仿真数据集分别是随机生成的简单数据集(样本数 300，聚类数 3，维数 12)，随机生成的双月型数据集(样本数 200，聚类数 2，维数 2)和随机生成的环形数据集(样本数 300，聚类数 3，维数 2)。传统的谱聚类算法在处理简单的数据集时，往往能得到较好结果。



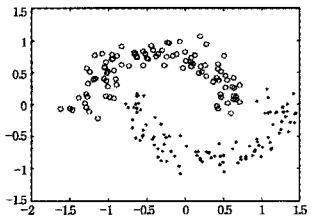
(a)简单数据集 (b)双月数据集

(c)环形数据集

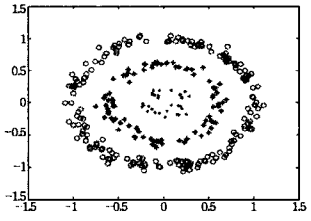
图 2 谱聚类算法的聚类结果



(a)简单数据集



(b)双月数据集



(c)环形数据集

图 3　RCDSD 算法聚类结果

如图 2(a)和图 3(a)所示，本文算法和传统谱聚类算法的聚类结果差异不大，均得到了较好的聚类结构。对于一些复杂数据集，谱聚类算法得到的聚类结果较差，如图 2(b)和图 2(c)所示，而基于密度敏感距离分配构造相似矩阵的方法，成功将数据点划分到正确类中，聚类结果如图 3(b)和图 3(c)所示。实验证明，基于秩约束密度敏感距离的自适应聚类算法是可行的，并且聚类效果较好。

5.2　真实数据集

实验选取了 8 个真实数据集，其中 seed，palm，compound，Isolet，New-thyroid，Iris，Ionosphere 均来自于 UCI 机器学习知识库，coil20 是图像数据集。表 1 总结了这 8 个数据集的描述。

表 1　真实数据集的描述

数据集	维数	聚类数	样本点数目
seed	7	3	210
palm	256	100	2000
compound	2	6	399
Isolet	617	26	1560
Coil20	1024	20	1440
New-thyroid	5	3	215
Iris	4	3	150
Ionosphere	34	2	351

在聚类实验中，设置每个数据集的类数为它的真实聚类数。为了比较不同算法的性能，使用两个广泛使用的评价指标：准确率(Accuracy，ACC)和标准化互信息量(Normalized Mutual Information，NMI)。

表 2　真实数据集上的准确率

数据集	SC	RCDSD
seed	0.7990	0.9052
palm	0.6430	0.8621
compound	0.6802	0.8895
Isolet	0.6147	0.8915
Coil20	0.6069	0.8458
New-thyroid	0.6234	0.9001
Iris	0.7535	0.9415
Ionosphere	0.5984	0.9123

表 3　真实数据集上的 NMI

数据集	SC	RCDSD
seed	0.6227	0.7335
palm	0.8627	0.9501
compound	0.6853	0.8052
Isolet	0.5429	0.6515
Coil20	0.7134	0.9200
New-thyroid	0.6849	0.8754
Iris	0.7953	0.8869
Ionosphere	0.6257	0.8723

表 2 和表 3 为每个数据集运行 50 次得到的最优 ACC 和 NMI 指标，可以看出，相比传统谱聚类算法，所提算法在真实数据集上的准确率和 NMI 指标均有很大提高，其中 New-thyroid，Iris，Ionosphere 为文献[7]中用于测量基于局部密度构造相似矩阵的谱聚类算法的准确率，经过对比，所提算法的聚类准确率更高一些，说明所提算法对数据集有更好的适应性，且其聚类性能也较高。

结束语　相似矩阵的构造对于能否得到一个更优的聚类结构至关重要，本文提出了一个新颖的聚类模型，同时得到数据相似矩阵和聚类结构，在该新的聚类方法中，数据相似度矩阵由每个数据点的密度敏感距离获得，在相似矩阵的拉普拉斯矩阵上施加秩约束，使相似矩阵的连通区域数等于聚类数，将数据点划分到正确的类。通过理论分析和对比实验表明，所提算法是有效可行的，相比传统的聚类算法，得到了准确的聚类结构，很大程度地提高了聚类性能。

参 考 文 献

[1] BERKHIN P. Survey of clustering data mining techniques[J]. Grouping Multidimensional Data, 2002, 43(1): 25-71.

[2] HU W Y, SUN Z H, WU Y J. Study of Sampling method on Data Mining and Stream Mining [J]. Journal of Computer Research and Development, 2011, 48(1): 45-54. (in Chinese)
胡文瑜，孙志挥，吴英杰. 数据挖掘取样方法研究[J]. 计算机研究与发展，2011，48(1)：45-54.

[3] LIU D Y, CHEN H L, QI H, et al. Advance in Spatiotemporal Data Mining[J]. Journal of Computer Research and Development, 2013, 50(2): 225-239. (in Chinese)
刘大有，陈慧灵，齐红，等. 时空数据挖掘研究进展[J]. 计算机研究与发展，2013，50(2)：225-239.

[4] HUANG Z H, XIANG Y, ZHANG B, et al. An Efficient Method for K-means Clustering[J]. Pattern Recognition and Artificial Intelligence, 2010, 23(4): 516-521. (in Chinese)
黄震华，向阳，张波，等. 一种进行 K-Means 聚类的有效方法[J]. 模式识别与人工智能，2010，23(4)：516-521.

[5] SARMA T H, VISWANATH P, REDDY B E. A hybrid approach to speed-up the k-means clustering method[J]. International Journal of Machine Learning & Cybernetics, 2013, 4(2): 107-117.

[6] YU H T, JIA M J, WANG H Q, et al. K-means lustering Algorithm Based on Artificial Fish Swarm[J]. Computer Science, 2012, 39(12): 60-64. (in Chinese)
于海涛，贾美娟，王慧强，等. 基于人工鱼群的优化 K-means 聚类算法[J]. 计算机科学，2012，39(12)：60-64.

(下转第 284 页)

参 考 文 献

[1] KJOS-HANSSEN B, Evangelista A J. Google distance between words[OL]. http://math. hawaii. edu/~ bjoern/Publications/Evangelista_Kjos-Hanssen. pdf.

[2] 姚双云. 复句关系标记的搭配研究[M]. 武汉:华中师范大学出版社,2008.

[3] YOU B. Measuring Semantic Relatedness between Words[D]. Wuhan:Central China Normal University Press,2013. (in Chinese)
游博. 词语语义相关度计算研究[D]. 武汉:华中师范大学,2013.

[4] XU Y, FAN X Z, ZHANG F. Semantic Relevancy Computing Based on Hownet[J]. Transactions of Beijing Institute of Technology,2005,25(5):411-414. (in Chinese)
许云,樊孝忠,张锋. 基于知网的语义相关度计算[J]. 北京理工大学学报,2005,25(5):411-414.

[5] WANG H L, LV Q, XU R. Computation model of Chinese semantic relevancy based on HowNet[C]// The National Academic Conference on Information Retrieval and Information Content Security. 2007. (in Chinese)
王红玲,吕强,徐瑞. 一种基于知网的中文语义相关度计算模型[C]//全国信息检索与内容安全学术会议. 2007.

[6] WANG J H, ZUO W L, YAN Z. Word Semantic Similarity Measurement Based on Naive Bayes Model[J]. Journal of Computer Research and Development, 2015, 52(7): 1499-1509. (in Chinese)
王俊华,左万利,闫昭. 基于朴素贝叶斯模型的单词语义相似度度量[J]. 计算机研究与发展,2015,52(7):1499-1509.

[7] AOUICHA M B, TAIEB M A H, HAMADOU A B. Taxonomy-based information content and wordnet-wiktionary-wikipedia glosses for semantic relatedness[J]. Applied Intelligence, 2016, 45(2):1-37.

[8] LI W, YANG C, FU X. Combining How Net and Extension Strategy Generation Method to Improve Customer Values[J]. Procedia Computer Science, 2015, 55:451-460.

[9] XIANG C C, SUI Z F, ZHAN W D. On Mapping between HowNet and CCD[J]. Journal of Chinese Information Processing, 2015, 29(3):44-51. (in Chinese)
向春丞,穗志方,詹卫东. HowNet 与 CCD 映射方法研究[J]. 中文信息学报,2015,29(3):44-51.

[10] KIMTANI D K, CHOUDHURY J, CHAKRABARTY A. Improvement in Word Sense Disambiguation by introducing enhancements in English WordNetStructure [J]. International Journal on Computer Science & Engineering, 2012, 4(7): 1366-1370.

[11] XIAO S, HU J Z, YAO S Y, et al. Objectorient ontology modeling for tag complex sentence[J]. Application Research of Computer, 2010, 27(2):552-554. (in Chinese)
肖升,胡金柱,姚双云,等. 面向对象有标复句本体建模[J]. 计算机应用研究,2010,27(2):552-554.

[12] WANG Z H, WANG L Y, DANG H, et al. Web Clustering Based on Hybrid Probabilistic Latent Semantic Analysis Model [J]. Journal of Computer Applications, 2012, 32(11): 3018-3022. (in Chinese)
王治和,王凌云,党辉,等. 基于混合概率潜在语义分析模型的 Web 聚类[J]. 计算机应用,2012,32(11):3018-3022.

[13] STRUBE B M, PONZETTO S P. WikiRelate! Computing semantic relatedness using Wikipedia[C]// Proc. of AAAI-06. 2015:1419-1424.

[14] WAN F Q, WU Y F. Computing Lexical Semantic relevancy with Chinese Wikipedia[J]. Journal of Chinese Information Processing, 2013, 27(6):31-37, 109. (in Chinese)
万富强,吴云芳. 基于中文维基百科的词语语义相关度计算[J]. 中文信息学报,2013,27(6):31-37,109.

[15] 邢福义. 汉语复句研究[M]. 北京:商务印书馆,2001.

[16] CRISTIANINI N, SHAWE-TAYLOR J, LODHI H. Latent semantic kernels[J]. Journal of Intelligent Information Systems, 2002, 18(2/3):127-152.

(上接第 279 页)

[7] WU J, CUI Z M, SHI Y J, et al. Local Density-based Similarity Matrix Construction for Spectral Clustering [J]. Journal on Communication, 2013(3):14-22. (in Chinese)
吴健,崔志明,时玉杰,等. 基于局部密度构造相似矩阵的谱聚类算法[J]. 通信学报,2013(3):14-22.

[8] HUANG L, LI R, CHEN H, et al. Detecting network com-munities using regularized spectral clustering algorithm[J]. Artificial Intelligence Review, 2014, 41(4):579-594.

[9] SHI J, MAILIK J. Normalized cuts and image segmentation[J]. IEEE transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8):888-905.

[10] LIANG H, XUEPING G U. Black-Start Network Partitioning Based on Spectral Clustering [J]. Power System Technology, 2013, 37(2):372-377.

[11] PAN XY, LIU F, JIAO L C. Density Sensitive Based Multi-Agent Evolutionary Clustering algorithm [J]. Journal of Software, 2010, 21(10):2420-2431. (in Chinese)
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] LIU Y, CAI D, LI C. Density Sensitive Hashing [J]. IEEE Transactions on Cybernetics, 2012, 44(8):1362-1371.

[13] YANG P, ZHU Q, HUANG B. Spectral clustering with density sensitive similarity function [J]. Knowledge-Based Systems, 2011, 24(5):621- 628.

[14] CHUNG F R K. Spectral graph Theory [J]. Regional Conference, 1997, 7(1):158.

[15] MOHAR B. The Laplacian spectrum of graphs[J]. Graph Theory, Combinatorics, and Applications, 1991, 2(7):871- 898.

[16] FAK K. On the theorem of weyl concerning eigenvalues of linear transformations[J]. Proceedings of National Academy of Sciences, 1950, 35(11):652-655.

[17] BOYD, STEPHEN, VANDENBERGHE, et al. Convex Optimization[M]//Cambridge University Press. 2004:1859.