密度自适应的半监督谱聚类算法

周海松 黄德才
(浙江工业大学计算机科学与技术学院 杭州 310023)

摘 要 谱聚类是一种新兴的聚类算法,数据点间的相似度定义对其聚类效果起着至关重要的作用。传统的谱聚类算法通常利用高斯核函数作为相似度函数,但是对于多密度的数据往往不能取得良好的效果。在定义新的相似度函数的基础上,提出了一种密度自适应的半监督聚类算法。该算法结合半监督聚类的成对约束理论,利用先验信息对样本点之间的相似度进行自适应调整,提高了聚类的精度。该算法在人工数据集和真实数据集上的仿真实验都取得了良好的效果。

关键词 密度,半监督,谱聚类
中图法分类号 TP393 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.12.038

Density Self-adaption Semi-supervised Spectral Clustering Algorithm

ZHOU Hai-song HUANG De-cai
(College of Computer Science & Technology, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract As an emerging clustering algorithm, the similarity definition of spectral clustering between data points plays an important role in its clustering results. Traditional spectral clustering algorithms typically use gaussian kernel function to be similarity function, but it doesn't make great effects on multidimensional data. On the basis of defining the new similarity function, a density self-adaption semi-supervised clustering algorithm was put forward which is sensitive with density. Combining with constraint theory in pairs of the semi-supervised clustering, the algorithm makes adaptations on similarity between sample points by using priori information, thus improving the accuracy of data. The algorithm achieves good results both in synthetic datasets and real-world datasets.

Keywords Density, Semi-supervised, Spectral clustering

1 引言

聚类作为一种有效的数据分析方法,在图像分割、语音识别、文本挖掘等领域有着广泛应用。传统的聚类算法如K-means算法、EM算法都是基于凸分布样本空间的,当样本空间非凸时,很容易陷入局部最优。而谱聚类具有识别非凸分布聚类的能力,能在任意形状的样本空间上聚类,且收敛于全局最优解。谱聚类算法通过矩阵谱分析理论导出聚类对象的特征,利用新特征对原数据进行聚类。

传统的基于密度的聚类算法如Dbsacn的缺点是对输入的邻域大小参数 e 比较敏感。参数 e 需要人为指定,但是却很难在算法运行前就确定。此外,该类算法无法发现密度相差较大的簇,由于选取的是全局的 e 值,数据对象领域大小是一致的,因此当数据的密度分布不均匀时,若根据较密的数据类选取 e 值,那么相对分布较疏的数据点来说,其领域中的数据对象数目就会小于MinPts,这些稀疏点就无法成为核心点,即无法进行进一步的聚类扩展。同理,根据较疏的数据类选取 e 值也会遇到相似的问题。针对这种情况,谭颖等人[1]提出了多密度阈值的Dbscan改进算法,通过构建网格密度矩阵绘制密度分布图,解决多密度层次的聚类。黄添强等人[2]提出了结构复杂数据的半监督聚类,利用成对限制反映多密度分布信息,计算出不同的领域半径参数 e,再用Dbscan进行聚类。

对于谱聚类的研究主要集中在相似度函数定义、聚类数目确定和特征向量选择等几个方向。其中相似度函数的定义对聚类结果起着决定性作用,Ng[3]等人选择高斯核函数即 $S(i,j)=\exp(-\|x_i-x_j\|^2/2\delta^2)$ 作为相似度函数提出了NJW算法,其中聚类数目 K 为人工指定参数。Zelnik-Manor等人[4]提出了一种自适应调节的谱聚类算法,在高斯核函数的基础上定义点 x_i 到其第 k 个近邻的欧氏距离作为自适应参数 δ_i,得到相似度函数 $S(i,j)=\exp(-\|x_i-x_j\|^2/\delta_i\delta_j)$。为了使相似度函数能自适应多密度数据,刘馨月等人[5]用两点的共享近邻数目表征局部密度,结合自适应调节的高斯核函数,提出了基于共享近邻的自适应谱聚类算法,该算法具有局部密度的自适应性,能有效识别数据点之间的内在联系。

传统的谱聚类算法通常被认为是无监督的,即在没有任何数据的先验信息下对数据进行聚类分析,但是这种聚类算法有时不能达到良好的聚类效果。在许多实际问题中,很容

到稿日期:2015-10-21 返修日期:2016-03-27 本文受水利部公益性行业科研专项(201401044)资助。
周海松(1991—),男,硕士生,主要研究方向为数据挖掘;黄德才(1958—),男,博士,教授,博士生导师,主要研究方向为数据挖掘、人工智能等,E-mail:hdc@zjut.edu.cn(通信作者)。

易获得少部分数据的先验知识，如类标签和数据点间的成对约束关系。因此可以利用这些具有先验知识的数据来辅助聚类，提高聚类的精度和效率。王玲等人[6]结合成对约束和谱聚类提出了密度敏感的半监督谱聚类。王娜等人[7]提出了基于监督信息特性的主动半监督谱聚类算法。

本文针对多密度数据，在基于共享近邻的自适应谱聚类算法上进行改进，提出了一种新的相似度函数，并且引入半监督聚类里的成对约束理论，利用数据点的先验信息进一步调整样本点间的相似度，提高了谱聚类中相似度的精确性。仿真实验表明，该算法在人工数据集和真实数据集上都能取得良好的效果。

2 相关算法和理论

2.1 谱聚类原理

谱聚类算法将数据集中的对象看作无向图中的顶点，对象之间的相似度作为相应连接边的权值，然后将聚类问题转化为图的划分问题。基于图论的最优划分准则使划分成的子图内部相似度最大，不同子图之间的相似度最小。

根据不同的准则函数及谱映射方法，谱聚类的实现方式有很多，但都可以归纳为以下 3 个步骤：1)构造相似度矩阵，得到 Laplacian 矩阵 L；2)计算 L 的前 K 个特征向量，构建特征向量空间；3)利用经典聚类算法对特征向量空间进行聚类。

与传统聚类算法相比，谱聚类算法能够识别非凸分布聚类，且收敛于全局最优解，非常适用于许多实际数据。

2.2 成对约束

Wagstaff 等人最早在文献[8]中引入 must-link 和 cannot-link 两类成对约束来进行半监督聚类。现有如下规定：若两数据点为 must-link，则这两个数据点必须在同一聚类中；若两数据点为 cannot-link，则它们在聚类时必须分配在不同类中。

must-link 和 cannot-link 这两类成对约束作为样本的先验信息在实际聚类中是非常容易得到的，但是一般数量有限。因此需要对这些成对约束进行扩展。Klein 等人[9]提出 must-link 和 cannot-link 在样本上具备一组二值传递关系：

$$\begin{cases}(x_i,x_j)\in must\text{-}link \,\&\&\, (x_j,x_k)\in must\text{-}link \Rightarrow \\ \quad (x_i,x_k)\in must\text{-}link \\ (x_i,x_j)\in must\text{-}link \,\&\&\, (x_j,x_k)\in cannot\text{-}link \Rightarrow \\ \quad (x_i,x_k)\in cannot\text{-}link\end{cases}$$

利用上述传递关系，通过数据集中的成对约束关系，可得数据集的闭包关系。

同类闭包：对 3 个样本 (a_1,a_2,a_3)，如果 (a_1,a_2) 属于 must-link，(a_2,a_3) 属于 must-link，则 (a_1,a_2,a_3) 构成的集合为同类闭包，简称闭包。

异类闭包：两个闭包 $A(a_1,a_2,\cdots,a_p)$，$B(b_1,b_2,\cdots,b_q)$，如果 (a_i,b_j) 属于 cannot-link，a_i 属于 A 且 b_j 属于 B，则称 A，B 互为异类闭包。

属于同类闭包的数据点互为 must-link，属于异类闭包的数据点互为 cannot-link。

通过对现有的成对约束求闭包和异类闭包，可以扩展 must-link 和 cant-link 这两类成对约束的数量，从而最大化地利用样本的先验信息。

3 密度自适应的半监督谱聚类

3.1 相似度定义

传统谱聚类通常用高斯核函数来定义相似度，但在面对多密度数据集时往往无法得到良好的聚类结果，针对这类问题，本文在相似度定义时加入了数据点邻域内的密度信息和数据点间的共享近邻信息，然后通过成对约束信息来更新数据点之间的相似度，进一步提高数据点相似度的准确性。

3.1.1 密度敏感的相似度

定义 1 数据集中具有成对约束关系的数据点集为约束集，其中的数据点称为约束点；数据集中没有成对约束关系的数据点集为普通集，其中的数据点称为普通点。

定义 2 约束集和普通集中数据点密度参数。

(1)约束点密度参数

$$\varepsilon_i=\frac{1}{K}\sum_{x_j'\in N_i'}d(x_i',x_j') \tag{1}$$

其中，x_i' 属于约束集，N_i' 为 x_i' 距离最近的 K 个数据点集合，$d(x_i',x_j')$ 为 x_i' 和 x_j' 的欧氏距离，ε_i 反映了 x_i' 周围的密度分布情况，ε_i 越大说明 x_i' 周围的点越分散，ε_i 越小说明越紧密。同理可得普通集中数据点的尺度参数 σ_i。

(2)普通点密度参数

$$\sigma_i=\frac{1}{K}\sum_{x_j\in N_i}d(x_i,x_j) \tag{2}$$

定义 3 约束点 x 的周围欧氏距离小于其密度参数的普通点集合为约束点 x 的普通邻域集。

定义 4 共享近邻数 $SNN(x_i,x_j)$ 为数据点 x_i 的前 d 个数据点集合 $N(x_i)$ 和数据点 x_j 的前 d 个数据点集合 $N(x_j)$ 的交集的点个数，即 $SNN(x_i,x_j)=|N(x_i)\cap N(x_j)|$。$SNN(x_i,x_j)$ 越大，说明两个数据点为同一类的概率也越大。

定义 5(密度敏感相似度)

$$A_{ij}=1/\left(1+\left(\frac{d(x_i,x_j)}{(1+SNN(x_i,x_j))}*\left(1+\frac{|\sigma_i-\sigma_j|}{\bar{\varepsilon}}\right)\right)\right) \tag{3}$$

其中，$d(x_i,x_j)$ 为数据点间的欧氏距离，$SNN(x_i,x_j)$ 为数据点间的共享近邻数，$|\sigma_i-\sigma_j|$ 表示普通点间的密度参数差，$\bar{\varepsilon}$ 表示约束集中约束点的密度参数均值。

3.1.2 相似度更新

(1)更新约束点间相似度

$$A_{ij}=\begin{cases}A_{ij}=1, & \text{if } (x_i,x_j)\in must\text{-}link \\ A_{ij}=0, & \text{if } (x_i,x_j)\in cannot\text{-}link\end{cases} \tag{4}$$

(2)更新不同约束点普通邻域集内的普通点之间的相似度

在半监督学习中，聚类假设和流形假设是两种最常用来建立联系的假设。聚类假设要求预测模型对相同聚类中的数据应该给出相同的类别标记；流形假设是指相邻的样本具有相似的性质，其标记也应该相似。

1)如果 x_i 和 x_j 属于 must-link，则更新 x_i 普通邻域集内的普通点和 x_j 普通邻域集内的普通点之间的相似度：

$$A_{ij}=1/\left(1+\left(\frac{\ln(1+d(x_i,x_j))}{SNN(x_i,x_j)}*\left(1+\frac{|\sigma_i-\sigma_j|}{\left(\frac{\varepsilon_m+\varepsilon_c}{2}\right)}\right)\right)\right) \tag{5}$$

其中，ε_m 和 ε_c 表示这对 must-link 中两个约束点的密度参数。因为根据流行假设，容易得到属于 must-link 的一对数据点，它们周围的点也非常有可能是属于同一类的，所以通过 $\ln(1+d(x_i,x_j))$ 来增大它们之间的相似度。

2)如果 x_i 和 x_j 属于 cannot-link，则更新 x_i 普通邻域集内的普通点和 x_j 普通邻域集内的普通点之间的相似度：

$$A_{ij}=1/(1+(\frac{e^{d(x_i,x_j)}}{SNN(x_i,x_j)}*(1+\frac{|\sigma_i-\sigma_j|}{(\frac{\varepsilon_m+\varepsilon_c}{2})})) \quad (6)$$

其中，ε_m 和 ε_c 表示这对 cannot-link 中两个约束点的密度参数。因为根据流行假设，容易得到属于 cannot-link 的一对数据点，它们周围的点也非常有可能不是属于同一类的，所以通过 $e^{d(x_i,x_j)}$ 来减小它们之间的相似度。

3.1.3 约束冲突处理

假设约束点 A 和约束点 D 为 must-link，约束点 C 和约束点 D 为 cannot-link，普通点 B 在约束点 A 的普通邻域集内，也在约束点 C 的普通邻域集内。这时计算普通点 B 与约束点 D 普通邻域集内普通点之间的相似度就遇到了困难，因为无法确定应用 must-link 公式还是 cannot-link 公式，称这种现象为约束冲突，如图 1 所示。本文的处理方法为：把普通点 B 分给距离其最近的约束点，若 $|AB|<|BC|$，则用 must-link 公式来计算。

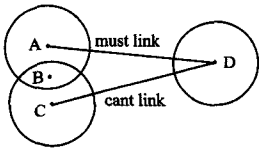


图 1 约束冲突处理

3.2 算法步骤

本文参考 NJW 提出的谱聚类算法步骤如下。

输入：n 个数据点 $S=\{s_i\}_{i=1}^n$，聚类数目 C

输出：C 个聚类 $C_1,\cdots,C_C$

Step1 利用式(3)计算数据集内点间的密度敏感相似度。

Step2 利用式(4)更新约束点间的相似度。

Step3 用式(5)、式(6)更新约束点周围的普通点间的相似度。

Step4 构造拉普拉斯矩阵 $L=D^{-1/2}AD^{-1/2}$，其中 D 为对角矩阵，对角元素为 $D_{ii}=\sum_{j=1}^{n}A_{ij}$。

Step5 计算 L 的前 K 个特征向量 $u_1,u_2,\cdots,u_k$，这些特征向量构成 $U=[u_1,u_2,\cdots,u_k]$，其中每一行分别对应原空间中的每个数据点，用 K-means 进行聚类，分成 C 类，即 $C_1,\cdots,C_C$。

Step6 当 U 中第 i 行属于第 c 类时，将数据点划分到第 c 类。

4 实验结果与分析

为了验证算法的有效性，本文分别在人工数据集和真实数据集上分别对标准谱聚类(SC)、基于共享近邻的自适应谱聚类(SNN-ASC)、密度自适应的半监督谱聚类(DS-SSC)进行了对比实验。

4.1 评价指标

为了对聚类算法进行准确评价，实验过程采用 3 种不同的评价指标：准确率指标、熵和 CRI 指标。

4.1.1 准确率

准确率描述为：

$$Accuracy=\sum_{i=1}^{k}X_i/n \quad (7)$$

其中，X_i 表示应该分到第 i 类中，且正确分到了第 i 类中的数据点个数，n 为原始数据集中样本数据点的个数。

4.1.2 熵

一个簇的 S_r 熵为：

$$E(S_i)=-\frac{1}{\log k}\sum_{i=1}^{k}\frac{X_i}{N_i}\log\frac{X_i}{N_i} \quad (8)$$

其中，S_i 是第 i 个簇，X_i 为正确分到 i 类的数据点数目，N_i 为第 i 类的数据点数目，k 是数据集的类数。各个簇的加权和即为聚类的整体熵：

$$Entropy=\sum_{i}^{k}\frac{N_i}{N}E(S_i) \quad (9)$$

由上述公式可知，熵越小，聚类效果越好。

4.1.3 CRI 指标

对半监督算法的评价指标往往采用文献[10]提出的评价函数：

$$CRI=\frac{\#CD-\#CN}{n(n-1)/2-\#CN} \quad (10)$$

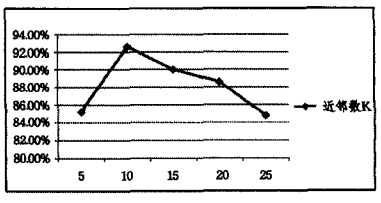
其中，$\#CD$ 表示由算法所获得的正确决策对数，$\#CN$ 表示已知成对约束的数目，包括用户直接提供的和通过计算闭包扩展而得的。$n(n-1)/2$ 表示数据集内的样本对数。在评价指标中减去已知成对约束数是因为半监督聚类算法中，已知的监督信息是不能反映聚类算法的效果的。

4.2 参数选择

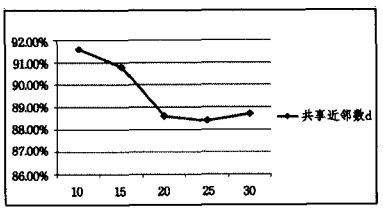
DS-SSC 主要涉及到的参数有近邻数 K 和共享近邻数 d。对 K 的选取既不能太大也不能太小，太小会因为邻域内点太少而无法充分反映其密度分布情况；选取太大会使约束冲突现象大量存在，对相似度的计算造成偏差。

4.2.1 近邻数 K 的选取

以数据量为 150 的 UCI 数据集 Iris 为例，计算取不同近邻数 K 时的聚类准确率。由图 2(a)可以看出，近邻数为 10，即数据总数量的 7%时，准确率最高。



(a)近邻数 K



(b)共享近邻数 d

图 2 参数选择

4.2.2 共享近邻数 d 的选取

仍以 Iris 数据集为例，计算共享近邻数为 d 时聚类的准确率。由图 2(b)可以看出，共享近邻数为 10 时，聚类准确率最高。

4.3 人工数据集

本文选择了两个多重密度的人工数据集，从图 3 可以看出传统的谱聚类在面对这类数据时都不能获得良好的效果，基于共享近邻的谱聚类虽然比传统谱聚类效果更佳，但是还存在误分点，而本文提出的密度自适应的谱聚类能够达到良好的效果。

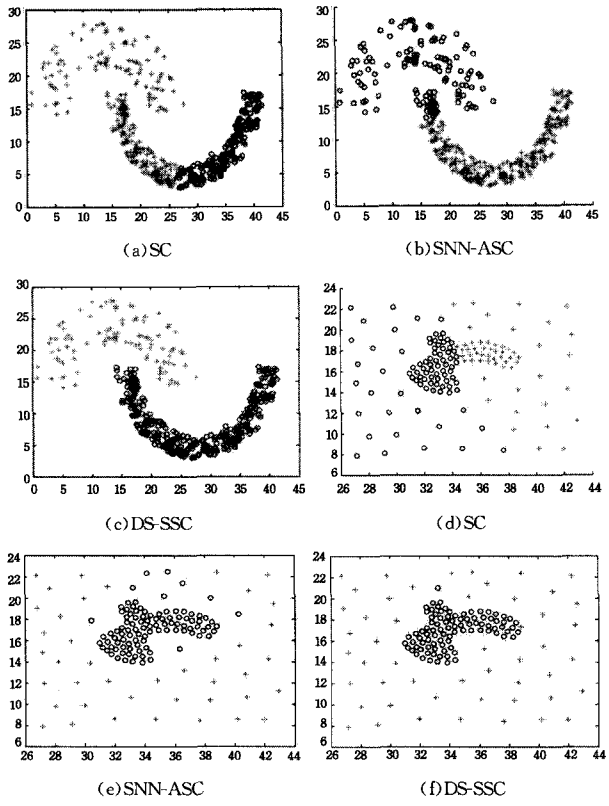


图 3　聚类结果对比

4.4　真实数据集

真实数据集来自 UCI 公共数据库，Iris 有 4 个属性，150 个数据；Wine 有 13 个属性，178 个数据；Ionosphere 有 34 个属性，351 个数据；Glass 有 10 个属性，214 个数据。因为这 4 个数据集原本都是用于分类研究的，因此每个数据集都带有分类标签属性。

本文针对上述两个数据，从正确率(Accuracy)、熵(Entropy)等方面比较 K-means，SC，SNN-ASC，DS-SSC 算法来评估聚类算法的效果，对比结果如表 1、表 2 所列。

表 1　真实数据聚类结果正确率对比(%)

数据集	K-means	SC	SNN-ASC	DS-SSC
Iris	89.3	89.3	90	92.6
Wine	70.2	69.5	71.4	72.4
Ionosphere	63.2	63.3	65.5	66.3
Glass	68.2	69.3	72.2	73.2

表 2　真实数据聚类结果熵对比

数据集	K-means	SC	SNN-ASC	DS-SSC
Iris	0.272	0.269	0.205	0.196
Wine	0.322	0.317	0.245	0.196
Ionosphere	0.706	0.703	0.691	0.672
Glass	0.399	0.382	0.334	0.322

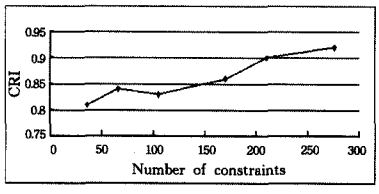


图 4　Iris 的 CRI 指标与成对约束数量变化的关系

从表 1 中可以看出，本文算法的准确率确实比其它几类谱聚类的准确率高。再取不同成对约束的数量采用 CRI 指标对 DS-SSC 算法进行评估，对图 4 分析可得，随着成对约束数量的增加，算法的聚类效果整体为上升趋势。但是由于约束冲突现象的存在，有时成对约束数量增加，聚类效果反而会下降。不过随着成对约束数量的增加，CRI 的值最终趋于 1。

结束语　由于谱聚类对输入的相似度矩阵非常敏感，因此得到一个能够真实反映数据点分布的相似度矩阵非常重要。本文提出了密度自适应的半监督谱聚类算法 DS-SSC，针对普通谱聚类无法得到良好效果的多密度数据，通过密度敏感相似度计算得到相似度矩阵。对于许多真实数据，获得少量样本点的成对约束关系往往很容易，结合这一点，本文对具有成对约束关系的数据点和其领域内数据点之间的相似度进行更新，从而获得了最符合实际数据分布的相似度矩阵，提高了聚类准确度。本文的算法复杂度和传统的谱聚类的算法复杂度同阶，均为 $O(n^3)$，面对大数据时的计算量较大。因此如何降低其计算复杂度，使其应用在大数据中将是下一步研究的重点。

参 考 文 献

[1] Tang Ying, Hu Rui-fei, Yin Guo-fu. Adapted DBSCAN with multi-threshold[J]. Journal of Computer Applications, 2008, 28(3): 745-748(in Chinese)
谭颖，胡瑞飞，殷国富. 多密度阈值的 DBSCAN 改进算法[J]. 计算机应用，2008，28(3)：745-748

[2] Huang Tian-qiang, Yu Yang-qiang, Qin Xiao-lin. Semi-supervised clustering for complicated data[J]. Control and Decision, 2010, 25(1): 14-19(in Chinese)
黄添强，余养强，秦小麟. 结构复杂数据的半监督聚类[J]. 控制与决策，2010，25(1)：14-19

[3] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: analysis and analgorithm[C]// Proceeding of NIPS. 2002: 849-856

[4] Zelnik-Manor L, Perona P. Self-tuning spectral clustering[C]// Proceeding of NIPS. 2005: 1601-1608

[5] Liu Xin-yue, Li Jing-wei, Yu Hong, et al. Adaptive Spectral Clustering Based on Shared Nearest Neighbors[J]. Journal of Chinese Computer Systems, 2011(9): 1876-1880(in Chinese)
刘馨月，李静伟，于红，等. 基于共享近邻的自适应谱聚类[J]. 小型微型计算机系统，2011(9)：1876-1880

[6] Wang Ling, Bo Lie-feng, Jiao Li-cheng. Density-Sensitive Semi-Supervised Spectral Clustering[J]. Journal of Software, 2007, 18(10): 2412-2422(in Chinese)
王玲，薄列峰，焦李成. 密度敏感的半监督谱聚类[J]. 软件学报，2007，18(10)：2412-2422

[7] Wang Na, Li Xia. Active Semi-supervised Spectral Clustering Based on Pairwise Constraints[J]. Acta Electronica Sinica, 2010(1): 172-176(in Chinese)
王娜，李霞. 基于监督信息特性的主动半监督谱聚类算法[J]. 电子学报，2010(1)：172-176

[8] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means clustering with background knowledge[C]// Proc. of the 18th Int'l Conf. on Machine Learning. 2001

[9] Klein D, Kamvar S D, Manning C D. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering[C]// Proc. of the 19th Int'l Conf. on Machine Learning. 2002

[10] Zhang Wei-jiao, Liu Chun-huang, Li Fang-yu. Method of Quality Evaluation for Clustering[J]. Computer Engineering, 2005, 31(20): 10-12(in Chinese)
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