文章编号: 1001-0920(2011)08-1147-06

一种基于近邻传播算法的最佳聚类数确定方法

周世兵[a], 徐振源[a,b], 唐旭清[b]

(江南大学 a. 物联网工程学院, b. 理学院, 江苏 无锡 214122)

摘 要: 在聚类分析中, 决定聚类质量的关键是确定最佳聚类数. 对此, 从样本几何结构的角度定义了样本聚类距离和样本聚类离差距离, 设计了一种新的聚类有效性指标. 在此基础上, 提出一种基于近邻传播算法确定样本最佳聚类数的方法. 理论研究和实验结果表明, 所提出的指标和方法能够有效地对聚类结果进行评估, 适合于确定样本的最佳聚类数.

关键词: 近邻传播; 聚类数; 聚类有效性指标; 聚类分析

中图分类号: TP18 **文献标识码:** A

Method for determining optimal number of clusters based on affinity propagation clustering

ZHOU Shi-bing[a], *XU Zhen-yuan*[a,b], *TANG Xu-qing*[b]

(a. School of IoT Engineering, b. School of Science, Jiangnan University, Wuxi 214122, China. Correspondent: ZHOU Shi-bing, E-mail: worldguard@sina.com)

Abstract: It is crucial to determine optimal number of clusters for the quality of clustering in cluster analysis. From the standpoint of sample geometry, two concepts of sample clustering distance and sample clustering deviation distance are defined, and a new clustering validity index is designed. In addition, a method for determining optimal number of clusters based on affinity propagation clustering algorithm is proposed. Theoretical research and experimental results show that the proposed index and method can evaluate the clustering results effectively, and be suitable for determining optimal number of clusters.

Key words: affinity propagation; number of clusters; clustering validity index; cluster analysis

1 引 言

聚类是数据分析中广泛使用的重要方法, 它按照特定标准, 对各个待分类的样本进行分类, 使得类内相似度尽可能大, 同时类间相似度尽可能小. 在聚类分析中, 决定聚类质量的关键是确定最佳聚类数. 目前, 大部分聚类算法需要预先给定聚类数, 才能对样本进行聚类分析. 而如何得到正确的最佳聚类数, 一直是聚类有效性研究的重要课题. 通常情况下, 确定最佳聚类数算法的基本思想是: 针对具体的数据集, 在确定的聚类数搜索范围内运行特定的聚类算法, 得到不同聚类数目的聚类结果, 选择合适的有效性指标对聚类结果进行评估, 并根据评估结果确定最佳聚类数. 由此可知, 确定最佳聚类数的核心环节是聚类算法和有效性指标的选择.

现有的最佳聚类数确定方法主要是围绕 K-均值算法和模糊 C-均值 (FCM) 算法等常用的聚类算法进行的. 由于这些聚类算法的聚类结果一般依赖于初始聚类中心的选择, 不同的初始聚类中心会产生不同的聚类结果, 从而导致聚类结果不稳定. 采用有效性指标对不稳定的聚类结果进行评估, 得到的最佳聚类数也是不确定的, 因此, 传统的 K-均值算法和FCM算法不太适合作为确定最佳聚类数的有效算法.

近年来, Frey 等人[1]提出了一种新的聚类算法, 称为近邻传播聚类算法 (AP算法). 与以往的聚类方法相比, 该方法能更快地处理大规模数据, 得到较好的聚类结果. 该算法通过数据点之间的消息传递产生高质量的聚类中心, 避免了聚类中心的初始选择, 使得聚类结果比较稳定. 文献 [1] 中将近邻传播聚类

收稿日期: 2010-04-24; **修回日期:** 2010-09-10.

基金项目: 国家自然科学基金项目(60703106); 中央高校基本科研业务费专项资金项目(JUSRP21012).

作者简介: 周世兵(1972–), 男, 讲师, 博士生, 从事人工智能、模式识别等研究; 徐振源(1946–), 男, 教授, 博士生导师, 从事混沌、同步控制和人工智能等研究.

算法应用于人脸图像聚类、基因表达数据的基因识别、手写体字符识别以及最优航空路线确定等问题. 实验结果表明, 近邻传播聚类算法在很短的时间内便能得到K中心算法花费很长时间才能达到的聚类结果[1-2].

目前, 有许多有效性评价指标可用来分析聚类结果并确定最佳聚类数. 其中性能较优的指标主要有Davies-Bouldin (DB) 指标[3], Krzanowski-Lai (KL) 指标[4], Homogeneity-Separation (HS) 指标[5]和In-Group Proportion (IGP) 指标[6]等. 但这些指标由于自身的缺陷, 对于聚类结构难以判别的情况, 其聚类有效性检验效果不够理想, 很难得到正确的最佳聚类数. 对此, 本文设计了一种新的基于样本几何结构的有效性指标, 在此基础上, 提出一种确定样本最佳聚类数的方法, 用来评估近邻传播聚类算法的聚类结果和确定样本的最佳聚类数. 理论研究和实验结果表明, 与其他指标和方法相比, 本文提出的新指标和方法具有更好的性能和可行性.

2 近邻传播聚类算法

近邻传播聚类 (AP)[1,7]算法是一种基于近邻信息传播的聚类算法, 其目的是找到最优的类代表的集合, 使得所有样本到最近的类代表的相似度之和最大. AP算法首先将数据集的所有N个样本都视为候选的类代表, 为每个样本建立与其他样本的吸引程度信息, 即任意2个样本x_i和x_k之间的相似度 (采用欧氏距离为测度时$s(i,k)=-\|x_i-x_k\|^2$) 被存储在$N\times N$的相似度矩阵中. AP算法用$s(i,k)$表示样本x_k在多大程度上适合作为样本x_i的类代表. 初始假设所有样本被选中成为类代表的可能性相同, 即设定所有$s(k,k)$为相同值p. 为选出合适的类代表, 需不断从样本中搜集有关证据, 为此, AP算法引入了两个重要的信息量参数: 可信度r和可用度a, 这两个信息量代表了不同的竞争目的. $r(i,k)$是从x_i指向x_k, 它代表x_k积累的证据, 用来表示x_k适合作为x_i的类代表的代表程度; $a(i,k)$是从x_k指向x_i, 它代表x_i积累的证据, 用来表示x_i选择x_k作为类代表的合适程度. 对于任意样本x_i, 计算所有样本的可信度$r(i,k)$与可用度$a(i,k)$之和, 则二者之和最大的样本x_k为类代表. AP算法的迭代过程就是两个信息量交替更新的过程.

为防止迭代过程中出现震荡, AP算法引入了防止震荡的因子λ, λ取0～1之间的值. 在本文实验中, 为避免震荡的发生, 设置$\lambda=0.9$. $r(i,k)$和$a(i,k)$的更新结果都是由当前迭代得到的值和上一步迭代的结果通过λ加权得到的. AP算法的基本步骤如下:

1) 初始化. 求解相似度矩阵$[s(i,k)]_{N\times N}$, 设定相似度矩阵对角线元素$s(k,k)$为相同值p, 在无先验知识时将p设定为统一的吸引度中值p_m; 设定初始可信度和可用度均为0, 即$r^{(0)}(i,k)=a^{(0)}(i,k)=0$.

2) 迭代过程如下:

①更新可用度和可信度.

②对所有样本求可信度与可用度之和, 根据$\arg\max\limits_k\{r(i,k)+a(i,k)\}$找到每个样本的类中心样本.

③判断信息迭代过程是否满足停止条件, 即: 超过最大迭代数; 信息改变量低于某一阈值; 类中心在连续几步迭代过程中保持稳定. 满足条件之一即可.

3) 输出聚类结果.

AP算法不能直接将指定类数K作为算法的输入参数, 以使算法产生K个聚类的聚类结果. 要获得指定类数的聚类结果, 一般采用搜索的方法. 文献[1]给出一种通过对分法搜索AP算法, 实现了指定类数的聚类分析.

3 新聚类有效性指标

评价聚类结果优劣的过程称为聚类有效性分析. 一般来说, 一个好的聚类划分应尽可能反映数据集的内在结构, 使类内样本尽可能相似, 类间样本尽可能不相似. 从距离测度考虑, 使类内距离极小化而类间距离最大化的聚类是最优聚类; 从相似测度考虑, 使类内相似度极大化而类间相似度最小化的聚类是最优聚类. 目前已提出了一些聚类有效性指标, 但由于这些指标自身的缺陷, 一般难以找到正确的最佳聚类数. 鉴于这种情况, 本文设计了一种新的聚类有效性指标, 该指标可以对AP算法的聚类结果进行评估, 并可用来确定样本的最佳聚类数.

3.1 新指标及相关概念定义

定义 1 令$K=\{X,R\}$为聚类空间, 其中$X=\{x_1,x_2,\cdots,x_n\}$. 假设n个样本对象被聚类为c类, 定义第j类的第i个样本的最小类间距离$\mathrm{bd}(j,i)$为该样本到其他每个类中样本平均距离的最小值, 即

$$\mathrm{bd}(j,i)=\min_{1\leqslant k\leqslant c,k\neq j}\left(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|\right). \tag{1}$$

其中: k和j表示类标, $x_i^{(j)}$表示第j类的第i个样本, $x_p^{(k)}$表示第k类的第p个样本, n_k表示第k类中的样本个数, $\|\cdot\|$表示欧氏距离.

定义 2 令$K=\{X,R\}$为聚类空间, 其中$X=\{x_1,x_2,\cdots,x_n\}$. 假设n个样本对象被聚类为c类, 定义第j类的第i个样本的类内距离$\mathrm{wd}(j,i)$为该样本到第j类中其他所有样本的平均距离, 即

$$\mathrm{wd}(j,i)=\frac{1}{n_j-1}\sum_{q=1,q\neq i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|. \tag{2}$$

其中: $x_q^{(j)}$表示第j类的第q个样本, 且$q\neq i$; n_j表示

第 j 类中的样本个数.

定义 3　令 $K=\{X,R\}$ 为聚类空间, 其中 $X=\{x_1,x_2,\cdots,x_n\}$. 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的聚类距离 $\mathrm{bawd}(j,i)$ 为该样本的最小类间距离和类内距离之和, 即

$$\mathrm{bawd}(j,i)=\mathrm{bd}(j,i)+\mathrm{wd}(j,i)=\min_{1\leqslant k\leqslant c,k\neq j}\Big(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|\Big)+\frac{1}{n_j-1}\sum_{q=1,q\neq i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|. \tag{3}$$

定义 4　令 $K=\{X,R\}$ 为聚类空间, 其中 $X=\{x_1,x_2,\cdots,x_n\}$. 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的聚类离差距离 $\mathrm{bswd}(j,i)$ 为该样本的最小类间距离和类内距离之差, 即

$$\mathrm{bswd}(j,i)=\mathrm{bd}(j,i)-\mathrm{wd}(j,i)=\min_{1\leqslant k\leqslant c,k\neq j}\Big(\frac{1}{n_k}\sum_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|\Big)-\frac{1}{n_j-1}\sum_{q=1,q\neq i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|. \tag{4}$$

定义 5　令 $K=\{X,R\}$ 为聚类空间, 其中 $X=\{x_1,x_2,\cdots,x_n\}$. 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的类间类内划分 (BWP) 指标 $\mathrm{BWP}_d(j,i)$ 为该样本的聚类离差距离与聚类距离的比值, 即

$$\mathrm{BWP}_d(j,i)=\frac{\mathrm{bswd}(j,i)}{\mathrm{bawd}(j,i)}=\frac{\mathrm{bd}(j,i)-\mathrm{wd}(j,i)}{\mathrm{bd}(j,i)+\mathrm{wd}(j,i)}=\frac{\min\limits_{1\leqslant k\leqslant c,k\neq j}\Big(\frac{1}{n_k}\sum\limits_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|\Big)-\frac{1}{n_j-1}\sum\limits_{q=1,q\neq i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|}{\min\limits_{1\leqslant k\leqslant c,k\neq j}\Big(\frac{1}{n_k}\sum\limits_{p=1}^{n_k}\|x_p^{(k)}-x_i^{(j)}\|\Big)+\frac{1}{n_j-1}\sum\limits_{q=1,q\neq i}^{n_j}\|x_q^{(j)}-x_i^{(j)}\|}\to\leftarrow. \tag{5}$$

以上指标及相关概念是基于距离测度进行定义的. 如果基于相似测度进行定义, 则应作适当调整. 由于距离与不相似度是对应的, 可将以上定义中的距离改为不相似度, 将以上所定义的最小类间距离、类内距离、聚类距离和聚类离差距离分别对应为最小类间不相似度、类内不相似度、聚类不相似度和聚类离差不相似度. 为了使指标的性能不受所采用的评估测度影响, 结合样本的不相似度作如下具体定义:

定义 6　令 $K=\{X,R\}$ 为聚类空间, 其中 $X=\{x_1,x_2,\cdots,x_n\}$. 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的最小类间不相似度 $\mathrm{bs}(j,i)$ 为该样本到其他每个类中样本平均不相似度的最小值, 即

$$\mathrm{bs}(j,i)=\min_{1\leqslant k\leqslant c,k\neq j}\Big(\frac{1}{n_k}\sum_{p=1}^{n_k}H(x_p^{(k)},x_i^{(j)})\Big), \tag{6}$$

其中 $H(\cdot)$ 表示不相似度.

定义 7　令 $K=\{X,R\}$ 为聚类空间, 其中 $X=\{x_1,x_2,\cdots,x_n\}$. 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的类内不相似度 $\mathrm{ws}(j,i)$ 为该样本到第 j 类中其他所有样本的平均不相似度, 即

$$\mathrm{ws}(j,i)=\frac{1}{n_j-1}\sum_{q=1,q\neq i}^{n_j}H(x_q^{(j)},x_i^{(j)}). \tag{7}$$

定义 8　令 $K=\{X,R\}$ 为聚类空间, 其中 $X=\{x_1,x_2,\cdots,x_n\}$. 假设 n 个样本对象被聚类为 c 类, 定义第 j 类的第 i 个样本的类间类内划分 (BWP) 指标 $\mathrm{BWP}_s(j,i)$ 为该样本的聚类离差不相似度与聚类不相似度的比值, 即

$$\mathrm{BWP}_s(j,i)=\frac{\mathrm{bsws}(j,i)}{\mathrm{baws}(j,i)}=\frac{\mathrm{bs}(j,i)-\mathrm{ws}(j,i)}{\mathrm{bs}(j,i)+\mathrm{ws}(j,i)}=\frac{\min\limits_{1\leqslant k\leqslant c,k\neq j}\Big(\frac{1}{n_k}\sum\limits_{p=1}^{n_k}H(x_p^{(k)},x_i^{(j)})\Big)-\frac{1}{n_j-1}\sum\limits_{q=1,q\neq i}^{n_j}H(x_q^{(j)},x_i^{(j)})}{\min\limits_{1\leqslant k\leqslant c,k\neq j}\Big(\frac{1}{n_k}\sum\limits_{p=1}^{n_k}H(x_p^{(k)},x_i^{(j)})\Big)+\frac{1}{n_j-1}\sum\limits_{q=1,q\neq i}^{n_j}H(x_q^{(j)},x_i^{(j)})}\to\leftarrow. \tag{8}$$

其中: $\mathrm{baws}(j,i)$ 表示第 j 类的第 i 个样本的聚类不相似度, 其值为 $\mathrm{bs}(j,i)+\mathrm{ws}(j,i)$; $\mathrm{bsws}(j,i)$ 表示第 j 类的第 i 个样本的聚类离差不相似度, 其值为 $\mathrm{bs}(j,i)-\mathrm{ws}(j,i)$.

3.2　新指标分析

为了反映聚类结构的类内紧密性和类间分离性, 本文提出了 BWP 指标. BWP 指标基于样本的几何结构, 以数据集中的某个样本作为研究对象, 对聚类结果进行有效性分析. 为便于说明该指标及相关概念的意义, 这里借鉴了文献 [8] 的指标说明方法, 结合图 1 的聚类结构分布示意图进行说明.

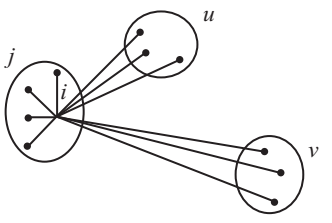

图 1　聚类结构分布示意图

在图 1 中, 数据集中的所有样本被分为 3 类, 分

别是 j,u,v, 在第 j 类中有一个样本 i. 在样本 i 的类内结构方面, 根据定义 2, 样本 i 到类 j 中所有样本距离的平均值, 称为样本 i 的类内距离. 相比于将样本 i 到类 j 的中心之间的距离作为类内距离而言, 定义 2 更准确, 更能反映样本 i 和类 j 中其他样本的结构关系. 在样本 i 的类间结构方面, 根据定义 1, 将样本 i 到其他每个类 (u 和 v) 中所有样本距离的平均值作为样本 i 的类间距离. 相比于将样本 i 到类 u 和类 v 的中心之间的距离作为类间距离而言, 定义 1 更准确, 更能反映样本 i 与类 u 和类 v 的结构关系. 同样, 相比于将样本 i 到每个类 (j, u 和 v) 的中心的不相似度作为计算依据而言, 定义 6 和定义 7 更准确, 更能反映样本 i 与每个类的结构关系. 为了反映类间的分离性, 研究样本 i 的近邻聚类与样本 i 的关系. 近邻聚类可通过样本 i 的最小类间距离或最小类间不相似度所对应的聚类得到, 在图 1 中类 u 便是样本 i 的近邻聚类. 因为如果类 u 满足类间远离的要求, 则其他聚类也一定满足要求. 另外, 如果样本 i 没有聚类到类 j, 则类 u 就是它的最佳选择. 因此, 研究样本 i 所属的聚类 j 和样本 i 的近邻聚类 u 具有重要意义.

确定聚类有效性的标准是使聚类结果达到类内紧密、类间远离. 基于距离测度考虑, 从类内紧密的角度出发, 希望样本的类内距离 $\mathrm{wd}\,(j,i)$ 越小越好; 从类间远离的角度出发, 希望样本离近邻聚类的距离, 即最小类间距离 $\mathrm{bd}\,(j,i)$ 越大越好. 为了综合这两种因素, 可使用线性组合方式平衡二者, 并使函数的目标一致. 使用聚类离差距离 $\mathrm{bswd}\,(j,i)$ (即 $\mathrm{bd}\,(j,i)+(-\mathrm{wd}\,(j,i))$) 来评价聚类结果, 显然 $\mathrm{bswd}\,(j,i)$ 越大, 说明该样本聚类效果越好. 为了使指标能够对所有样本进行有效性分析, 并使指标不受量纲影响, 这里引入了样本聚类距离的概念. 通过样本聚类距离对单个样本的聚类离差距离进行压缩, 使指标成为无量纲量, 指标的值为样本单位聚类距离上的离差距离, 指标值的范围为 $[-1,1]$. 基于距离测度的分析表明, BWP 指标具有一定的合理性. 同样, 基于相似测度分析 BWP 指标, 可以得到和距离测度相同的结果.

定义 5 和定义 8 所定义的 BWP 指标, 两者的基本原理相同, 都是基于某一样本所属的聚类与其近邻聚类的关系进行有效性分析, 并且都是指标值越大, 聚类效果越好. 区别在于前者是采用距离测度对聚类结果进行评估, 后者则是采用相似测度进行的.

3.3 新指标与最佳聚类数确定

BWP 指标反映了单个样本的聚类有效性情况, BWP 指标值越大, 说明单个样本的聚类效果越好. 本文通过求某个数据集中所有样本的 BWP 指标值的平均值来分析该数据集的聚类效果. 显然, 平均值越大, 说明该数据集的聚类效果越好, 其最大值所对应的聚类数即为最佳聚类数. 通常情况下, 指标评估测度的选择与数据集的类型有关. 对于一般数据集, 可以基于距离测度采用 $\mathrm{BWP}_d(j,i)$ 进行聚类有效性分析; 对于基因表达数据集, 通常要求基于相似测度, 可以采用 $\mathrm{BWP}_s(j,i)$ 进行有效性分析. 为方便表述, 这里统一用 $\mathrm{BWP}(j,i)$ 表示 $\mathrm{BWP}_d(j,i)$ 和 $\mathrm{BWP}_s(j,i)$. 由此可得到如下公式:

$$\mathrm{avgBWP}(k)=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\mathrm{BWP}(j,i), \tag{9}$$

$$k_{\mathrm{opt}}=\arg\max_{2\leqslant k<n}\{\mathrm{avgBWP}(k)\}. \tag{10}$$

其中: $\mathrm{avgBWP}\,(k)$ 表示数据集聚成 k 类时的平均 BWP 指标值, k_{opt} 表示最佳聚类数.

4 确定最佳聚类数的算法

本文基于 AP 算法以及 BWP 聚类有效性指标, 提出一种新的分析聚类效果、确定最佳聚类数的算法. 算法归纳如下:

1) 选择聚类数的搜索范围 $[k_{\min},k_{\max}]$.

2) For $k=k_{\min}$ to $k_{\max}$:

① 采用对分法调用 AP 算法;

② 针对一般数据集和基因表达数据集, 分别采用式 (5) 和式 (8) 计算单个样本的 BWP 指标值;

③ 利用式 (9) 计算平均 BWP 指标值.

3) 利用式 (10) 计算最佳聚类数.

4) 输出最佳聚类数、有效性指标值和聚类结果.

5 仿真实验与分析

为了检验有效性指标 BWP 和最佳聚类数确定算法的性能, 本文通过 3 组实验共 8 个数据集进行测试, 并与 DB 指标, KL 指标, HS 指标以及 IGP 指标等常用指标进行比较. 实验中聚类数的搜索范围为 $[2,k_{\max}]$, 根据普遍使用的经验规则 $k_{\max}\leqslant\sqrt{n}$, 取 $k_{\max}=\mathrm{Int}\,(\sqrt{n})$.

AP 算法是以 n 个样本形成的相似度矩阵为基础进行聚类的, 通常情况下, AP 算法中相似度的值为负数. 对于一般数据集 (包括 UCI 真实数据集和人工数据集), 两个维数为 d 的样本 x_i 和 x_k 之间的相似度采用欧氏距离为测度, 即 $s(i,k)=-\|x_i-x_k\|^2$. 对于基因表达数据集, 采用普遍使用的 Pearson 相关系数作为相似性测度, 即两个样本 x_i 和 x_k 之间的线性相关系数为

$$R(x_i,x_k)=\frac{\sum_{m=1}^{d}(x_{im}-\overline{x}_i)(x_{km}-\overline{x}_k)}{\sqrt{\sum_{m=1}^{d}(x_{im}-\overline{x}_i)^2}\sqrt{\sum_{m=1}^{d}(x_{km}-\overline{x}_k)^2}}.$$

为避免负数引起计算混乱, 将 $R(x_i, x_k) \in [-1, 1]$ 进行转换, 即设 $P(x_i, x_k) = 1 - (1 + R(x_i, x_k))/2$, 从而使 Pearson 相关系数转换为正的 Pearson 距离[9]. $P(x_i, x_k)$ 越大, 表示两个样本相距越远, 这样, 基因表达数据的相似度可表示为 $s(i, k) = -P(x_i, x_k)$.

使用 BWP 指标对 AP 算法的聚类结果进行有效性分析时, 对于一般数据集 (包括 UCI 真实数据集和人工数据集), 应采用基于欧氏距离测度的 BWP 指标进行分析; 对于基因表达数据集, 则采用基于相似测度的 BWP 指标进行分析, 并采用 Pearson 相关系数作为相似性测度, 两个样本 x_i 和 x_k 之间的不相似度表示为 $H(x_i, x_k) = 1 - R(x_i, x_k)$.

实验 1　UCI 真实数据集实验. 该实验包括 3 个 UCI 真实数据集, 分别是 BUPA, Pima-indians-diabetes (简称 Pid) 和 Breast-cancer-wisconsin (简称 Bcw), 来源于 UCI Machine Learning Repository(http://www.ics.uci.edu/∼mlearn/MLRepository.html). BUPA 数据集是肝病诊断数据集, 该数据集的结构特征为轻微重叠、松散的聚类结构; Pid 数据集是糖尿病诊断数据集, 其结构特征为轻微分离、松散的聚类结构; Bcw 数据集是乳癌诊断数据集, 其结构特征为完全分离、松散的聚类结构. 数据集信息以及 5 种有效性指标估计出的最佳聚类数实验结果如表 1 所示. 从表 1 可知, BWP 指标和 IGP 指标对 3 个真实数据集都能得到正确的最佳聚类数, KL 指标仅对 BUPA 数据集有效, 而 DB 指标和 HS 指标对每个真实数据集都无法得到正确的最佳聚类数.

表 1　5 种有效性指标估计出的 UCI 数据集最佳聚类数

数据集	样本数目	样本维数	正确类数	最佳聚类数				
				DB	KL	HS	IGP	BWP
BUPA	345	6	2	14	2	18	2	2
Pid	768	8	2	20	18	27	2	2
Bcw	699	9	2	18	7	26	2	2

对正确类数为 2 类的 BUPA 数据集, 运用 BWP 指标确定最佳聚类数的实验结果如图 2 所示. 从中可以看出, BWP 指标得到的最佳聚类数 2 是正确的.

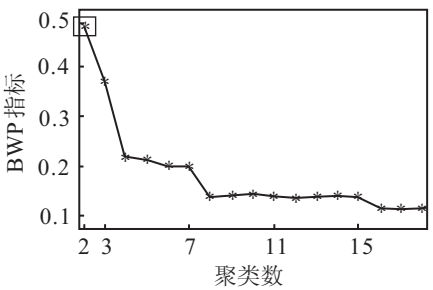

图 2　BUPA 的聚类数-BWP 指标关系图

实验 2　人工数据集实验. 该实验包括 3 个人工数据集, 分别是 Model 2, Y3c 和 Kes3. Model 2 数据集由中心分别为 $(0, 0), (0, 5), (5, -3)$ 的二维三高斯分布数据组成. 其中: $(5, -3)$ 类有 50 个样本, 其余类各有 25 个样本, 每个类的协方差矩阵为 I_2, I_2 为 2 阶单位矩阵. 该数据集的结构特征为轻微分离、松散的聚类结构. Y3c 数据集[10]是二维三类的人工合成数据集, 其结构特征为轻微重叠、松散的聚类结构. Kes3 数据集 (其分布结构参见图 3) 是二维三类的人工数据集, 其结构特征为完全分离、松散的聚类结构, 其中有一个类中样本较多, 并且部分类内样本之间的距离大于类间样本之间的距离.

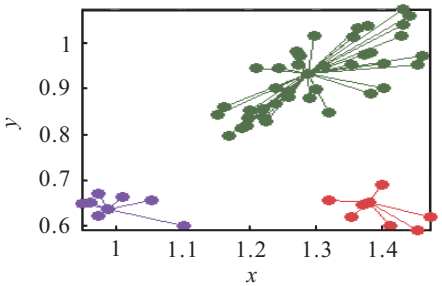

图 3　数据集 Kes3 的聚类结果

基于 AP 算法, 对 Model 2 数据集采用 5 种有效性指标得到的聚类结果如表 2 所示. 其中粗体数字所对应的聚类数为该列指标得到的最佳聚类数. 因为 Model 2 数据集的正确类数为 3 类, 所以 KL 指标, IGP 指标和 BWP 指标得到的最佳聚类数是正确的; DB 指标和 HS 指标得到的最佳聚类数是错误的.

表 2　Model 2 数据集的聚类有效性指标值

聚类数	DB	KL	HS	IGP	BWP
2	0.560 4	1.329 2	−54.898 1	0.980 8	0.509 0
3	1.842 4	**44.118 3**	−51.239 4	**1.000 0**	**0.535 0**
4	1.439 5	0.249 1	−42.728 9	0.978 8	0.343 9
5	1.613 6	1.331 4	−38.912 9	0.954 2	0.344 3
6	0.491 4	3.714 6	−36.794 5	0.914 9	0.199 6
7	0.454 8	0.779 7	−33.970 8	0.912 0	0.205 9
8	0.440 9	0.450 8	−31.756 2	0.919 6	0.214 1
9	**0.331 0**	2.8743	−30.357 6	0.870 2	0.223 3
10	0.335 7	2.874 3	**−28.241 4**	0.836 0	0.226 6

人工数据集信息以及 5 种有效性指标估计出的最佳聚类数实验结果如表 3 所示. 从表 3 可知, BWP 指标对 3 个人工数据集都能得到正确的最佳聚类数; IGP 指标和 KL 指标仅对 Model 2 数据集能得到正确的最佳聚类数; DB 指标仅对 Kes3 数据集有效; 而 HS 指标对每个人工数据集都无法得到正确的最佳聚类数.

基于 AP 算法, 设定 Kes3 数据集的类数为 3 类, 数据集 Kes3 的聚类结果如图 3 所示. 从中可以看出 AP 算法的聚类效果很好; 而在提供正确类数的前提

下, 采用 K-均值算法或FCM算法，对该数据集仍然无法正确聚类.

表 3 5种有效性指标估计出的人工数据集最佳聚类数

数据集	样本数目	样本维数	正确类数	最佳聚类数				
				DB	KL	HS	IGP	BWP
Model 2	100	2	3	9	3	10	3	3
Y3c	300	2	3	7	12	17	2	3
Kes3	57	2	3	3	4	7	2	3

实验 3 基因表达数据集实验. 该实验包括两个真实的基因表达数据集, 分别是Leukemia[11]和Yeast[12]. 其中: Leukemia是白血病的基因表达数据, 其结构特征是轻微分离、松散的聚类结构; Yeast是酵母的基因表达数据, 其结构特征是完全分离、松散的聚类结构.

基于AP算法, 对Leukemia数据集采用5种有效性指标得到的聚类结果如表4所示, 其中粗体数字所对应的聚类数为该列指标得到的最佳聚类数. 因为Leukemia数据集的正确类数为3类, 所以KL指标, HS指标, IGP指标和BWP指标得到了正确的最佳聚类数; DB指标得到的最佳聚类数是错误的.

表 4 Leukemia数据集的聚类有效性指标值

聚类数	DB	KL	HS	IGP	BWP
2	**0.3100**	3.9303	0.6126	0.9167	0.3152
3	0.3230	**12.0812**	**0.6939**	**1.0000**	**0.3421**
4	0.3557	0.9820	0.6556	0.8068	0.3034
5	0.3924	1.4197	0.6096	0.8479	0.2550
6	0.6093	0.9636	0.5463	0.7236	0.1281
7	0.4973	0.8322	0.5127	0.7804	0.1151
8	0.4938	0.8322	0.4855	0.7465	0.1345

对正确类数为4类的Yeast数据集, 运用BWP指标确定最佳聚类数的实验结果如图4所示. 从中可以看出, BWP指标得到的最佳聚类数4是正确的.

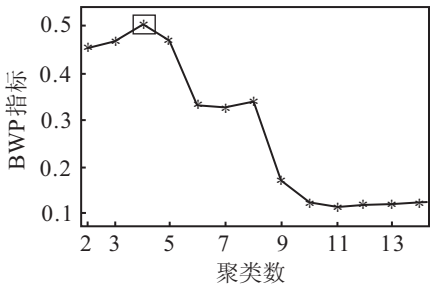

图 4 Yeast的聚类数-BWP指标关系图

真实的基因表达数据集信息以及5种有效性指标估计出的最佳聚类数实验结果如表5所示. 从表5可知, HS指标和BWP指标对两个真实的基因表达数据集都能得到正确的最佳聚类数; KL指标和IGP指标仅对Leukemia数据集能够得到正确的最佳聚类数; DB指标对每个真实的基因表达数据集都无法得到正确的最佳聚类数.

表 5 5种有效性指标估计出的基因表达数据集最佳聚类数

数据集	样本数目	样本维数	正确类数	最佳聚类数				
				DB	KL	HS	IGP	BWP
Leukemia	72	39	3	2	3	3	3	3
Yeast	208	79	4	7	8	4	3	4

6 结 论

AP算法能在较短的时间里得到很好的聚类结果, 且不需要初始化聚类中心, 这使得AP算法比较稳定, 适合进行聚类分析. 从样本几何结构的角度出发, 本文设计了一种新的聚类有效性指标——BWP指标, 并基于AP算法, 提出了一种新的确定样本最佳聚类数的方法. 理论研究和实验结果表明, 该方法能够有效地对聚类结果进行评估, 适用于确定样本的最佳聚类数.

参考文献(References)

[1] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315(5814): 972-976.

[2] Mézard Marc. Where are the exemplars?[J]. Science, 2007, 315(5814): 949-951.

[3] Davies D L, Bouldin D W. A cluster separation measure[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1979, 1(2): 224-227.

[4] Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset[J]. Genome Biology, 2002, 3(7): 1-21.

[5] Chen G, Jaradat S A, Banerjee N, et al. Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data[J]. Statistica Sinica, 2002, 12(1): 241-262.

[6] Kapp A V, Tibshirani R. Are clusters found in one dataset present in another dataset?[J]. Biostatistics, 2007, 8(1): 9-31.

[7] 肖宇, 于剑. 基于近邻传播算法的半监督聚类[J]. 软件学报, 2008, 19(11): 2803-2813.
(Xiao Y, Yu J. Semi-supervised clustering based on affinity propagation algorithm[J]. J of Software, 2008, 19(11): 2803-2813.)

[8] Rousseeuw P J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis[J]. J of Computational and Applied Mathematics, 1987, 20(1): 53-65.

(下转第1157页)