doi:10.3969/j.issn.1001-4616.2018.03.002

基于三支决策的谱聚类算法研究

施　虹[1],刘　强[1],王平心[1,2],杨习贝[1]

(1.江苏科技大学计算机学院,江苏 镇江 212003)
(2.河北师范大学数学与信息科学学院,河北 石家庄 050024)

[**摘要**]　硬聚类要求聚类的结果必须具有清晰的边界,即每个对象要么属于一个类,要么不属于一个类. 然而,将某些不确定的对象强制分配到某个类中往往容易带来较高的决策风险. 三支聚类将确定的元素放入核心域中,将不确定的元素放入边界域中延迟决策,可以有效地降低决策风险. 本文将三支决策理论与传统的谱聚类算法相结合给出了三支谱聚类的聚类算法. 该方法通过修改谱聚类算法的聚类过程并获得任一类簇的上界. 然后通过扰动分析从该类簇的上界分离出该类簇的核心域,同时上界与核心域的差值认为是该类簇的边界域. 在 UCI 数据集上的实验结果显示,该方法能有效提高聚类结果的 ACC、AS、ARI 值,并且降低 DBI 值.

[**关键词**]　谱聚类,三支决策,三支聚类,三支谱聚类
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2018)03-0006-08

Research on Spectral Clustering Algorithm Based on Three-way Decision

Shi Hong[1], Liu Qiang[1], Wang Pingxin[1,2], Yang Xibei[1]

(1.School of Computer, Jiangsu University of Science and Technology, Zhenjiang 212003, China)
(2.College of Mathematics and Information Science, Hebei Normal University, Shijiazhuang 050024, China)

Abstract: Hard clustering based on the assumption that a cluster must be represented by a set with crisp boundary. However, assigning uncertain points into a cluster will increase decision risk. Three-way clustering assigns the identified elements into the core region and the uncertain elements into the fringe region to reduce decision risk. In this paper, we present a new three-way spectral clustering by combining three-way decision and spectral clustering. In the proposed algorithm, we revise the process of spectral clustering and obtain an upper bound of each cluster. Perturbation analysis is applied to separate the core region from upper bound and the differences between upper bound and core region are regarded as the fringe region of specific cluster. The results on UCI data sets show that such strategy is effective in reducing the value of DBI and improving the values of ACC and AS.

Key words: spectral clustering, three-way decision, three-way clustering, three-way spectral clustering

聚类一直以来都是机器学习和模式识别等领域的一个极具挑战性的研究内容,所谓的聚类就是将数据集中的样本划分为不同的类,使得同类样本相似度高,异类样本相似度低. 经过不断的发展和完善,在图像处理[1]、网页搜索[2]、安全保障[3]和生物技术[4-5]等领域聚类分析已经得到了成功的应用,同时这一领域的研究人员也提出了很多相应的聚类算法[6-7]. 大致可以分为两大类:划分式聚类和层次聚类[8]. 其中划分式的聚类算法 k-means[9] 只能建立在凸球形样本空间的基础上,而面对非凸球形的样本空间就会陷入局部最优,但划分式聚类算法中的谱聚类算法[10-12]却能够很好地解决 k-means 算法的不足之处. 谱聚类算法是基于谱图划分理论的聚类算法,能对任意形状的数据进行划分且收敛于全局最优解,已经成功地运用到计算机视觉[13]、VLSI 设计[14]和机器学习[15]等领域.

k-means 算法和谱聚类算法都是硬聚类算法[6-7],即任意对象至多属于某一类簇,类与类之间有明确的界限. 而在面对某一个对象信息不确定或者不充分的情况下,强制将该对象划分到某类簇中,往往会带来较高的风险[16].

面对这样的难题,本文提出了一种新型的聚类算法即三支谱聚类算法. 三支谱聚类算法通过将三支

收稿日期:2018-04-16.
基金项目:国家自然科学基金(61503160、61572242).
通讯联系人:王平心,博士,副教授,研究方向:粗糙集、粒计算. E-mail:pingxin_wang@ hotmail.com

决策理论和谱聚类算法相结合,使得聚类结果由原来的上近似区域进一步细分为核心域和边界域. 核心域中的样本必定属于该类,边界域中的样本可能属于该类,核心域样本和边界域样本的并集组成上近似区域的样本. 该算法不仅可以进一步提高聚类结果的准确性,而且可以有效规避二支聚类所带来的风险. 对聚类结果的进一步划分使得各聚类性能指标:ACC 更大、ARI[17] 更大、AS[18] 更大、DBI[19-20] 更小. 通过仿真实验得出的结论可以验证该算法是行之有效的.

1　相关工作

1.1　谱聚类

谱聚类(Spectral Clustering)算法[10-12]近年来已经成为了机器学习领域的一个研究热点,它是一种基于图论的聚类方法,能在任意的样本空间上进行聚类并且获得收敛于全局最优解. 主要思想是把样本空间中的所有数据看成是空间中的点,这些点之间可以用边连接起来,距离较远(相似度较低)的两个点之间边的权重值较低,距离较近(相似度较高)的两个点之间边的权重值较高,通过对这些数据点组成的图进行切割,让切割后的不同子图之间的权重和尽可能的低,而子图内的权重和尽可能的高,这样就可以达到聚类的目的. 相比于 k-means 聚类算法[9],谱聚类算法[10-12]有很多的优势,其简单易于实现,只需通过标准的线性代数的方法便可以有效地求解.

本文主要研究谱聚类算法的经典算法之一即 NJW 算法[21],它是由 Ng 等人提出的多类实现算法.

该算法的关键部分就是通过高斯核函数 $\boldsymbol{A}_{ij}=\exp(-\|s_i-s_j\|^2/2\sigma^2)$($s_i$ 与 s_j 表示样本)构造相似矩阵 $\boldsymbol{A}$;然后根据公式 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$($\boldsymbol{D}$ 为度矩阵)得出矩阵 $\boldsymbol{L}$;计算矩阵 $\boldsymbol{L}$ 的前 k 个相互正交的特征向量组成矩阵 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k\}$;单位化矩阵 $\boldsymbol{X}$ 得到矩阵 $\boldsymbol{Y}$;最后将 $\boldsymbol{Y}$ 中的每一行看成是 R^k 空间中的一点,使用 k-means 聚类算法将其聚为 k 类. 主要实现步骤如算法 1 所示.

算法 1:NJW 算法

Step 1:给定数据集 $S=\{s_1,s_2,\cdots,s_n\}\in R^l$,聚类数目 k;

Step 2:令 $\boldsymbol{A}_{ij}=\exp(-\|s_i-s_j\|^2/2\sigma^2)$,$i\neq j$ 且 $\boldsymbol{A}_{ii}=0$,得到相似矩阵 $\boldsymbol{A}$;

Step 3:定义对角矩阵 $\boldsymbol{D}$,其中 d_{ii} 的值等于矩阵 $\boldsymbol{A}$ 第 i 行元素的和($i=1,2,\cdots,n$),构造矩阵 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$;

Step 4:求解矩阵 $\boldsymbol{L}$ 前 k 个相互正交的特征向量 $\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k$,得到矩阵 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_k\}\in R^{n\times k}$,其中 $\boldsymbol{x}_i$ 为列向量;

Step 5:单位化矩阵 $\boldsymbol{X}$ 得到矩阵 $\boldsymbol{Y}$,其中 $\boldsymbol{Y}_{ij}=\boldsymbol{X}_{ij}/\left(\sum_j \boldsymbol{X}_{ij}^2\right)^{1/2}$;

Step 6:将 $\boldsymbol{Y}$ 中的每一行看成是 R^k 空间中的一点,使用 k-means 聚类算法将其聚为 k 类;

Step 7:原始点 s_i 聚于类 j 中当且仅当矩阵 $\boldsymbol{Y}$ 的第 i 行被聚到类 j 中.

1.2　三支决策

三支决策理论[22-25]于 2010 年由 Regina 大学姚一豫教授通过扩展决策粗糙集的知识及整合其他学科领域的相关知识总结出来的,是自粗糙集理论发展以来的一种全新的决策思想. 三支决策是在研究对象信息不确定或者不充分的情况下所使用的一种决策方式,其核心思想是将决策项分成 3 种决策规则,分别是:正域决策、负域决策以及边界域决策,使其能够应用到实际生活领域中符合人类认知的一种决策模式. 相比于二支决策,三支决策具有很多优势,其中三支决策可以有效地降低因样本信息不足而强制做出决策所带来的风险[16]是其显著的优势之一.

文献[23]对三支决策的描述:设 U 是一个有限、非空实体集,其中 A 是有限条件集. 基于有限条件集,三支决策主要的任务是将 U 划分成 3 个两两互不相交的域,这 3 个域分别称之为 POS(正域)、NEG(负域)、BND(边界域). 依据这 3 个域给出与之对应的三支决策规则:接受、拒绝以及不承诺规则. 根据实际的应用可以给出不同的三支决策规则以及相应的解释(不局限于接受、拒绝以及不承诺). 图 1 给出了三支决策模型的简单描述.

1.3　三支聚类

三支决策思想自提出以来,广受各领域学者们的关注. 其中 Yu[26-29]将三支决策的思想引入到聚类中,提出了三支聚类的方法. 王[30-31]等人提出了关于三支聚类的 CE3 框架以及基于动态邻域的三支聚类方法.

文献[26-27]对三支聚类给出如下的描述:设数据集 $U=\{x_1,x_2,\cdots,x_n\}$,其中 n 表示数据集中含有 n 个对象,令 $C=\{c_1,c_2,\cdots,c_k\}$,其中 k 表示聚类数目即 n 个对象可以被分为 k 类. 基于三支聚类思想,其聚类结果以区间集的形式来表示即:

$$T=\{(Co(c_1),Fr(c_1)),(Co(c_2),Fr(c_2)),\cdots,(Co(c_k),Fr(c_k))\}.$$

其中 $Co(c_i)(i=1,2,\cdots,k)$表示第 i 类的核心域,核心域中的对象肯定属于该类;$Fr(c_i)(i=1,2,\cdots,k)$表示第 i 类的边界域,边界域中的对象可能属于该类也有可能不属于该类;核心域与边界域的并集组成该类的上界即 $Co(c_i)\cup Fr(c_i)=C_i^u$.

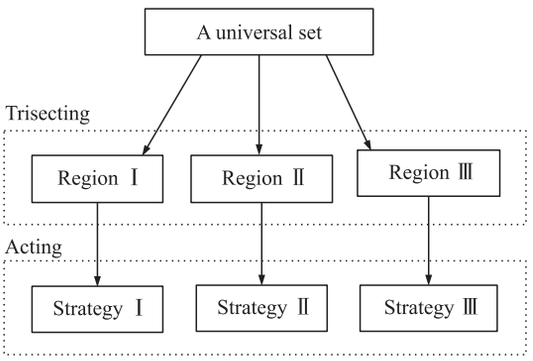

图 1　三支决策模型

Fig. 1　Three-way decisions model

用 C_i^l 和 C_i^u 分别表示类 i 下近似区域和上近似区域. C_i^l 表示类 i 下近似区域,该区域中的元素肯定属于类 i 即类 i 的核心域 $Co(c_i)(i=1,2,\cdots,k)$. $C_i^u-C_i^l$ 表示类 i 的边界域即 $Fr(c_i)(i=1,2,\cdots,k)$. 表示类 i 的琐碎域即 $Tr(c_i)(i=1,2,\cdots,k)$,该区域中的元素肯定不属于类 i. 根据聚类结果的定义,C_i^l和 C_i^u 须满足以下 3 个条件:

(1) $C_i^l\neq\varnothing(i=1,2,\cdots,k)$;

(2) $\bigcup_{i=1}^{k}C_i^u=U$;

(3) $C_i^l\cap C_j^l=\varnothing(i\neq j)$.

其中条件(1)表示任意一个类都是非空的,条件(2)表示其中一个对象 $x_i\in U$ 至少属于一个类,条件(3)表示任意一个类中的下近似区域中的对象都互不相交.

2　三支谱聚类

三支谱聚类算法的主要思想是:将三支决策理论[22-25]与谱聚类算法[10-12]相结合得到的一种聚类方法. 在三支谱聚类的聚类结果中任意类簇都是由核心域和边界域组成,核心域与边界域的并集组成该类簇的上界,核心域中的样本比较密集而边界域中的样本相对而言比较稀疏.

三支谱聚类算法主要分为两步:第一步通过谱聚类算法获取每一类簇的上界. 例如对于某一对象 x_i,任意选取 k 个聚类中心. 计算对象 x_i 到这 k 个聚类中心的最短距离即 $d(x_i,z_h)=\min\limits_{1\leqslant c\leqslant k}d(x_i,z_c)$,得到集合 $F=\{j:d(x_i,z_j)-d(x_i,z_h)\leqslant p\wedge j\neq h\}$(其中 p 是给定的参数)则有以下两种情形:

(1)如果 $F=\varnothing$则对象 $x_i\in C_h^u$;

(2)如果 $F\neq\varnothing$则对象 $x_i\in C_h^u$ 且 $x_i\in C_j^u$.

通过上述两种情形获取每一类簇的上近似区域的样本,接下来就需要重新计算每一类簇的聚类中心,计算公式如式(1)所示:

$$Z_i=\left(\sum_{x_k\in C_i^u}x_k\right)/|C_i^u|. \tag{1}$$

式中,$|C_i^u|$表示第 i 类上近似区域中的样本个数.

第二步将每一类簇的上界进一步细分为核心域和边界域(其中核心域与边界域的并集组成该类簇的上界). 对于每一类的上界 C_i^u,将进一步划分为核心域和边界域,操作如下:

(1)情形 1$=\{x_i\in C_i^u\mid\exists j=1,\cdots,k,j\neq i,x_i\in C_j^u\}$;

(2)情形 2$=\{x_i\in C_i^u\mid\forall j=1,\cdots,k,j\neq i,x_i\notin C_j^u\}$.

若对象 x 属于情形 1,即该对象至少属于某一类簇,则将该对象划分到类 i 的上界;若对象 x 属于情形 2,即该对象至多属于某一类 i 簇,则在类的上界中增加 m_i(m_i 表示类 i 上界的样本个数)个相同的样本 x,得到类 i 新的上界记为 $C_i^{U^*}$,使用公式(1)计算 $C_i^{U^*}$ 的聚类中心 z_i^*,并且计算新旧聚类中心的距离 $|z_i-z_i^*|$,若 $|z_i-z_i^*|\leqslant q$(q 为给定参数)则将 x 聚到类 i 的核心域,否则将 x 聚到 i 类的边界域.

通过三支谱聚类方法可以将数据集中的样本划分为类内相似度高,类间相异度高的聚类结果,并且利用 ACC、ARI[20]、DBI[19-20]、AS[18]等性能评价指标对该算法的聚类结果进行评价,以此来验证该聚类算法是行之有效的. 主要实现步骤如算法 2 所示.

算法 2:三支谱聚类算法

输入数据集 $U=\{x_1,x_2,\cdots,x_n\}$,聚类数目 k,参数 p、q、σ;
输出聚类结果 $T=\{(Co(c_1),Fr(c_1)),\cdots,(Co(c_k),Fr(c_k))\}$;
令 $\boldsymbol{A}_{ij}=\exp(-\|x_i-x_j\|^2/2\sigma^2)(i\neq j)$ 且 $\boldsymbol{A}_{ii}=0$,得到相似矩阵 $\boldsymbol{A}=\sum_{i=1}^{n}\sum_{j=1}^{n}a_{ij}$;
令 $d_{ii}=\sum_{j=1}^{n}a_{ij}$,$d_{ij}=0(i\neq j)$,得到矩阵 $\boldsymbol{D}=\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}$;
令 $\boldsymbol{L}=\boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$,得到矩阵 $\boldsymbol{L}$;
计算矩阵 $\boldsymbol{L}$ 的前 k 个相互正交的特征向量 $\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_k$,得到矩阵 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_k\}$;
令 $\boldsymbol{Z}_{ij}=\boldsymbol{Y}_{ij}/\left(\sum_j\boldsymbol{Y}_{ij}^2\right)^{1/2}$,得到矩阵 $\boldsymbol{Z}=\{z_1,z_2,\cdots,z_k\}$;
随机选择 k 个聚类中心 $z_1,z_2,\cdots,z_k$;
for $i=1,2,\cdots,n$ do
　repeat
　　计算距离对象 x 最近的聚类中心 z_h:$d(x_i,z_h)=\min\limits_{1\leqslant c\leqslant k}d(x_i,z_c)$;得到集合 $F=\{j:d(x_i,z_j)-d(x_i,z_h)\leqslant p\wedge j\neq h\}$;
　　If $F=\varnothing$ then
　　将样本聚到类 h 的上界即 $x_i\in C_h^u$;
　　else
　　将样本 x_i 同时归到类 h 和类 j 的上界即 $x_i\in C_h^u$ 且 $x_i\in C_j^u$;
　　end if
　　使用公式(1)重新计算聚类心;
　until 聚类中心值不再变化
end for
for $i=1,2,\cdots,k$ do
　　样本 $x\in C_i^u$,定义集合 $G=\{j:j\neq i\wedge x\in C_j^u\}$;
　if $G\neq\varnothing$ then
　　将样本 x 归到 C_i 的边界域即 $x\in Fr(c_i)$;
　　else
　　在类 i 的上界中增加 m_i 个相同的样本 x 得到类 i 新的上界记为 $C_i^{u^*}$,其中 m_i 表示类 i 上界的样本个数,使用公式(1)计算 $C_i^{u^*}$ 的聚类中心 z_i^*,并且计算新旧聚类中心的距离 $|z_i-z_i^*|$;
　　If $|z_i-z_i^*|\leqslant q$ then
　　将 x 聚到类 i 的核心域即 $x\in Co(c_i)$;
　　else
　　将 x 聚到类 i 的边界域即 $x\in Fr(c_i)$;
　　end if
　end if
end for
return $T=\{(Co(c_1),Fr(c_1)),\cdots,(Co(c_k),Fr(c_k))\}$

3　聚类结果评价指标

3.1　准确率

准确率(Accuracy)是一种常见的评价聚类结果好坏的外部指标. 其基本思想是准确率越高,聚类的效果越好.

定义 1(ACC)

$$ACC=\frac{1}{N}\sum_{i=1}^{k}\theta_i,$$

式中,N 表示已确定划分到所属类别的样本个数,θ_i 表示正确划分到类 i 的样本个数,k 表示聚类数.

3.2 Adjusted Rand Index 评价指标

ARI[17]是经 RI 推广而来的,在 RI 不能保证类别标签是随机分布的情况下,其值接近于 0,并且 RI 的惩罚力度不够,区分度不高. 为了实现在聚类结果随机产生的情况下,指标应接近于零的目的,由此提出了调整兰德系数(Adjusted rand index),它具有更高的区分度.

对数据集 $S=\{P_1,P_2,\cdots,P_n\}$,假定通过聚类给出的簇划分为 $X=\{X_1,X_2,\cdots,X_r\}$,参考模型给出的簇划分为 $Y=\{Y_1,Y_2,\cdots,Y_s\}$.

定义 2(RI)

$$RI=\frac{a+b}{a+b+c+d},$$

$$a=|S^*|,S^*=\{(P_i,P_j)\mid P_i,P_j\in X_k,P_i,P_j\in Y_l\};$$

$$b=|S^*|,S^*=\{(P_i,P_j)\mid P_i\in X_{k_1},P_j\in X_{k_2},P_i\in Y_{l_1},P_j\in Y_{l_2}\};$$

$$c=|S^*|,S^*=\{(P_i,P_j)\mid P_i,P_j\in X_k,P_i\in Y_{l_1},P_j\in Y_{l_2}\};$$

$$d=|S^*|,S^*=\{(P_i,P_j)\mid P_i\in X_{k_1},P_j\in X_{k_2},P_i,P_j\in Y_l\};$$

$$(1\leqslant i,j\leqslant n,i\neq j,1\leqslant k,k_1,k_2\leqslant r,k_1\neq k_2,1\leqslant l,l_1,l_2\leqslant s,l_1\neq l_2)$$

a 表示属于同一类的样本最后被分到同一类中的样本总数;b 表示不属于同一类的样本并且最后没有分到同一类中的样本总数;c 表示属于同一类但最后没有被分到同一类的样本总数;d 表示不属于同一类但最后被分到同一类的样本总数. $a+b$ 可以表示划分结果相一致的样本总数;$c+d$ 可以表示划分结果不一致的样本总数.

定义 3(ARI)

$$ARI=\frac{\overbrace{\sum_{ij}\binom{n_{ij}}{2}}^{\text{Index}}-\overbrace{[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}]/\binom{n}{2}}^{\text{Expected\_Index}}}{\underbrace{\frac{1}{2}[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}]}_{\text{Max\_Index}}-\underbrace{[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}]/\binom{n}{2}}_{\text{Expected\_Index}}},$$

式中,变量 a_i,b_j 的值如表 1 所示,$n_{ij}=|X_i\cap Y_j|$. ARI 取值范围为[-1,1],ARI 值越大表示聚类结果越符合真实情况. 从广义角度来讲,ARI 衡量的是两个样本分布的契合程度.

表 1　情形分析表

Table 1　The contingency table

X	Y				Sums
	Y_1	Y_2	$\cdots$	Y_s	
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
Sums	b_1	b_2	$\cdots$	b_s	

3.3 Davies-Bouldin Index 评价指标

Davies-Bouldin Index,即 DB_Index[19-20]. 由 Davide L Davies 和 Donald W Bouldin 于 1979 年提出来的,是一种评估度量聚类结果的方法.

定义 4(DBI)

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{i\neq j}\left(\frac{\bar{c}_i+\bar{c}_j}{\|w_i-w_j\|_2}\right),$$

式中,$\bar{c}_i$ 表示第 i 类中的所有样本元素到聚类中心 w_i 的平均距离,$\|w_i-w_j\|_2$ 表示类 i 与类 j 之间的欧式距离,k 表示聚类数.

DBI 的基本思想评价一个聚类结果的好坏的依据是:要求类内元素相似度高,类间元素相似度低. DBI 计算任意两类的类内平均距离之和除以两聚类中心距离求出最大值,DBI 越小,意味着类内距离越小,同时类间距离越大.

3.4 平均轮廓系数

轮廓系数(Silhouette Coefficient)[18],是聚类效果好坏的一种评价方式,最早由 Peter J Rousseeuw 在 1986 年提出. 它结合内聚度和分离度两种因素,可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响.

定义 5(单个样本 d_i 的轮廓系数 S_i)

$$S_i=\frac{b_i-a_i}{\max(a_i,b_i)},$$

式中,a_i 表示样本 d_i 与同类中其他所有样本的平均距离,a_i 称为类内相异度,a_i 越小说明该样本属于该类的可能性越大. 式中 $b_i=\min\{D(d_i-c_j)\}$,$D(d_i-c_j)$ 表示样本 d_i 到类 c_j 所有样本的平均距离,b_i 称为类间相异度,b_i 越大说明该样本属于其他类的可能性越小.

定义 6(平均轮廓系数 AS)

$$AS=\frac{1}{N}\sum_{i=1}^{N}S_i,$$

式中,i 表示样本总数,S_i 表示第 i 个样本的轮廓系数. 平均轮廓系数是用所有样本的轮廓系数的均值来表示,是该聚类结果是否有效、合理的度量.

4　实验结果

本文选用五组标准的 UCI[32] 数据集,如表 2 所示. 首先通过 k-means 算法[9] 和谱聚类算法[10-12] 分别对这 5 组标准的 UCI 数据集进行二支聚类,经实验测试得出参数 $p=0.005$,$q=0.26$;然后使用本文所提出的三支谱聚类算法分别对这 5 组标准的 UCI 数据集进行三支聚类;最后通过聚类结果性能评价指标 DBI[19-20]、ARI[17]、ACC、AS[18] 分别对二支聚类结果和三支聚类结果进行评价.

表 2　实验中使用的数据集

Table 2　Datasets used in experiments

Datasets	Sample numbers	Sample dimensions	Categories
Iris	150	4	3
Wdbc	569	30	2
Wine	178	13	3
Hill	1212	100	2
Bank	1372	4	2

表 3　UCI 数据集上的实验结果

Table 3　Experimental results on UCI datasets

Datasets	Algorithm	Average value				Best value			
		DBI	AS	ARI	ACC	DBI	AS	ARI	ACC
Iris	k-means	0.780 0	0.683 5	0.680 1	0.850 2	0.761 0	0.695 9	0.716 3	**0.886 7**
	Spectral clustering	0.769 1	0.689 7	0.684 7	0.865 9	0.722 8	0.690 0	0.689 8	0.873 3
	Three-way spectral clustering	**0.717 7**	**0.730 8**	**0.719 5**	**0.883 0**	**0.694 4**	**0.731 1**	**0.722 2**	0.886 5
Wdbc	k-means	1.136 3	0.576 5	0.730 2	0.927 9	1.136 3	0.576 5	0.730 2	0.927 9
	Spectral clustering	1.175 4	0.540 6	0.725 1	0.926 2	1.175 4	0.540 6	0.725 1	0.926 2
	Three-way spectral clustering	**1.088 1**	**0.585 1**	**0.799 5**	**0.947 4**	**1.088 1**	**0.585 1**	**0.799 5**	**0.947 4**
Wine	k-means	1.312 8	0.474 8	0.838 0	0.943 5	1.072 9	0.476 4	0.899 2	0.966 3
	Spectral clustering	1.306 9	0.474 8	0.899 2	0.966 3	1.306 9	0.474 8	0.899 2	0.966 3
	Three-way spectral clustering	**1.191 1**	**0.537 0**	**0.942 2**	**0.981 3**	**1.191 1**	**0.537 0**	**0.942 2**	**0.981 2**
Hill	k-means	**0.404 8**	**0.937 6**	0.000 0	0.509 1	**0.404 8**	0.937 6	0.000 0	0.509 1
	Spectral clustering	0.499 2	0.892 6	0.000 3	0.513 2	0.499 2	0.892 6	0.000 3	0.513 2
	Three-way spectral clustering	0.461 2	0.912 4	**0.000 4**	**0.514 4**	0.460 5	0.912 8	**0.000 4**	**0.514 7**
Bank	k-means	1.191 3	0.500 2	0.021 6	0.574 7	1.191 1	0.500 4	0.022 3	0.575 8
	Spectral clustering	1.200 2	0.490 8	0.051 7	0.614 4	1.200 2	0.490 8	0.051 7	0.614 4
	Three-way spectral clustering	**1.125 6**	**0.530 5**	**0.058 0**	**0.621 1**	**1.125 6**	**0.530 5**	**0.058 0**	**0.621 1**

通过对每组数据集进行 100 次实验,得到这 100 组实验数据的平均值和最好值,其中平均值用来比较该算法的总体性能,最好值用来比较该算法的最好性能,实验结果如表 3 所示.

通过比较表 3 的实验结果发现:三支谱聚类算法分别在数据集 Iris、Wdbc、wine、Bank 上有较好的聚类效果,各性能指标 DBI、AS、ARI、ACC 的平均值与最好值都显著优于 k-means 和谱聚类算法. 其中,由于 k-means 算法的不稳定性出现了个别性能指标的最好值要优于三支谱聚类算法的性能指标的最好值. 但是,我们发现三支谱聚类算法在 Hill 数据集上的聚类性能并没有达到预期的效果,k-means 和谱聚类算法在数据集 Hill 上的性能指标 DBI 和 AS 的平均值和最好值要优于三支谱聚类,尽管 ARI 和 ACC 两个指标的平均值与最好值略高于 k-means 和谱聚类算法. 虽然三支谱聚类在数据集 Hill 上的聚类结果不如

k-means 算法和谱聚类算法,但是在大部分的数据集上,三支谱聚类算法的聚类结果相比于二支聚类 k-means 算法和谱聚类算法的聚类结果,不管在总体性能上还是在最好性能上都得到了有效的提升. 综上所述,可以证实三支谱聚类算法对于大部分的数据集是可以有效降低聚类结果的 DBI 值,提高聚类结果的 AS、ARI、ACC 值.

5 结语

目前,信息化时代的快速发展伴随着大量不完备不确定信息的涌现,硬聚类已然无法解决这些实际问题. 考虑到能进一步提高聚类性能,本文提出了三支谱聚类算法,通过结合三支决策理论和谱聚类算法. 面对那些信息不确定以及不充分的对象,采取延迟决策的规则可以有效降低决策风险. 通过三支谱聚类算法将类中样本进一步划分为核心域和边界域,以此来提高聚类结果的准确性. 实验结果表明,相比较于硬聚类算法的聚类结果,本文提出的算法较大程度地提高了其性能. 但是该聚类算法有待进一步地修改与完善,可以进一步提高聚类结果的性能,这也将是我们接下来研究的主要内容.

[参考文献]

[1] ELALAMI M E. Supporting image retrieval framework with rule base system[J]. Knowledge-based systems,2011,24(2):331-340.

[2] MARTiN G J D,PALOMARES A,BALAGUER B E,et al. Studying the feasibility of a recommender in a citizen web portal based on user modeling and clustering algorithms[J]. Expert systems with applications,2006,30(2):299-312.

[3] KALYANI S,SWARUP K S. Particle swarm optimization based k-means clustering approach for security assessment in power systems[J]. Expert systems with applications,2011,38(9):10839-10846.

[4] SHI J,LUO Z. Nonlinear dimensionality reduction of gene expression data for visualization and clustering analysis of cancer tissue samples[J]. Computers in biology & medicine,2010,40(8):723.

[5] SEBISKVERADZE D,VRABIE V,GOBINET C,et al. Automation of an algorithm based on fuzzy clustering for analyzing tumoral heterogeneity in human skin carcinoma tissue sections[J]. Technical methods and pathology,2011,91(5):799-811.

[6] XU R. Survey of clustering algorithms[J]. IEEE transactions on neural networks,2005,16(3):645-678.

[7] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008,19(1):48-61.

[8] KAUFMAN L,ROUSSEEUW P J. Finding groups in data:an introduction to cluster analysis[M]. New Jersey:John Wiley & Sons Inc,1990.

[9] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]//Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability. California:University of California Press,1967.

[10] LUXBURG U V. A tutorial on spectral clustering[J]. Statistics & computing,2007,17(4):395-416.

[11] 蔡晓妍,戴冠中,杨黎斌. 谱聚类算法综述[J]. 计算机科学,2008,35(7):14-18.

[12] CAI Y,JIAO Y Y,ZHUGE W Z,et al. Partial multi-view spectral clustering[J]. Neurocomputing,2018,311:316-324.

[13] MALIK J,BELONGIE S,LEUNG T,et al. Contour and texture analysis for image segmentation[J]. International journal of computer vision,2001,43(1):7-27.

[14] WEISS Y. Segmentation using eigenvectors:a unifying view[C]//International Conference on Computer Vision. United States:IEEE Computer Society,1999.

[15] SHI J,MALIK J. Normalized cuts and image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence,2000.

[16] 李丽红,李言,刘保相. 三支决策中不承诺决策的转化代价与风险控制[J]. 计算机科学,2016,43(1):77-80.

[17] FAHAD A,ALSHATRI N,TARI Z,et al. A survey of clustering algorithms for big data:taxonomy and empirical analysis[J]. IEEE transactions on emerging topics in computing,2014,2(3):267-279.

[18] ROUSSEEUW P J. Silhouettes:a graphical aid to the interpretation and validation of cluster analysis[J]. Journal of computational & applied mathematics,1987,20(20):53-65.

[19] BEZDEK J C,PAL N R. Some new indexes of cluster validity[J]. IEEE transactions on systems man & cybernetics society,1998,28(3):301-315.

[20] MAULIK U,BANDYOPADHYAY S. Performance evaluation of some clustering algorithms and validity indices[J]. IEEE

transactions on pattern analysis & machine intelligence,2002,24(12):1650-1654.
[21] NG A Y,JORDAN M I,WEISS Y. On spectral clustering:analysis and an algorithm[J]. Proc Nips,2001,14:849-856.
[22] YAO Y Y. The superiority of three-way decisions in probabilistic rough set models[J]. Information sciences,2011,181(6):1080-1096.
[23] YAO Y Y. An outline of a theory of three-way decisions[C]//International Conference on Rough Sets and Current Trends in Computing. Berlin,Heidelberg:Springer,2012.
[24] GAO C,YAO Y Y. Actionable strategies in three-way decisions[J]. Knowledge-based systems,2017,133:183-199.
[25] 李金海,邓硕. 概念格与三支决策及其研究展望[J]. 西北大学学报(自然科学版),2017,47(03):321-329.
[26] YU H,ZHANG C,WANG G Y. A tree-based incremental overlapping clustering method using the three-way decision theory[J]. Knowledge-based systems,2016,91(C):189-203.
[27] YU H,JIAO P,YAO Y Y,et al. Detecting and refining overlapping regions in complex networks with three-way decisions[J]. Information sciences,2016,373:21-41.
[28] YU H. A framework of three-way cluster analysis[C]//International Joint Conference on Rough Sets. Cham:Springer,2017.
[29] YU H,WANG X,WANG G,et al. An active three-way clustering method via low-rank matrices for multi-view data[J/OL]. Information sciences,2018. https://doi.org/10.1016/j.ins.2018.03.009.
[30] WANG P X,YAO Y Y. CE3:A three-way clustering method based on mathematical morphology[J]. Knowledge-based systems,2018,155:54-65.
[31] 王平心,刘强,杨习贝,等. 基于动态邻域的三支聚类分析[J]. 计算机科学,2018,45(1):62-66.
[32] BLAKE C L,MERZ C J. UCI machine learning repository[J/OL]. Html,2005. http://www.ics.uci.edu/mlearn/MLRepository.

[责任编辑:黄　敏]