文章编号:1671-9352(2024)05-0045-07 DOI:10.6040/j.issn.1671-9352.4.2023.137

基于自然最近邻的样本扰动三支聚类

朱金[1],付玉[2*],管文瑞[3],王平心[4]
(1.江苏科技大学经济管理学院, 江苏 镇江 212100; 2.南京中医药大学镇江附属医院(镇江中医院), 江苏 镇江 212000; 3.江苏科技大学自动化学院, 江苏 镇江 212100; 4.江苏科技大学理学院, 江苏 镇江 212100)

摘要:利用数据样本的自然最近邻信息,给出了一种基于样本扰动理论的三支聚类算法,结合自然最近邻信息生成 2 组扰动数据集,随机提取特征子集并使用 K-means 聚类算法获得不同的聚类结果,利用共现概率矩阵和确定函数获得样本的稳定性,根据样本稳定性阈值将样本划分为稳定区域和不稳定区域,再对 2 个区域的样本使用不同的策略获得每个类簇的核心域和边界域。实验采用 5 个公开数据集与 2 种传统的聚类算法进行对比, 结果验证了所提算法的有效性。

关键词:三支决策;三支聚类;样本扰动;自然最近邻
中图分类号:TP181 **文献标志码**:A
引用格式:朱金,付玉,管文瑞,等.基于自然最近邻的样本扰动三支聚类[J]. 山东大学学报(理学版),2024,59(5):45-51,62.

Perturbation three-way clustering based on natural nearest neighbors

ZHU Jin[1], FU Yu[2*], GUAN Wenrui[3], WANG Pingxin[4]
(1. School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang 212100, Jiangsu, China; 2. Zhenjiang Hospital Affiliated to Nanjing University of Chinese Medicine (Zhenjiang Hospital of Traditional Chinese Medicine), Zhenjiang 212000, Jiangsu, China; 3. School of Automation, Jiangsu University of Science and Technology, Zhenjiang 212100, Jiangsu, China; 4. School of Science, Jiangsu University of Science and Technology, Zhenjiang 212100, Jiangsu, China)

Abstract: By using sample's natural nearest neighbors, a three-way clustering algorithm is proposed based on sample's perturbation theory. The proposed algorithm combines natural nearest neighbor information with sample's perturbation to generate two datasets. By randomly selecting parts of the sample's feature, different clustering results are obtained through K-means clustering algorithms. The stability of each sample is calculated based on the defined frequencies. The universe is divided into stable set and unstable set based on the sample's stability. Then, we use different strategies to obtain the core region and fringe region of each cluster. The testing results on five open datasets verify the effectiveness of the proposed algorithm through comparative tests with two traditional clustering methods.

Key words: three-way decision; three-way clustering; sample's perturbation; natural nearest neighbor

0 引言

如何从海量的数据中挖掘有用的信息一直是信息科学和人工智能领域中一项具有挑战性的任务。粒计算作为一种新兴的智能信息处理技术,可以在多个粒度或抽象层次上处理信息和解决问题[1-2]。粒计算的主要任务是通过各种信息粒化方法构建不同的粒结构,聚类是信息粒化中最广泛使用的方法之一[3]。聚类的目的是将一个数据对象的集合划分成为若干个通常是不相交的簇,使得簇内对象之间有较高的相似性,不

收稿日期:2023-06-02; **网络出版时间**:2024-04-09 21:47:33
网络出版地址:https://link.cnki.net/urlid/37.1389.N.20240401.1729.002
基金项目:国家自然科学基金资助项目(62076111, 61773012); 江苏省高校自然科学基金资助项目(15KJB110004)
第一作者:朱金(1979—),男,副教授,博士,研究领域为三支决策、粒计算. E-mail:zhujjust@ 163.com
*通信作者:付玉(1978—),女,副主任中医师,研究领域为医学三支决策、粒计算. E-mail:fumima0511@ hotmail.com

同簇中的对象有较高的差异性[4]。目前，聚类已广泛应用于生物信息学[5]、图像处理[6]、属性约简[7]等各个领域。

现有的聚类算法大多都是硬聚类，也就是元素只能属于一个簇或者不属于一个簇，并且簇之间有明确的边界。当信息不准确或数据不充分时，如果将数据对象强行划分到某一类簇，会增加误分类的概率，带来较高的决策风险。为了克服不确定性信息带来的决策风险，Yao[8]在研究决策粗糙集的基础上提出了三支决策概念，3个域分别对应接受、拒绝以及不承诺3种决策规则。相比于二支决策，三支决策更符合人类认知的模式，同时还能有效规避由样本信息不充分带来的决策风险。三支决策是将一个有限、非空实体集在一个有限条件集的基础下划分成3个两两互不相交的域[9]，这3个域分别称为正域、负域、边界域，给出这3个域对应的决策规则为接受、拒绝以及不承诺规则。Yu等[10]将三支决策和聚类结合，提出了三支聚类框架。三支聚类用核心域和边界域来描述一个簇，这2个域将数据集分成了3个部分，且定义了属于、不属于和可能属于3种关系。Wang等[11]结合数学形态学思想提出了基于收缩和扩张策略的三支聚类理论；Yu等[12]提出了具有噪声并基于密度的三支聚类算法，利用优化相似度聚类；凡嘉琛等[13]将三支聚类引入到密度敏感谱聚类；姜春茂等[14]利用阴影集构造三支聚类的核心域和边界域，提出了一种基于阴影集的三支集成聚类算法。

本文利用数据样本的自然最近邻信息，提出了一种基于样本扰动理论的三支聚类算法，结合自然最近邻信息生成2组扰动数据集，随机提取特征子集并使用K-means聚类算法获得不同的聚类结果，通过使用协关联矩阵和确定函数获得样本的稳定性，根据样本稳定性阈值将样本划分为稳定区域和不稳定区域，对2个区域的样本使用不同的策略，获得每个类簇的核心域和边界域。

1 相关工作

1.1 三支决策聚类

设数据集包含 n 个样本对象，传统的聚类方法是用一个集合表示一个类，即寻找一组集合 $C^1, C^2, \cdots, C^k$，满足

$$\begin{cases} C^i \neq \varnothing, & i=1,2,\cdots,k, \\ \bigcup\limits_{i=1}^{k} C^i = U, & \\ C^i \cap C^j = \varnothing, & i \neq j, \end{cases}$$

式中 k 为类簇的个数。上述条件要求一个对象要么属于某个类簇，要么不属于某个类簇，聚类的结果具有清晰的边界，然而，某些不确定的对象强制分配到某个类中会降低聚类结果的结构和精度。三支聚类用 C_{core}、C_{fringe} 和 C_{trivial} 表示一个类，即核心域、边界域和琐碎域，C_{core} 中的样本点一定属于类 C，C_{fringe} 中的样本可能属于类 C，C_{trivial} 中的对象一定不属于类 C，由于 $C_{\text{trivial}} = U - C_{\text{core}} - C_{\text{fringe}}$，因此三支聚类为

$$\mathscr{C} = \{ (C^1_{\text{core}}, C^1_{\text{fringe}}), (C^2_{\text{core}}, C^2_{\text{fringe}}), \cdots, (C^k_{\text{core}}, C^k_{\text{fringe}}) \}。$$

三支聚类用核心域与边界域来表示一个类，这种聚类方法充分考虑了那些因信息不充分而无法确定类簇归属的样本对象，降低了决策风险，广泛应用于不确定性分析等领域[15]。

1.2 自然最近邻

定义 1（自然最近邻） 对于数据点 x，若有数据对象 y 认为 x 是其邻居，且当数据集中所有的数据样本在其他样本的邻域中出现的次数至少为1次时，称数据对象 y 为数据对象 x 的自然最近邻居。

定义 2（自然特征值） 数据集的自然特征值为使任意的数据点 x 都至少被另一个数据点 y 的 r-邻域包含的最小 r 值，自然特征值定义为

$$K_{\text{sup}} = \min \{ r \mid \forall x, \exists y, y \neq x, \text{ s.t. } x \in N_r(y) \}, \tag{1}$$

式中 $N_r(y)$ 表示点 y 的第 r 最近邻域。

自然最近邻[16]是一种新的最近邻概念，是一种无尺度的最近邻，这也是自然最近邻与K-近邻和 ε-邻域最大的不同之处。自然最近邻的基本思想是数据集中密集较高区域中的元素拥有较多的邻居，密集较低区域中的元素拥有较少的邻居，而数据集中最离群的数据点只有一个最近邻居。自然最近邻居的计算过程不

需要任何的参数,在寻找邻居的过程中不断地适应数据集的分布结构。具体算法步骤如算法 1 所示。

算法 1　自然最近邻算法。

输入　数据集 $U=\{v_1,v_2,\cdots,v_n\}$ (维数为 $N*p$);

输出　自然邻居特征值 $K_{\sup}$;数据点的自然最近邻 $\mathcal{N}(v_i)$;每个点自然最近邻的数量 $nb(v_i)$。

初始化参数:$r=0$, $N=\varnothing$, $\forall v_i\in U$, $nb(v_i)=0$, flag = 0

将数据集 X 放入 k-d 树中;

While flag = 0 do

If $nb(v_i)\neq 0$ 任意的 $v_i\in U$

flag = 1

Else

$r=r+1$

For 任意的 $v_i\in U$

利用 k-d 数寻找 v_i 第 r 个邻居:$N_r(v_i)$

$nb(N_r(v_i))=nb(N_r(v_i))+1$

$\mathcal{N}(N_r(v_i))=\mathcal{N}(N_r(v_i))\cup\{i\}$

End For

End If

End while

$K_{\sup}=r$

返回　自然邻居特征值 $K_{\sup}$;每个点的自然最近邻 $\mathcal{N}(v_i)$;每一个点自然最近邻个数 $nb(v_i)$。

1.3　样本稳定性

文献[17]利用样本的共现概率和确定性函数,给出了样本稳定性的定义,样本的共现概率是基于一组聚类结果。共现概率为 1 时,样本始终被分在一类,具有较高的稳定性;共现概率为 0 时, 样本始终没有被分在一类,2 个样本之间具有较高的稳定性;而共现概率在[0,1]时,说明样本在某些聚类结果是聚在一类,某些聚类结果没有聚在一类。文献[18]给出了确定函数的概念来评价样本关系的确定度,并用一个样本与其他样本的平均确定度来定义样本的稳定性。

定义 3[17] (确定性函数)　设 t 为[0,1]上一个常数,样本的确定性函数为关于共现概率变量 p 的函数 $f(p)$,式中 $p\in[0,1]$,$f(p)$ 满足:

(1) 如果 $p<t$, $f'(p)<0$;如果 $p>t$, $f'(p)>0$;

(2) 如果 $p_i<t<p_j$ 且 $\dfrac{t-p_i}{p_j-t}=\dfrac{t}{1-t}$,有 $f(p_i)=f(p_j)$。

当 $p<t$ 时,函数 f 的导数小于 0,函数单调递减;当 $p>t$ 时,函数的导数大于 0,函数 f 单调递增,因此,确定性函数在 t 处取得最小值且 p 距 t 越远,函数值越大。确定性函数在 t 的两侧成比例对称,即当 $p_i<t<p_j$ 且 $\dfrac{t-p_i}{p_j-t}=\dfrac{t}{1-t}$ 时,函数 $f(p_i)=f(p_j)$。满足条件(1)、(2)的确定性函数有很多,本文采用线性的方法定义确定性函数,

$$f_l(p)=\begin{cases}|(p-t)/t|, & p<t,\\ |(p-t)/(1-t)|, & p\geqslant t。\end{cases}\tag{2}$$

定义 4[17] (样本稳定性)　假定一个数据集 $U=\{v_1,v_2,\ldots,v_n\}$ 含有 n 个样本,p_{ij} 表示样本 v_i 与 v_j 的共现概率,基于确定性函数 f,对于每一个点 v_i,样本稳定性 $s(v_i)$ 为

$$s(v_i)=\frac{1}{n}\sum_{j=1}^{n}f(p_{ij})。\tag{3}$$

特别地,针对线性函数 $fl(p)$,样本点 v_i 的稳定性记为

$$s_l(v_i)=\frac{1}{n}\sum_{j=1}^{n}f_l(p_{ij})。\tag{4}$$

2 基于样本扰动的三支聚类方法

本文给出一种基于样本扰动理论的三支聚类算法，该算法结合自然最近邻算法生成2组扰动数据集，通过随机提取样本的部分特征子集，利用传统的K-means聚类，算法得到不同的聚类结果。利用协关联矩阵和确定性函数得到样本的稳定性，根据样本稳定性阈值将样本划分为稳定区域和不稳定区域。对稳定区域中的元素，使用K-means算法将其划分为每个簇的核心区域，对不稳定区域中的样本，将其分配给每个簇的边缘区域，从而获得三支聚类的结果。

2.1 样本扰动

本文利用对数据集和扰动数据集在不同的特征子集上进行聚类，生成不同的聚类结果，利用自然最近邻生成扰动数据集。给定一个数据集 $U=\{v_1,v_2,\cdots,v_n\}$，每个元素的自然最近邻居可以通过算法1得到。对于元素 v，通过以下公式得到2个新元素：

$$v^1=v+\alpha d_v, \tag{5}$$

$$v^2=v-\alpha d_v, \tag{6}$$

式中：α 为给定参数，d_v 为元素 v 的自然最近邻的标准差。详细的算法实现步骤见算法2。

算法2 扰动数据集生成算法。

输入 数据集 $U=\{v_1,v_2,\cdots,v_n\}$，参数 α。

输出 U,U_1,U_2。

根据算法1获取每个元素的自然最近邻 $\mathcal{N}(v)$；

For $i=1:n$ do

通过 $\mathcal{N}(v_i)$ 计算 d_{v_i}；

$v_i^1=v_i+\alpha\times d_{v_i}$；

$v_i^2=v_i-\alpha\times d_{v_i}$；

End

返回 $U=\{v_1,v_2,\cdots,v_n\}$，$U_1=\{v_1^1,v_2^1,\cdots,v_n^1\}$，$U_2=\{v_1^2,v_2^2,\cdots,v_n^2\}$。

2.2 基于样本扰动的三支聚类算法

对于一个多维的数据集，不同的特征子集能够从不同的视角描述数据集。当利用样本不同的特征的数据子集时，得到不同的聚类结果。假设 U 是一个有 m 个特征的数据集，随机抽取部分特征，使用传统聚类算法得到一个聚类结果，重复这个过程 t 次，最终可以获得一组聚类结果 C_1、C_2、$\cdots$、C_t。将上述过程应用于扰动数据集 U_1 与 U_2 上，获得 $3t$ 个不同的聚类结果。聚类生成过程见算法3。

算法3 基聚类生成算法。

输入 数据集 $U=\{v_1,v_2,\cdots,v_n\}$，聚类次数 t，类别数 k。

输出 $\mathscr{C}_1^1,\mathscr{C}_1^2,\mathscr{C}_1^3,\cdots,\mathscr{C}_t^1,\mathscr{C}_t^2,\mathscr{C}_t^3$。

根据算法2获取 U、U_1、U_2。

For $i=1:t$ do

随机提取部分特征；

使用K-means算法得到聚类 $\mathscr{C}_i^1$、$\mathscr{C}_i^2$、$\mathscr{C}_i^3$；

End

返回 $\mathscr{C}_1^1,\mathscr{C}_1^2,\mathscr{C}_1^3,\cdots,\mathscr{C}_t^1,\mathscr{C}_t^2,\mathscr{C}_t^3$。

假设数据样本 $U=\{v_1,v_2,\cdots,v_n\}$，通过样本扰动和抽取部分特征子集利用K-means算法得到了一组聚类结果 $\Pi=(\mathscr{C}_1,\mathscr{C}_2,\cdots,\mathscr{C}_M)$。构建关系矩阵[18]，任意两点共现概率为

$$p_{ij}=\frac{1}{M}\sum_{m=1}^{M}\prod(\mathscr{C}_m(x_i),\mathscr{C}_m(x_j)), \tag{7}$$

式中：M 代表不同的聚类结果；v_i 和 v_j 表示2个不同的数据样本；$\mathscr{C}_m(v_i)$ 表示点 v_i 在第 m 个聚类结果中的

簇数,即

$$\prod(\mathscr{C}_m(v_i),\mathscr{C}_m(v_j))=\begin{cases}1, & \mathscr{C}_m(v_i)=\mathscr{C}_m(v_j),\\ 0, & \mathscr{C}_m(v_i)\neq\mathscr{C}_m(v_j)。\end{cases} \tag{8}$$

基于以上的讨论,由式(4)得每个样本的稳定性,根据稳定性阈值 t_s 将数据集分为 2 个区域,由 Otsu 算法[19],得

$$O=\{i\mid s_i^M>t_s,\ i=1,2,\cdots,n\}, \tag{9}$$

$$H=\{i\mid s_i^M\leqslant t_s,\ i=1,2,\cdots,n\}, \tag{10}$$

式中,O 中的元素具有较高的稳定性,而 H 中的元素具有较低的稳定性。

对于稳定数据集 O,通过 K-means 算法获得稳定集 O 的元素的簇,这个簇是三支聚类结果的核心区域,即 C_{core}。对于不稳定的数据集 U 中的每个元素 v,计算 v 和 O 中 k 个类簇中心的最小距离 $d(v,x_i)=\min\limits_{1\leqslant j\leqslant k}d(v,x_j)$,$d(v,x_j)$是 v 和 x_j 之间的欧氏距离。给定一个参数 p_1,得到集合 $T=\{j:d(v,x_j)-d(v,x_i)\leqslant p_1\wedge j\neq i\}$,集合 T 有 2 种情况:如果 $T=\varnothing$,则 $v\in C^i_{\text{fringe}}$;如果 $T\neq\varnothing$,则 $v\in C^i_{\text{fringe}}$且 $v\in C^j_{\text{fringe}}$,这样就得到了每个类的核心域和边界域。算法 4 给出了基于样本扰动理论的三支聚类的过程。

算法 4　基于样本扰动的三支聚类算法。

输入　数据集 $U=\{v_1,v_2,\cdots,v_n\}$,参数 α、p_1,基聚类大小 t,聚类数 k。

输出　$\mathscr{C}=\{(C^1_{\text{core}},C^2_{\text{fringe}}),(C^2_{\text{core}},C^2_{\text{fringe}}),\cdots,(C^k_{\text{core}},C^k_{\text{fringe}})\}$。

① 利用算法 1 获得样本的自然最近邻;
② 利用算法 2 得到扰动数据集;
③ 利用算法 3 生成不同的聚类结果;
④ 利用式(7) 计算任意 2 个样本的共现概率;
⑤ 利用稳定性定义(4)式计算每个样本的稳定性;
⑥ 使用 Otus 算法获取阈值 t_s;
⑦ 分别通过公式(9)、(10)得到稳定集 O 和不稳定集 H;
⑧ 通过在稳定集 O 上使用 K-means 算法得到核心区域,即 $C^i_{\text{core}}(i=1,2,\cdots,k)$;
⑨ For $h=1,2,\cdots,|H|$ do
⑩　计算元素 v_h 与距其最近的聚类中心 x^i 之间的距离 $d(v_h,x_i)=\min\limits_{1\leqslant j\leqslant k}d(v_h,x_j)$ 和集合 $T=\{j:d(v_h,x_j)-d(v_h,x_i)\leqslant p_1\wedge j\neq i\}$;
⑪　If $T=\varnothing$,则;
⑫　　将元素 v_h 分配到边界域 $v\in C^i_{\text{fringe}}$;
⑬　　Else
⑭　　将元素 v_h 同时分配到边界域 $v\in C^i_{\text{fringe}}$和 $v\in C^j_{\text{fringe}}$;
⑮　End
⑯ End

返回　核心域与边界域

3　实验分析

为了验证本文提出的自然最近邻的样本扰动三支聚类算法的有效性,选取 5 组常见的公开数据集。表 1 列出了实验中使用的数据集信息。

为了评价聚类的效果,本文采用调整互信息、调整兰德指数、标准互信息和准确率 4 个指标检验聚类效果,这 4 个指标使用所有的核心区域来表示聚类结果。选取全部特征 60%的特征子集,参数 $\alpha=3$,阈值 $p=0.8$,在每个数据集和其对应的扰动数据集上各运行 60 次,结果如表 2 所示。为了对比本文算法的有效性,表中还给出了 K-means 算法与 Voting 算法的结果。

表1 实验中使用的数据集

Table 1 Datasets used in experiments

数据集	样本数	属性数	类别数
Iris	150	4	3
Wine	178	13	3
Seeds	210	7	3
Wdbc	569	30	2
Waveform	5 000	21	3

表2 不同算法不同数据集的实验结果

Table 2 Experimental results of different algorithms

算法	数据集	调整互信息	调整兰德指数	标准互信息	准确率
本文	Iris	**0.823 1**	**0.801 7**	**0.831 9**	**0.926 8**
	Wine	**0.885 1**	**0.906 8**	**0.887 3**	**0.969 7**
	Seeds	**0.665 5**	**0.700 8**	**0.690 0**	**0.895 0**
	Wdbc	**0.624 2**	**0.758 9**	**0.645 6**	**0.935 1**
	Waveform	**0.365 0**	**0.255 5**	**0.365 3**	0.540 1
K-means	Iris	0.738 7	0.716 3	0.741 9	0.886 7
	Wine	0.789 3	0.784 2	0.791 5	0.927 0
	Seeds	0.341 2	0.318 0	0.348 8	0.885 7
	Wdbc	0.612 6	0.730 2	0.623 1	0.927 9
	Waveform	0.354 0	0.243 6	0.364 2	**0.553 0**
Voting	Iris	0.738 7	0.723 4	0.723 5	0.876 7
	Wine	0.568 3	0.462 5	0.571 2	0.907 1
	Seeds	0.526 3	0.425 6	0.528 6	0.880 5
	Wdbc	0.602 4	0.720 3	0.614 0	0.925 6
	Waveform	0.153 5	0.094 3	0.153 7	0.517 8

在大多数数据集上,本文算法在调整互信息、调整兰德指数、标准互信息和准确率4个评价指标上都具有优势。例如在Wine数据集上,本文算法的4个指标结果分别为0.885 1、0.906 8、0.887 3和0.969 7,优于其他的算法。

为了分析随机抽取不同比例的特征对聚类集成结果的影响,在5个不同的数据集上,分析了抽取不同特征比例时算法的综合效果。分别随机选择50%、60%、70%、80%和90%的特征,利用本文提出的样本扰动理论的三支聚类算法进行三支聚类,在不同的特征子集中进行50次,并对评价指标的结果进行平均。表3分别给出了5个数据集上4个聚类指标随抽取特征比例的变化规律。

表3 抽取不同特征比例时的实验结果

Table 3 Results of different feature proportions on five datasets

数据集	特征比例/%	调整互信息	调整兰德指数	标准互信息	准确率
Iris	50	0.831 5	0.802 9	0.835 0	0.916 6
	60	0.823 1	0.801 7	0.831 9	0.926 8
	70	0.823 3	0.801 9	0.834 1	0.926 9
	80	0.822 7	0.800 0	0.831 9	0.926 9
	90	0.823 0	0.800 1	0.831 8	0.927 0
Wine	50	0.885 8	0.907 4	0.887 0	0.961 5
	60	0.885 1	0.906 8	0.887 3	0.969 7
	70	0.885 0	0.906 9	0.887 3	0.969 8
	80	0.885 2	0.910 0	0.890 0	0.970 2
	90	0.885 4	0.885 5	0.886 8	0.970 0
Seeds	50	0.665 4	0.700 4	0.700 2	0.900 0
	60	0.665 5	0.700 8	0.690 0	0.895 0
	70	0.654 7	0.701 2	0.700 3	0.895 8
	80	0.663 3	0.701 4	0.701 6	0.880 5
	90	0.661 8	0.701 9	0.699 5	0.896 2

续表

数据集	特征比例/%	调整互信息	调整兰德指数	标准互信息	准确率
Wdbc	50	0.620 1	0.728 5	0.648 9	0.920 1
	60	0.624 2	0.758 9	0.645 6	0.935 1
	70	0.621 4	0.751 5	0.638 3	0.935 8
	80	0.618 9	0.751 9	0.645 5	0.936 1
	90	0.620 0	0.752 4	0.646 7	0.940 2
Waveform	50	0.375 1	0.265 2	0.375 4	0.538 9
	60	0.365 0	0.255 5	0.365 3	0.540 1
	70	0.364 9	0.255 8	0.365 2	0.545 3
	80	0.365 1	0.256 0	0.365 5	0.545 5
	90	0.364 8	0.255 1	0.365 2	0.543 8

从表 3 中每一列可以发现,不同的数据集在不同特征比例时获得最佳性能。随着特征子集比例的增加,该算法中使用的 4 个评价指标在每个数据集上都会有少许波动。特征比例发生变化时,单个数据集的聚类结果会受到轻微影响,但总体而言变化很小。

4　结论

本文给出了一种基于自然最近邻的样本扰动三支聚类算法,该算法使用自然最近邻和样本的扰动理论获得每个扰动数据集,随机提取样本的特征子集,并使用传统的聚类算法来获得不同的基聚类,通过共现概率矩阵和确定性函数来获得每个样本的稳定性,利用 Otus 算法获得的稳定阈值将样本划分为稳定区域和不稳定区域,不同的区域采用不同的策略。稳定区域由具有高稳定性的样本组成,K-means 算法将高稳定性的样本分配给每个聚类的核心域;稳定区域由稳定性低的样本组成,稳定性低的样本分配给每个类簇的边界域。通过以上分配策略得到三支集成聚类结果。实验结果表明,与传统的集成聚类算法相比,本文的算法能够有效地揭示数据结构。

参考文献:

[1] FUJITA H, LI T R, YAO Y Y. Advances in three-way decisions and granular computing[J]. Knowledge-based Systems, 2016, 91:1-3.

[2] QIAN Yuhua, CHENG Honghong, WANG Jieting, et al. Grouping granular structures in human granulation intelligence[J]. Information Sciences, 2017, 382/383:150-169.

[3] XU Weuhua, YUAN Kehua, LI Weitao. Dynamic updating approximations of local generalized multi-granulation neighborhood rough set[J]. Applied Intelligence, 2022, 52(8):9148-9173.

[4] JI Xia, LIU Shuaishuai, ZHAO Peng, et al. Clustering ensemble based on sample's certainty[J]. Cognitive Computation, 2021, 13(4):1034-1046.

[5] RAO Liang, JIA Ningxin, HU Jun, et al. ATPdock: a template-based method for ATP-specific protein-ligand docking[J]. Bioinformatics, 2022, 38(2):556-558.

[6] NIU Chuang, SHAN Hongming, WANG Ge. SPICE: semantic pseudo-labeling for image clustering[J]. IEEE Transactions on Image Processing, 2022, 31:7264-7278.

[7] LIU Keyu, YANG Xibei, YU Hualong, et al. Supervised information granulation strategy for attribute reduction[J]. International Journal of Machine Learning and Cybernetics, 2020, 11(9):2149-2163.

[8] YAO Yiyu. Three-way decisions with probabilistic rough sets[J]. Information Sciences, 2010, 180(3):341-353.

[9] 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(3):321-329.
LI Jinhai, DENG Shuo. Concept lattice, three-way decisions and their research outlooks[J]. Journal of Northwest University (Natural Science Edition), 2017, 47(3):321-329.

[10] YU Hong, WANG Xinchen, WANG Guoying, et al. An active three-way clustering method via low-rank matrices for multi-view data[J]. Information Sciences, 2020, 507:823-839.

(下转第 62 页)

sets[J]. Journal of Nanjing University(Natural Science), 2021, 57(1):141-149.
[11] 徐怡,唐静昕. 基于优化可辨识矩阵和改进差别信息树的属性约简算法[J]. 计算机科学, 2020, 47(3):73-78.
XU Yi, TANG Jingxin. Attribute reduction algorithm based on optimized discernibility matrix and improving discernibility information tree[J]. Computer Science, 2020, 47(3):73-78.
[12] 翁冉,王俊红,魏巍,等. 基于区分矩阵的多粒度属性约简[J]. 南京航空航天大学学报, 2019, 51(5): 636-642.
WENG Ran, WANG Junhong, WEI Wei, et al. Multi-granulation attribute reduction based on discernibility matrix[J]. Journal of Nanjing University of Aeronautics & Astronautics, 2019, 51(5):636-642.
[13] 王国胤. 决策表核属性的计算方法[J]. 计算机学报, 2003(5): 611-615.
WANG Guoyin. Calculation methods for core attributes of decision table[J]. Chinese Journal of Computers, 2003(5): 611-615.
[14] YAO Jingtao, VASILAKOS A V, PEDRYCZ W. Granular computing: perspectives and challenges[J]. IEEE Transactions on Cybernetics, 2013, 43(6):1977-1989.
[15] YAO Jingtao. Information granulation and granular relationships[C]//IEEE International Conference on Granular Computing. Beijing: IEEE, 2005.
[16] 李丹. 多粒度粗糙集模型下的矩阵属性约简算法[J]. 计算机工程与应用, 2017, 53(19):168-172.
LI Dan. Matrix-based attribute reduction approach under multigranulation rough set[J]. Computer Engineering and Applications, 2017, 53(19):168-172.
[17] QIAN Yuhua, ZHANG Hu, SANG Yanli, et al. Multigranulation decision-theoretic rough sets[J]. International Journal of Approximate Reasoning, 2014, 55:225-237.
[18] SKOWRON A, RAUSZER C. The discernibility matrices and functions in information systems[M]//SLOWINSKI R. Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory. Dordrecht: Kluwer Academic Publisher, 1991:331-362.
[19] GE Meijun, FAN Nianbai, SUN Tao. Attribute reduction algorithm based on structure discernibility matrix in composite information systems[C]//International Conference on Information Science and Technology. Wuhan: ITM Web of Conferences, 2017, 11:01016.

(编辑:陈丽萍)

(上接第51页)

[11] WANG Pingxin, YAO Yiyu. CE3: a three-way clustering method based on mathematical morphology[J]. Konwledge-based Systems, 2018, 155:54-65.
[12] YU Hui, CHEN Luyuan, YAO Jingtao, et al. A three-way clustering method based on an improved DBSCAN algorithm[J]. Physica A,Statistical Mechanics and Its Applications, 2019, 535:122289.
[13] 凡嘉琛,王平心,杨习贝. 基于三支决策的密度敏感谱聚类[J]. 山东大学学报(理学版), 2022, 57(11):10-17.
FAN Jiachen, WANG Pingxin, YANG Xibei. Density sensitive spectral clustering based on three-way decision[J]. Journal of Shandong University (Natural Science) 2022, 57(11):10-17.
[14] 姜春茂,赵书宝. 基于阴影集的多粒度三支聚类集成[J]. 电子学报, 2021, 49(8):1524-1532.
JIANG Chunmao, ZHAO Shubao. Multi-granulation three-way clustering ensemble based on shadowed sets[J]. Acta Electronica Sinica, 2021, 49(8):1524-1532.
[15] FAN Jiachen, WANG Pingxin, JIANG Chunmao, et al. Ensemble learning using three-way density-sensitive spectral clustering[J]. International Journal of Approximate Reasoning, 2022, 149:70-84.
[16] ZOU Xianlin, ZHU Qingsheng, YANG Ruilong. Natural nearest neighbor for isomap algorithm without free-paramater[J]. Advanced Materials Research, 2011, 219/220:994-998.
[17] LI Feijiang, QIAN Yuhua, WANG Jieting, et al. Clustering ensemble based on sample's stability[J]. Artificial Intelligence, 2019, 273:37-55.
[18] 李飞江,钱宇华,王婕婷,等. 基于样本稳定性的聚类方法[J]. 中国科学(信息科学), 2020, 50(8):1239-1254.
LI Feijiang, QIAN Yuhua, WANG Jieting, et al. Clustering method based on sample's stability[J]. Scientia Sinica Informationis, 2020, 50(8):1239-1254.
[19] OTUS N. A threshold selection method from gray-level histogarms[J]. IEEE Transcations on Systems, Man, and Cybernetics, 1979, 9:62-66.

(编辑:陈丽萍)