Advances in Applied Mathematics 应用数学进展, 2022, 11(2), 858-865
Published Online February 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.112092

Hans汉斯

基于三支决策的密度聚类算法

姜　凡

江苏科技大学理学院，江苏　镇江

收稿日期：2022年1月23日；录用日期：2022年2月21日；发布日期：2022年2月28日

摘　要

三支聚类使用核心域，边界域和琐碎域三个集合来表示类簇，将确定的元素放入核心域中，不确定的元素放入边界域中延迟决策，降低了决策风险。本文将含有噪声的基于密度的聚类算法(Density Based Spatial Clustering of Application with Noise, DBSCAN)与三支聚类进行结合，利用数学形态学中的腐蚀和膨胀思想，用自然最近邻算法定义了一个结构算子，对二支聚类的结果通过收缩和膨胀得到核心域和边界域。在UCI数据集和Shape数据集上的实验结果显示，该方法可以有效降低DBI的值，同时提高ACC和AS的值。

关键词

三支聚类，DBSCAN，数学形态学，自然最近邻

Density Based Spatial Clustering of Application with Noise Based on Three-Way Decision

Fan Jiang

School of Science, Jiangsu University of Science and Technology, Zhenjiang Jiangsu

Received: Jan. 23rd, 2022; accepted: Feb. 21st, 2022; published: Feb. 28th, 2022

Abstract

Three-way clustering uses three sets of core region, boundary region and trivial region to represent the clusters. The determined elements are put into the core region, while the uncertain elements are put into the boundary region to delay the decision, thus reducing the decision risk. In this paper, DBSCAN (Density Based Spatial Clustering of Application with Noise) is combined with

文章引用: 姜凡. 基于三支决策的密度聚类算法[J]. 应用数学进展, 2022, 11(2): 858-865.
DOI: 10.12677/aam.2022.112092

the three-way clustering, and a structural operator is defined by using the natural nearest neighbor algorithm based on the corrosion and expansion ideas in mathematical morphology. The core region and boundary region are obtained by shrinking and expanding the results of the two-way clustering. Experimental results on UCI datasets and shape datasets show that this method can effectively reduce the value of DBI and improve the value of ACC and AS.

Keywords

Three-Way Clustering, DBSCAN, Mathematical Morphology, Natural Nearest Neighbor

1. 引言

聚类作为一种无监督的学习方式，广泛应用于图像处理[1]、社会网络[2]等领域。简单来说，聚类就是将数据集划分成多个不同的类簇，使得同类簇中的数据点相似性较高，不同类簇中的数据点相似性较低。按聚类原理，目前流行的聚类算法可分为基于密度聚类、基于层次聚类、基于网格聚类、基于划分聚类、基于模型聚类五种[3]。DBSCAN (Density Based Spatial Clustering of Application with Noise) [4]就是基于密度的聚类算法，它可以在含有噪声的数据中识别出任意形状的类簇，在医学诊断[5]、目标检测[6]等领域广泛应用。

传统聚类算法都属于硬聚类，聚类结果中对象要么属于一个类，要么不属于一个类，类簇之间有清晰的边界，但在处理不确定信息时，如果强制将某个对象划分到类簇中，会带来较高决策风险，降低聚类精度。为了解决传统聚类方法存在的问题，很多新的聚类方法被提出，Lingras [7]提出粗糙聚类方法，用粗糙集的正域、负域和边界域来表示聚类结果。Yao [8]等人用区间集来表示聚类结果中的一个类。Yu [9] [10] [11]将三支决策引入聚类中，提出三支聚类的方法，Wang 和 Yao 等人[12]提出了基于三支聚类的 CE3 理论，Wang 等人[13]融合了 k-means 算法和三支决策理论，提出了三支 k-means 算法，Yang [14]等人用复杂网络来分析三支决策文章中的数据集，构建了科学合作网络、大学合作网络、科学论文网络和关键词网络，分别揭示了作者、单位、论文以及关键词之间的关系，并解答了三支决策的相关问题。

最近，Yu [15]等人提出了基于 3W-DBSCAN 算法，通过改进相似度的计算获得硬聚类结果，进而进行三支聚类。本文利用数学形态学中的腐蚀和膨胀思想，对 DBSCAN 获得的硬聚类结果，采用基于样本的自然最近邻域对其进行收缩和膨胀得到三支聚类的核心域与边界域，将不确定的元素划分到边界域中延迟决策，待信息充分时再做决策，以此来降低决策风险。仿真实验也验证了该方法的有效性和可行性。

2. 相关工作

2.1. 三支决策聚类

2010 年，Yao [16] [17]等人提出三支决策理论，核心思想是将研究对象分为正域、负域、边界域，使其更符合实际生活中人们认知的一种决策模式，三个域分别对应三种决策规则：接受、拒绝以及不承诺规则。相比于二支决策，三支决策能够有效降低信息不充分时的决策风险。

三支决策理论提出以来，广受各领域关注，刘强[18]等人提出基于 ε 邻域的三支决策聚类的方法。Yu

[19]等人提出用三支决策来检测和完善复杂网络中的重叠区域。Zhang [20]等人建立了一个动态的三支决策模型。

三支聚类用 $Co(C)$、$Fr(C)$、$Tr(C)$ 三个集合来表示一个类簇，即核心域、边界域和琐碎域。其中核心域中的元素确定属于类 C，边界域中的元素可能属于类 C，琐碎域中的元素确定不属于类 C。结果表示如下：

$$CS=\left\{\left\{Co(C_1),Fr(C_1)\right\},\cdots,\left\{Co(C_k),Fr(C_k)\right\}\right\}$$

式中这三个集合满足如下性质：

$$CS=\left\{\left\{Co(C_1),Fr(C_1)\right\},\cdots,\left\{Co(C_k),Fr(C_k)\right\}\right\}.$$

式中这三个集合满足如下性质：

1) $Co(C_i)\neq\varnothing, i=1,2,\cdots,k$

2) $\bigcup_{i=1}^{k}\left(Co(C_i)\cup Fr(C_i)\right)=U$

3) $Co(C_i)\cup Fr(C_i)\cup Tr(C_i)=U$

性质 1) 表示 $Co(C)$ 不为空集，即每个类中至少要有一个对象；

性质 2) 确保集合 U 中每个对象至少被划分到一个类中；

性质 3) 确保任何一个类簇的三个集合的并集为 U。

2.2. DBSCAN 算法

DBSCAN 是一种典型的基于密度的聚类算法，其核心思想[21]是用一个点的 ε 邻域内的邻居点个数来衡量该点所在空间的密度，优点是聚类时不需要事先确定类簇的个数。DBSCAN 算法的主要定义如下：

定义 1 (*Eps* 邻域)：给定空间中的一点 p，p 的 *Eps* 邻域点集为以 p 为中心，*Eps* 为半径的区域内包含样本点的集合，即

$$N_{Eps}(p)=\left\{q\in D \mid dist(p,q)\leq Eps\right\} \tag{1}$$

其中：D 为样本数据集，$dist(p,q)$ 为点 p 和点 q 之间的距离。

定义 2 (*Minpts* 密度阈值)：形成一个高密度区域时，*Eps* 邻域内至少需要的点的个数。

DBSCAN 算法步骤如下：

算法 1 DBSCAN 算法

输入：数据集 D。

输出：聚类结果 $\{C_1,C_2,\cdots,C_k\}\cup\{NO\}$。

1：初始化参数：邻域半径 ε，邻域密度阈值 *Minpts*，标记所有对象为 **unvisited**。

2：while 存在标记为 unvisited 的对象：

随机选择一个 unvisited 对象 p，标记为 visited；

if p 的 ε -邻域至少有 *Minpts* 个对象：

创建一个新簇 C,并把 p 添加到 C 中，令 N 为 p 的 ε -邻域的对象集合；

for N 中的每个点 p' do：

if p' 是 unvisited：标记 p' 为 visited；

if p' 的 ε -邻域至少有 *Minpts* 个对象：把这些对象添加到 N，如果 p' 还不是任何簇的成员，把 p' 添加到 C 中；

else 标记 p 为噪声；

3：输出：聚类结果 $\{C_1,C_2,\cdots,C_k\}\cup\{NO\}$。

2.3. 自然最近邻

传统最近邻算法主要是 k 最近邻算法和 ε 最近邻算法，k 最近邻算法中每个数据点都有 k 个元素，ε 最近邻算法中每个数据点有相同的邻域半径 ε，但这两种方法都需要进行参数的选取，参数设置准确程度会直接影响算法结果。为此，我们采用自然最近邻居算法，自然最近邻居[22]不同于传统的近邻算法，它不需要邻域参数，因而避免了人为因素对算法结果的影响，有如下定义：

定义 1 (最近邻)：$NN_r(x_i)$ 表示样本点 x_i 的 r 最近邻，r 由自然最近邻搜索算法自动产生。

定义 2 (逆近邻)：$RNN_r(x_i)$ 表示样本点 x_i 的 r 逆最近邻：

$$RNN_r(x_i)=\left\{x_j \in X \mid x_i \in NN_r(x_j), i \neq j\right\} \tag{2}$$

定义 3 (自然最近邻)：$NNN(x_i)$ 表示样本点 x_i 的自然最近邻：

$$NNN(x_i)=\left\{x_j \in X \middle| x_i \in NN_r(x_j), x_i \in RNN_r(x_i)\right\} \tag{3}$$

定义 4 (自然邻居特征值 sup_k)：当所有的样本点都有逆近邻或者逆近邻个数为 0 的样本点不变时，自然近邻搜索过程到达自然稳定状态，此时的搜索次数称为自然特征值，记为

$$sup_k=\left\{r \mid \forall x\, \exists y\left(y \neq x \cap x \in NN_r(y)\right)\right\} \tag{4}$$

2.4. 数学形态学

数学形态学(Mathematical Morphology)由 Matheron 和 Serra 提出[23]，是图像处理中应用最为广泛的技术之一，主要用于从图像中提取对表达和描述区域形状有意义的图像分量，使后续的识别工作能够抓住目标对象最为本质的形状特征。其基本思想是用具有一点形态的结构元素去度量和提取图像中对应形状以达到图像分析和识别的目的[24]。Bloch [25]将数学形态学与粗糙集联系起来，下近似对应腐蚀操作，上近似对应膨胀操作，在此我们引入腐蚀和膨胀操作，有如下定义：

定义 1 (腐蚀)：用集合 B 来腐蚀集合 A，记作 $A \Theta B$：

$$A \Theta B=\left\{x \middle| (B)x \subseteq A\right\} \tag{5}$$

其中 B 也称为结构元素，$(B)x$ 表示 B 平移 x 后得到的新集合，由上式可知，腐蚀可以使目标区域范围变小，其实质是造成图像的边界收缩，可以用来去除没有意义的目标物。

定义 2 (膨胀)：用结构算子 B 来膨胀 A，记作 $A \oplus B$：

$$A \oplus B=\left\{x \middle| \left(\hat{B}\right)x \cap A \neq \varnothing\right\} \tag{6}$$

其中集合 B 也称为结构元素，$\left(\hat{B}\right)x$ 表示 B 反射平移 x 后得到的新集合，由上式可知，膨胀会使目标边界向外部扩张，作用是用来填补目标区域中某些空洞以及消除包含在目标区域中的小颗粒噪声。

3. 基于三支决策的密度聚类算法

DBSCAN 硬聚类的结果包括了噪声数据，因此在进行三支聚类的时候，应该将噪声数据单独进行处理，首先给出如下定义：

定义 1：设 $\{C_1, C_2, \cdots, C_k\}$ 是非噪声集合，$x_j \in C_i$，$x_t \in NNN_{nb}(x_j)$，x_j 的平均自然最近邻居距离 $avg(x_j)=\dfrac{1}{nb(x_j)}\sum_{t=1}^{nb(x_j)} d(x_t, x_j)$，类 C_i 的平均自然最近邻距离为 $avg(C_i)=\dfrac{1}{|C_i|}\sum_{j=1}^{|C_i|} avg(x_j)$。

定义 2： 令 $Allpos=\bigcup_{i=1}^{k} pos\left(C_i\right)$ 为所有核心点的集合，则有如下定义：

$$NCN\left(x\right)=y\leftarrow \arg\min d\left(x,y\right),y\in Allpos$$

三支 DBSCAN 算法分为以下几步：

1) 对于非噪声集合 $\left\{C_1,C_2,\cdots,C_k\right\}$，任取 $x_b\notin C_i$，如果 x_b 的自然最近邻域与 C_i 有交集，那么就把 x_b 划分到 C_i 的边界域中；如果对于 $x_j\in C_i$，x_j 的自然最近邻域不包含于 C_i 中，则把 x_j 划分到 C_i 的边界域中。

2) 给定一个参数 ρ $\left(\rho>1\right)$，设 $x_j\in C_i$，计算 x_j 的平均自然最近邻居距离，以及类 C_i 的平均自然最近邻距离，比较 $\rho avg\left(C_i\right)$ 与 $avg\left(x_j\right)$ 大小，如果前者比较大，则划分到对应的核心域，反之划分到边界域中。

3) 处理噪声数据。计算噪声点到各个核心域中的元素的最小距离，将噪声点分配到核心点所在类簇边界域中。

算法 2　基于 DBSCAN 的三支聚类

输入：数据硬聚类结果 $\left\{C_1,C_2,\cdots,C_k\right\}\cup\left\{NO\right\}$。

输出：聚类结果。

1：初始化参数：参数 ρ，排除噪声点数据集 D。

2 算法步骤：

2.1 处理非噪声点：

for each $x_i\in D$：

for each $C_j\in\left\{C_1,C_2,\cdots,C_k\right\}$：

if $x_i\notin C_j$：

if $NNN_{nb\left(x_i\right)}\left(x_i\right)\cap C_j\neq\varnothing$：$x_i\in Fr\left(C_j\right)$；

if $x_i\in C_j$：

if $NNN_{nb\left(x_i\right)}\left(x_i\right)\subsetneq C_j:x_i\in Fr\left(C_j\right)$；

if $avg\left(x_i\right)>\rho avg\left(C_j\right):x_i\in Fr\left(C_j\right)$；

else $x_i\in Co\left(C_j\right)$；

2.2 处理噪声点：

for each $x_i\in NO$：

for each $C_j\in\left\{C_1,C_2,\cdots,C_k\right\}$：

if $NCN\left(x_i\right)\in Co\left(C_j\right)$：

$Fr\left(C_j\right)=Fr\left(C_j\right)\cup x$；

3：输出：聚类结果

4. 实验结果

论文采用三组 UCI 数据集和五组 shape [26]数据集，如表 1 及表 2。利用 ACC、DBI、AS 指标，将三支 DBSCAN 聚类与 DBSCAN 进行比较，得出三支 DBSCAN 聚类可以提高聚类精度、改善聚类性能的结论。

本实验先对每组数据进行遍历，选取一个最佳邻域半径和阈值，然后选取 DBI、ACC、AS 的值作为

评价指标，实验结果如表 3 所示。Congressional 和 R15 上的实验结果可以看出，三支 DBSCAN 的准确率达到了 99%，从表 3 可以看出，与 DBSCAN 算法相比较，该聚类算法在数据集上有较好的聚类效果，DBI 的值明显下降，ACC 和 AS 的值也均有上升，虽然 ACC 的上升并不明显，但相对于二支聚类，三支 DBSCAN 算法在结果上都得到了有效提升。综上所述，可以证实基于 DBSCAN 的三支聚类算法是有效的。

Table 1. UCI dataset
表 1. UCI 数据集

数据集	样本个数	样本维数	类别数
Iris	150	4	3
Congressional	435	16	2
Forest	523	27	4

Table 2. Shape dataset
表 2. Shape 数据集

数据集	样本个数	样本维数	类别数
Flame	240	2	2
Jain	373	2	2
R15	600	2	15
Aggregation	788	2	7
D31	3100	2	31

Table 3. Experimental results on datasets
表 3. 数据集上的实验结果

Datasets	Algorithm	DBI	AS	ACC
Iris	DBSCAN	0.4474	0.6180	0.9866
	Three-way DBSCAN	**0.3664**	**0.8698**	**1.0000**
Congressional	DBSCAN	0.8107	0.3979	0.9902
	Three-way DBSCAN	**0.5877**	**0.4613**	**0.9973**
Forest	DBSCAN	0.6025	0.6351	0.9812
	Three-way DBSCAN	**0.3459**	**0.7327**	**0.9976**
Aggregation	DBSCAN	0.5046	0.6498	0.7821
	Three-way DBSCAN	**0.4570**	**0.7189**	**0.7872**
Jain	DBSCAN	0.6018	0.4618	0.9302
	Three-way DBSCAN	**0.4456**	**0.6337**	**0.9377**
Flame	DBSCAN	0.8168	0.4417	0.9132
	Three-way DBSCAN	**0.7175**	**0.5136**	**0.9486**
R15	DBSCAN	0.3163	0.8971	0.9916
	Three-way DBSCAN	**0.1248**	**0.9532**	**0.9978**
D31	DBSCAN	0.5177	0.8001	0.9524
	Three-way DBSCAN	**0.4017**	**0.8911**	**0.9686**

5. 结束语

由于传统硬聚类在处理不确定信息时，强将元素划分到某一类别中，可能会带来决策风险。本文提出了借助样本的自然最近邻域将硬聚类结果转换成三支聚类，将不确定的元素划分到边界域中延迟决策，待信息充分时再做决策，以此提高聚类结果的准确性，实验结果也表明本文提出的算法可以提高其性能，可以有效降低 DBI 的值，同时提高 ACC 和 AS 的值。但由于 DBSCAN 聚类算法参数的配置需要人工判定，存在了一定的人为因素，同时由于本身的算法特点，DBSCAN 在处理低维数据时性能更优越，但随着数据维度的提高，该算法性能表现变差。考虑到这个问题，下一阶段使用深度神经网络对样本进行特征提取，以达到数据降维的目的，保证了 DBSCAN 算法的可靠性。

参考文献

[1] Elalami, M.E. (2011) Supporting Image Retrieval Framework with Rule Base System. *Knowledge-Based Systems*, **24**, 331-340. https://doi.org/10.1016/j.knosys.2010.10.005

[2] Chang, M.S., Chen, L.H., Hung, L.J., *et al*. (2014) Exact Algorithms for Problems Related to the Densest *k*-Set Problem. *Information Processing Letters*, **114**, 510-513. https://doi.org/10.1016/j.ipl.2014.04.009

[3] Xu, D. and Tian, Y. (2015) A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*, **2**, 165-193. https://doi.org/10.1007/s40745-015-0040-1

[4] Ester, M., Kriegel, H.P., Sander, J., *et al*. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the* 2*nd International Conference on Knowledge Discovery and Data Mining*, Portland, 2-4 August 1996, 226-231.

[5] Rangaprakash, D., Odemuyiwa, T., Narayana, D.N., *et al*. (2020) Density-Based Clustering of Static and Dynamic Functional MRI Connectivity Features Obtained from Subjects with Cognitive Impairment. *Brain Informatics*, **7**, Article No. 19. https://doi.org/10.1186/s40708-020-00120-2

[6] 岳晓新, 贾君霞, 陈喜东, 李广安. 改进 YOLO V3 的道路小目标检测[J]. 计算机工程与应用, 2020, 56(21): 218-223.

[7] Lingras, P. and West, C. (2004) Interval Set Clustering of Web Users with Rough *k*-Means. *Journal of Intelligent Information Systems*, **23**, 5-16. https://doi.org/10.1023/B:JIIS.0000029668.88665.1a

[8] Yao, Y.Y., Lingras, P., Wang, R.Z., *et al*. (2009) Interval Set Cluster Analysis: A Re-Formulation. *International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular Soft Computing*, Delhi, 15-18 December 2009, 398-405. https://doi.org/10.1007/978-3-642-10646-0_48

[9] Yu, H., Wang, X., Wang, G., *et al*. (2018) An Active Three-Way Clustering Method via Low-Rank Matrices for Multi-View Data. *Information Sciences*, **507**, 823-839. https://doi.org/10.1016/j.ins.2018.03.009

[10] Yu, H., Zhang, C. and Wang, G.Y. (2016) A Tree-Based Incremental Overlapping Clustering Method Using the Three-Way Decision Theory. *Knowledge-Based Systems*, **91**, 189-203. https://doi.org/10.1016/j.knosys.2015.05.028

[11] Yu, H. (2017) A Framework of Three-Way Cluster Analysis. *International Joint Conference on Rough Sets*, Olsztyn, 3-7 July 2017, 300-312. https://doi.org/10.1007/978-3-319-60840-2_22

[12] Wang, P.X. and Yao, Y.Y. (2018) CE3: A Three-Way Clustering Method Based on Mathematical Morphology. *Knowledge-Based Systems*, **155**, 54-65. https://doi.org/10.1016/j.knosys.2018.04.029

[13] Wang, P.X., Shi, H., Yang, X.B. and Mi, J. (2019) Three-Way *k*-Means: Integrating *k*-Means and Three-Way Decision. *International Journal of Machine Learning & Cybernetics*, **10**, 2767-2777. https://doi.org/10.1007/s13042-018-0901-y

[14] Yang, B. and Li, J.H. (2020) Complex Network Analysis of Three-Way Decision Researches. *International Journal of Machine Learning and Cybernetics*, **11**, 973-987. https://doi.org/10.1007/s13042-020-01082-x

[15] Yu, H., Chen, L.Y., Yao, J.T., *et al*. (2019) A Three-Way Clustering Method Based on an Improved DBSCAN Algorithm. *Physica A*: *Statistical Mechanics and its Applications*, **535**, Article ID: 122289. https://doi.org/10.1016/j.physa.2019.122289

[16] Yao, Y.Y. (2011) The Superiority of Three-Way Decisions in Probabilistic Rough Set Models. *Information Sciences*, **181**, 1080-1096. https://doi.org/10.1016/j.ins.2010.11.019

[17] Yao, Y.Y. (2012) An Outline of a Theory of Three-Way Decisions. *International Conference on Rough Sets and Current Trends in Computing*, Chengdu, 17-20 August, 1-17. https://doi.org/10.1007/978-3-642-32115-3_1

[18] 刘强, 施虹, 王平心, 杨习贝. 基于 ε 邻域的三支决策聚类分析[J]. 计算机工程与应用, 2019, 55(6): 140-144.

[19] Yu, H., Jiao, P., Yao, Y.Y., *et al*. (2016) Detecting and Refining Overlapping Regions in Complex Networks with Three-Way Decisions. *Information Sciences*, **373**, 21-41. https://doi.org/10.1016/j.ins.2016.08.087

[20] Zhang, Q., Lyu, G., Chen, Y., *et al*. (2018) A Dynamic Three-Way Decision Model Based on the Updating of Attribute Values. *Knowledge-Based Systems*, **142**, 71-84. https://doi.org/10.1016/j.knosys.2017.11.026

[21] 周红芳, 王鹏. DBSCAN 算法中参数自适应确定方法的研究[J]. 西安理工大学学报, 2012, 28(3): 291-293.

[22] Huang, J., Zhu, Q., Yang, L., *et al*. (2016) A Non-Parameter Outlier Detection Algorithm Based on Natural Neighbor. *Knowledge-Based Systems*, **92**, 71-77. https://doi.org/10.1016/j.knosys.2015.10.014

[23] Serra, J. (1986) Introduction to Mathematical Morphology. *Computer Vision Graphics & Image Processing*, **35**, 283-305. https://doi.org/10.1016/0734-189X(86)90002-2

[24] Banerji, A. (2000) An Introduction to Image Analysis Using Mathematical Morphology. *IEEE Engineering in Medicine and Biology Magazine*, **19**, 13-14. https://doi.org/10.1016/S0031-3203(99)00129-6

[25] Bloch, I. (2000) On Links between Mathematical Morphology and Rough Sets. *Pattern Recognition*, **33**, 1487-1496.

[26] Pasi, F. and Sami, S. (2018) K-Means Properties on Six Clustering Benchmark Datasets. *Applied Intelligence*, **48**, 4743-4759. https://doi.org/10.1007/s10489-018-1238-7