Advances in Applied Mathematics 应用数学进展, 2017, 6(8), 985-991
Published Online November 2017 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2017.68118

Hans汉斯

A Three-Way Spectral Clustering Algorithm Based on Perturbation Analysis

Xiaolei Wang, Pingxin Wang*

School of Science, Jiangsu University of Science and Technology, Zhenjiang Jiangsu
Email: *pingxin_wang@hotmail.com

Received: Nov. 8th, 2017; accepted: Nov. 21st, 2017; published: Nov. 27th, 2017

Abstract

We structure a three-way Clustering Algorithm Based on NJW Algorithm. The main idea is that each object, making weighted arithmetic, repeats clustering arithmetic to obtain the weighted results, which benefits from the stability of NJW Algorithm. According to the perturbation for all objects, we can classify the object into core or fringe regions. The compact degree of core and fringe regions can be described by the modified *S_Dbw*(*c*) index. Testing the algorithm on the artificial and UCI data sets, the results meet expectation.

Keywords

Three-Way Clustering, Disturbance Degree, NJW Algorithm

一种基于扰动分析的三支谱聚类算法

王晓磊，王平心*

江苏科技大学理学院，江苏 镇江
Email: *pingxin_wang@hotmail.com

收稿日期：2017年11月8日；录用日期：2017年11月21日；发布日期：2017年11月27日

摘 要

本文在多路谱聚类算法(NJW算法)的基础上进行三支聚类算法构造。主要思想是针对每个对象进行加权运算，利用多路谱聚类算法的稳定性，重复进行聚类运算从而获得加权结果。根据其加权后对全体对象的扰动影响，将其划入核心域或边界域。对*S_Dbw*(*c*)指标进行改造使其可以表示核心域与边界域的离散程度。最后分别在人工与UCI数据集上对算法进行测试，有较好效果。

*通讯作者。

文章引用: 王晓磊, 王平心. 一种基于扰动分析的三支谱聚类算法[J]. 应用数学进展, 2017, 6(8): 985-991.
DOI: 10.12677/aam.2017.68118

关键词

三支聚类，扰动影响，谱聚类

1. 引言

聚类[1] [2] [3]可以看作一种分类，是将具有一定相似度的对象按照其自身的特性进行分类，是当代分类学与多元分析的结合，并已用于多个领域[4]。由于聚类所生成的簇是一组数据对象的集合，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。在聚类的研究和应用中存在两个主要问题[5]。首先，如何有效的划分一个数据集。其次，如何确定划分出的类簇个数。使用不同的算法思想对同一个数据集进行操作往往会产生不同的聚类结果。产生分歧的原因往往是一些元素特征并不明显或过于明显，使得在分类过程中既可以被分入这个类也可以被分入那个类。因此会降低聚类的效果。

聚类结果的效果如何，取决于数据集中各个元素能否依照算法准确地被划分入相应类簇。无法被准确分类的元素除特征明显，与其余点明显相异的噪点(孤立点)，还有介于两类簇之间，既可以被划入此类簇也可以被划入彼类簇的边界点。因此引入三支聚类的概念，依照核心域，边界域，琐碎域对数据集进行划分，解决了由于特征不清导致分类不准的问题。

2. 相关工作与概念

目前，大多二支聚类结果都是一种硬划分，即对某一个元素都有唯一确定的类簇与之匹配。而在某些数据集中，由于数据本身特征存在模糊与相异度较低的情况，无法对其进行准确的划分，此时被强制分配的元素会大大降低二支聚类结果的精度，破坏其结构。例如图 1 所示，位于中间的元素在两类簇的边界处出现交叉，其依照算法被划分为两类，但是从观察中可得这些元素并未明显差异。

以下介绍三支聚类的有关概念：

设给定一个数据集 $U=\{x_1,x_2,\cdots,x_n\}$，依照二支聚类的定义，由一个集合代表一个类簇，即寻找一组集合 $C_1,C_2,\cdots,C_k$ 满足 $U=\bigcup_{i=1}^{k}C_i$，且满足 $C_i\cap C_j=\varnothing$，$(i,j=1,2,\cdots,k,i\neq j)$，其中 k 为聚类的个数。

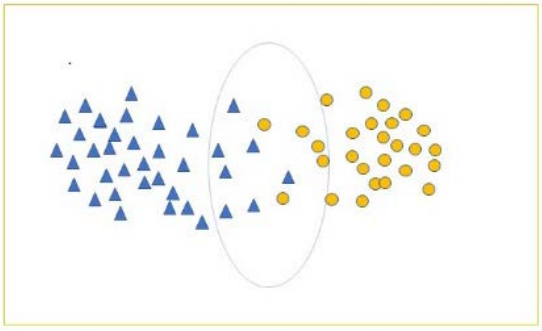

Figure 1. A schematic diagram of data set
图 1. 数据集示意图

在三支聚类中，使用三个互不相交的集合表示一个类，即 C_i^P，C_i^B 与 C_i^N，分别称为类的核心域，边界域和琐碎域，其中

$$C_i^P \cup C_i^B \cup C_i^N = U \tag{1}$$

核心域中的元素确定属于这个类，边界域中的元素可能属于也可能不属于这个类，而琐碎域中的元素则肯定不属于这个类。由关系式(1)，我们可以通过 C_i^P 和 C_i^B 来表示一个类。反过来，如果给定一组集合 C_i^P，C_i^B，$i=1,2,\cdots,k$ 满足

$$C_i^p \neq \varnothing, i=1,2,\cdots,k \tag{2}$$

$$\bigcup_{i=1}^{k}\left(C_i^p \cup C_i^B\right)=U \tag{3}$$

我们称其为数据集 U 三支聚类。其中(2)要求正域非空，即每个类中至少有一个对象，而(3)保证每个对象至少被分到一个类中。和传统的硬聚类的结果 $C=\{C_1,C_2,\cdots,C_k\}$ 不同，三支聚类的结果应为如下形式，

$$TC=\left\{\left(C_1^P,C_1^B\right),\cdots,\left(C_k^P,C_k^B\right)\right\}$$

显然在三支聚类中如果有 $C_i^B=\varnothing$，$(i=1,2,\cdots,k)$，就变成了 传统二支决策的聚类形式。因此，三支决策聚类形式是传统二支聚类方法的推广，使用边界域去解决针对目前知识体系下难以聚类的对象，或是特征不明显难以准确归类的元素，即为三支聚类的主要思想。

3. 基于扰动的三支聚类算法

本算法的主要思想是对每个对象进行加权处理，之后依照前后对比进行分类。首先使用 NJW 算法对数据集进行一次划分，之后在此次划分基础上对选定对象加权重新聚类，比较两次聚类的扰动情况。从图 1 中可以看出属于边界域的点往往会呈现一种边缘性，正是因为这种边缘性使得划分结果的不准确。边缘性反映在数据上即为元素距类簇中心的距离，但是单独判断显然不是一个可行方法。首先。作为一种无监督学习，在分析之前不知道判断距离的标准。其次，单纯的判断距离无法识别出所判断点所处区域的数据稠密情况，疏密情况对于域的划分有着显著的影响，最后，距离判断误差极大，一次性会将某一范围外的所有点归为边界域，失去参考价值。所以本文采用扰动的方法，设立标准扰动量，计算每个点加权后对全局的扰动，与标准扰动量进行比较判断所属域。

参数 m 为对某点添加新点的个数，通常为聚类总对象数量的特定倍数。通常这个倍数不易过大，一旦过大，新的类簇中心必然出现在选定点上，此时就转换为单纯的比较距离。失去了扰动算法的本质意义。

标准扰动距离并非一成不变，而是针对不同的数据集特意选定。并且标准扰动距离与参数 m 呈正相关(当添加的点愈多，偏移愈大)。两者需要考量数据集的形状，类型，才能较好的选定。这里给出一种通常的选定方式，通常把数据集分位数进行加权处理，以此作为标准扰动距离。

算法 1： 基于全局扰动的三支谱聚类算法

Step1 通过 NJW 算法对数据集 U 进行一次聚类操作，得到 $C=\{C_1,C_2,\cdots,C_k\}$；

Step 2 在数据集中任找一个尚未被判断的点 q，在原本数据集中添加 m 个数据 q 得到新数据集 U_1，对 U_1 应用 NJW 算法重新聚类，得到 $C^*=\{C_1^*,C_2^*,C_3^*,\cdots\}$。

Step3 在 C^* 中找到 q 所对应的聚类 C_i^*，比较两聚类中心的欧氏距离，当其距离差值大于标准扰动聚类 w 时，将 q 划分至边界域，反之，将 q 划分至核心域。

Step4 若 q 以外所有点都已被判断则转入 Step5，否则转入 Step2。

Step5 输出核心域与边界域。

由于算法本身的不足，造成此算法的时间复杂度极大。这是由于在聚类过程中对于每个点都要进行一次聚类操作，并且每此参与聚类的数据量还大于原数据集。为了解决上述问题，以下考虑一种基于局部的聚类算法。仅在初始时进行一次聚类运算，之后在判断聚类中心时仅仅在一个类簇内进行考虑而非对全部数据进行操作。以此大大简化了运算量，当然精度不可避免的造成一定损失。但是经实验检测，此算法依然有较好的实验效果。

算法 2：基于局部扰动的三支谱聚类算法

Step1 通过 NJW 算法对数据集 U 进行一次聚类操作，得到 $C=\{C_1,C_2,\cdots,C_k\}$；

Step 2 在数据集中任找一个尚未被判断的点 q， $q\in C_i$，在原本数据集 U 中添加 m 个数据得到新数据集 U_1，找到 q 原本所在类簇 C_i^*。记 $\frac{\text{添加的对象数}}{\text{数据集总对象数}}=\text{扰动系数}$。

Step3 计算 C_i^* 的聚类中心，将其与原本 C 的中心进行比较，若欧式距离差大于标准扰动距离与扰动系数的乘积则将其划分至边界域，反之，将其划分至核心域。

Step4 若 q 以外所有点都已被判断则转入 Step5，否则转入 Step2

Step5 输出核心域与边界域。

基于局部的算法中，将原本需要重新聚类获得的聚类中心转变为直接在原本类簇中增加相应的点进行求中心处理。由于不需要考虑全局，可以取较小值，将添加的类簇数据量倍数记为扰动系数。这是因为单个聚类本身具有极好的扰动特性(完全的物理特性)。此时标准扰动距离与扰动系数的乘积在实验中有良好的表现。

4. 实验结果

4.1. 人工数据集

本节采用一组人工数据集 U，U 为从均值为 μ，协方差为 σ 的正态分布中抽取 $n*d$ 的矩阵 $\boldsymbol{R}$(n 为样本数，d 为样本维数，本实验 d 为 2)。

其中，$\mu=\begin{bmatrix} a_1 & a_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ a_7 & a_8 \end{bmatrix}$，$R$ 中的每一行以中对应的行为均值的正态分布中抽取的一个样本。同时 $a_i=\left(\left(rand\left(0,1\right)-1\right)*11\right)+1,\ i=1,\cdots,8$，$rand\left(0,1\right)$ 表示在 0~1 中取随机数。

$\sigma=\begin{bmatrix} b_1 & b_2 \\ \vdots & \vdots \\ \vdots & \vdots \\ b_7 & b_8 \end{bmatrix}$，对于 b_j 有 $b_j=\left(\left(rand\left(0,1\right)*1\right)/4\right)+2,\ j=1,\cdots,8$。

此为基于局部的 NJW 算法的结果，使用局部处理的结果在人工数据集上(图 2)与全局处理并无明显差异。但是在局部算法中并不能彻底发挥 NJW 算法的稳定性优势。

4.2. 评价标准

由于三支聚类的特殊性，需要将原本二支聚类评价指标进行修改，使其可以对三支聚类的结果进行正确评价，本指标利用 *S_Dbw*(*c*) [6]指标的衡量方法，分别对核心域，边界域进行评价，将两者结果与原本的离散程度进行对比。

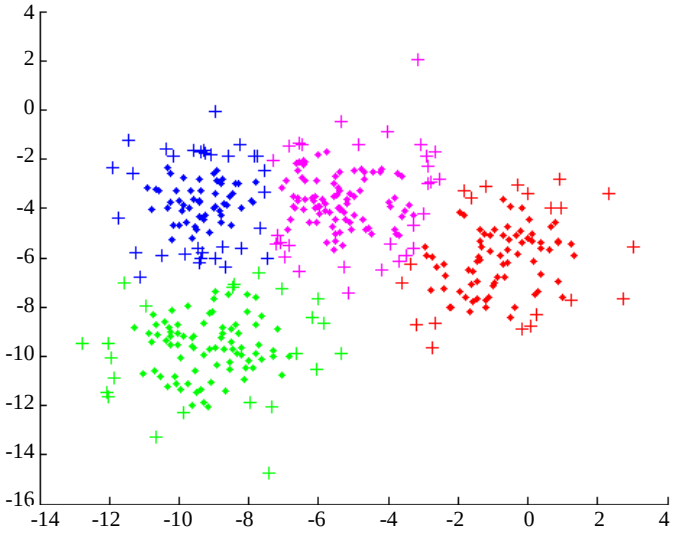

Figure 2. A synthetic data set with 400 points
图 2. 包含 400 个数据的人工数据集

$$Scat(c)=\frac{1}{c}\sum_{i=1}^{c}\frac{\|\sigma(v_i)\|}{\|\sigma(U)\|} \tag{4}$$

其中 c 为二支聚类结果中的类簇个数，S 为数据集。$\sigma(U)$ 的方差，其 P 阶距定义如下：

$$\sigma_x^P=\frac{1}{n}\sum_{k=1}^{n}\left(X_k^P-\bar{X}^P\right)^2 \tag{5}$$

$$\bar{X}=\frac{1}{n}\sum_{k=1}^{n}X_k,\ \forall X_k\in U \tag{6}$$

$\sigma(v_i)$ 是类簇 C_i 的方差：

$$\sigma_{v_i}^P=\frac{1}{n_i}\sum_{k=1}^{n_i}\left(X_k^P-v_i^P\right)^2 \tag{7}$$

对 $\|x\|$ 的定义为：如果 x 是一个向量，则

$$\|x\|=\sqrt{x*x'} \tag{8}$$

记 $P=\{c_1^P,c_2^P,\cdots\}$，$B=\{c_1^B,c_2^B,\cdots\}$，$C=\{c_1,c_2,\cdots\}$，其中 P 为每个已划分类簇核心域构成的集合，B 为每个已划分类簇边界域构成的集合，C 为原二支聚类的结果。根据 $Scat(c)$指标，从理论上可以得出

$$Scat(c^P)=\frac{1}{c^P}\sum_{i=1}^{c^P}\frac{\|\sigma(v_i)\|}{\|\sigma(P)\|},$$

$$Scat(c)=\frac{1}{c}\sum_{i=1}^{c}\frac{\|\sigma(v_i)\|}{\|\sigma(U)\|},$$

$$Scat(c^B)=\frac{1}{c^B}\sum_{i=1}^{c^B}\frac{\|\sigma(v_i)\|}{\|\sigma(B)\|},$$

$$Scat(c^P)<Scat(c)<Scat(c^B). \tag{9}$$

Table 1. UCI data sets used in experiments
表 1. UCI 数据集

数据集	样本数	样本维数	类别
Pi	768	8	2
EN	336	7	8
CV	435	16	2
IN	150	4	3
DRD	1151	19	2

Table 2. Results of experiments
表 2. 实验结果

数据集	$Scat(c^P)$	$Scat(c)$	$Scat(c^B)$
Pi	0.6145	0.7973	1.5178
EN	0.1930	0.3835	1.3446
CV	0.7028	0.7705	1.2564
IN	0.1182	0.1812	0.3054
DRD	0.2656	0.4302	1.1432

4.3. UCI 数据集测试

本节使用 5 组 UCI 数据集 Pima (Pi)，Ecoli_Nor(EN)，Congressional Voting(CV)，Iris_Nor(IN)，Diabetic Retinopathy Debrecen (DRD)对所算法进行测试。

数据介绍如表 1。

实验结果如表 2。

从实验结果来看，符合理论预期结果。该算法较好的解决了三支聚类分析中边界域和核心域的划分问题。

5. 总结

通过对每个元素进行加权处理获得个体对全体的影响，确定对象所属区域。思想来源于物理中质心的概念。最后使用 $S_Dbw(c)$进行离散度判断，确定算法的有效性。下一步工作将处理扰动系数，标准扰动量等参数具体的关系，从定性转化为定量分析。更换判定方式，诸如密度等，从而克服维数过高导致的结果不可知。

基金项目

本文受国家自然科学基金(61503160，61572242)，江苏省高校自然科学基金(15KJB110004)，江苏科技大学本科生创新计划项目资助。

参考文献 (References)

[1] Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) Data Clustering: A Review. *ACM Computing Survey*, **31**, 264-323. https://doi.org/10.1145/331499.331504

[2] Anderberg, M.R. (1973) Cluster Analysis for Application. Academic Press, New York.

[3] Omran, M.G.H., Engelbrecht, A.P. and Salman, A. (2007) An Overview of Clustering Methods. *Intelligent Data Analysis*, **11**, 583-605.

[4] Xu, R. and Wunsch, D. (2005) Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, **16**, 645- 678. https://doi.org/10.1109/TNN.2005.845141

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

[6] Halkidi, M. and Vazirgiannis, M. (2002) Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set. *Proceedings IEEE International Conference on Data Mining*, 29 November-2 December 2001. http://ieeexplore.ieee.org/abstract/document/989517/