Computer Science and Application 计算机科学与应用, 2019, 9(12), 2349-2356
Published Online December 2019 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2019.912261

Hans汉斯

Three-Way Clustering Analysis Based on Voting Theory

Liuquan Gu, Ruilin Chai, Pingxin Wang*

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

Received: Dec. 2nd, 2019; accepted: Dec. 13th, 2019; published: Dec. 20th, 2019

Abstract

At present, most of existed clustering methods are two-way clustering which are based on the assumption that a cluster must be represented by a set with crisp boundary. However, assigning uncertain objects into a certain cluster will reduce the accuracy of clustering results. Three-way clustering is an overlapping clustering which describes each cluster by core region and fringe region. It handles the category problem of uncertain objects and reduces the decision risk effectively. This paper mainly introduces a model of three-way clustering, and gives three-way clustering algorithm based on k-means as an example for analysis. Firstly, different clustering results of the same data set are obtained by ensemble clustering. Then, the label matching method is used. Finally, the cluster of objects is determined according to the voting rules. Through the analysis of experimental results, it is verified that the effect of the clustering method has been significantly improved.

Keywords

K-Means, Three-Way Clustering, Label Matching, Voting Rules

基于投票理论的三支聚类分析

谷留全，柴瑞林，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：2019年12月2日；录用日期：2019年12月13日；发布日期：2019年12月20日

摘　要

目前，大多数聚类方法是二支聚类，即每个对象要么属于一个类，要么不属于一个类，聚类结果具有清晰的边界。然而，将某些不确定的对象强制分配到某个类簇中会降低聚类结果精度。而三支聚类是一种

*通讯作者。

文章引用: 谷留全, 柴瑞林, 王平心. 基于投票理论的三支聚类分析[J]. 计算机科学与应用, 2019, 9(12): 2349-2356.
DOI: 10.12677/csa.2019.912261

重叠聚类方法，它采用核心域和边界域来表示每个类簇，较好地处理了含有不确定性信息对象的类别归属问题，同时有效地降低了决策风险。本文主要介绍一种三支决策聚类模型，并给出基于k-means的三支决策聚类算法作为实例进行分析。首先，通过已有的聚类算法得到相同数据集的不同二支决策聚类结果，然后对聚类成员的类簇标签进行匹配，最后制定投票规则确定样本的类簇归属。通过实验结果分析，本文所提出的聚类方法的聚类结果在各聚类性能度量指标上有了明显的提升。

关键词

K-Means，三支聚类，标签匹配，投票规则

1. 引言

聚类是一种无监督的学习方法，是机器学习、数据挖掘、模式识别和图像分析等领域中一个极具挑战性的研究方向。聚类反映了在不同层面做出相应决策的过程，即聚类是在一定粒度层[1]面确定样本对象属于或者不属于某一类簇的过程。聚类分析的一个广泛潜在假设是一个类簇可以用单一的集合来表示，或者说类间的边界是确定的、清晰的。虽然在聚类分析中有许多聚类算法且都有自己特有的发现潜在数据结构的方法，但是不同的聚类算法可能会有不同的聚类结果。因此，在没有任何监督信息的情况下，很难判断出哪一种聚类结果适合当前的数据集。而集成聚类很好地解决了该问题，聚类集成主要包含两个过程：生成过程与一致性划分，本文方法主要是解决已生成的聚类结果的一致性划分问题。

为了解决传统聚类方法存在的问题以及三支决策理论的提出，许多新的决策聚类方法被提出。Lingras和Yan [2]认为使用区间集可以更好地表示聚类结果。在此基础上本文提出的三支聚类算法是在二支聚类结果的基础上进行收缩和扩张分别得到三支聚类结果的核心域和边界域。该算法的聚类结果相比经典的聚类算法有较大的改进。

本文针对二支决策聚类结果如何转换为三支决策聚类结果的问题，提出了基于集成聚类框架并采用投票方式的一种解决方案。具体来说就是：在二支集成聚类的过程中，首先对聚类结果中所有类簇进行交集运算，交集部分的样本对象可以明确其归属类簇，即将这些样本都划分在同一类簇。然后，将这部分对象划分到相应标签对应类簇的核心域。最后，根据投票规则确定剩余样本对象的所属类簇。根据本文提出的基于投票的三支决策聚类方法，得到类簇的核心域与边界域确定全部数据对象的类簇归属，最终将二支决策聚类结果转换为一个三支决策聚类结果。

2. 相关工作

2.1. 三支聚类

人们通常根据事物已有的信息做出相应的决策，然而，信息的获取通常是一个动态的过程。现实生活中，由于样本对象信息不充分且不能确切地确定其类别归属，则需要其他方法来处理含有不确定性特征的聚类任务即三支决策聚类。对于信息充分的对象可以做出确切的决策，而对于信息不够充分的对象则需要进一步获取相关信息后作出决策，这是一种典型的三支决策思想[3] [4]。

三支决策是姚一豫教授[5] [6] [7]于2010年提出的一种基于粗糙集和决策粗糙集知识的理论。三支决

策理论是对传统二支决策理论的推广和完善，其核心思想是将粗糙集中的正域、负域和边界域对应决策中的接受、拒绝和延迟。三支决策更符合人们在实际生活中作决策时的思维模式，现已受到国内外学者的广泛关注。

在三支决策理论的基础上，Yu [8]将聚类思想与三支决策理论相结合提出了三支决策聚类方法。二支聚类即硬聚类方法的聚类结果是由单一的集合表示某个类簇，而三支聚类结果是使用嵌套集合来表示某个类簇。样本对象和类簇之间存在三种关系即：对象确定属于这个类；对象确定不属于这个类；对象可能属于这个类。Yu, Liu 等人[9] [10] [11]将三支聚类的三个区域给出确切定义为：采用区间集表示一个类，核心域为区间集的下界，边界域由区间集的上下界之间的对象组成，琐碎域为区间集的上界的补集。核心域的对象确定属于这个类，边界域的对象不一定属于这个类，而琐碎域的对象一定不属于这个类。

关于三支聚类[12]的描述如下：给定样本数据集 $U=\{x_1,x_2,\cdots,x_n\}$，其中 n 表示样本对象个数，令 $M=\{m_1,m_2,\cdots,m_k\}$，其中 k 表示聚类数，即 n 个对象可以被划分为 k 类。用 $C(m_i)(i=1,2,\cdots,k)$ 表示第 i 类的核心域；$F(m_i)(i=1,2,\cdots,k)$ 表示第 i 类的边界域：$T(m_i)(i=1,2,\cdots,k)$ 表示第 i 类的琐碎域。相关性质有：

$$C(m_i)\cup F(m_i)\cup T(m_i)=U\ \ (i=1,2,\cdots,k) \tag{1}$$

式(1)表示任意类簇的核心域、边界域和琐碎域的并集为全集。

$$\bigcup_{i=1}^{k}\left(C(m_i)\cup F(m_i)\right)=U\ \ (i=1,2,\cdots,k) \tag{2}$$

式(2)表示所有类簇的核心域、边界域的并集为全集。

$$C(m_i)\neq\phi\ \ (i=1,2,\cdots,k) \tag{3}$$

式(3)表示任意类簇的核心域非空。

2.2. 匹配原则

聚类分析是一种无监督学习方法，简单地说就是把相似的样本划分到同一组。聚类的目标是使得类内样本相似度高而类间样本相似度低。因此，一个聚类算法通常只需要知道如何计算相似度就可以开始工作了。聚类通常并不需要使用训练对象进行学习，但是，在聚类过程中，每次给出不同的聚类规则就会得出一个与之相对应的聚类结果。聚类结果中的每个类簇，都可以用符号来标记，如 r、s、t 等，这些符号文本称之为标签，其实也就是每个类的类号。不同于分类，在研究过程中发现这些标签，往往不能相同，也就是说类簇标签不唯一，不同的标签却表示同样的结果，这就造成了对结果分析的困扰。研究人员在此情况下，优化多种聚类方式，其中比较具有代表性的聚类方法，例如 k-means [13]聚类、均值漂移聚类、基于密度的聚类方法、凝聚层次聚类等，却始终改变不了无监督学习的本质，没有唯一与之对应的类簇标签，意味着无法对结果进行精准的分类与识别。

由于是无监督学习的特点，类簇标签就不再对聚类结果具有代表性，因此需要更加准确的类簇标签，这个标签不再是聚类结果中随机分配给的 r、s、t 等，而是通过匹配原则，利用记录匹配重新定义的 $P_1,P_2,\cdots,P_n$ 等，这种类簇标签的特点是唯一对应的，若 $P_i\neq P_j$，类簇也不相同，反过来，类簇不相同，一定可以推出 $P_i\neq P_j$。

首先对数据集进行多次聚类，得到多组聚类结果；在每次聚类结果中，都会随机给其分配类簇标签 r、s、t 等，由于数据量大，这里选用 1、2、3 等代替，方便数据展示。实质上，类簇标签 1 只是代表某类簇在第一次聚类结果，是随机排序的第一位，第二次聚类结果中，某类簇可能就不是随机排序的第一位，可能是在任何一个位次，从而不再是类簇标签 1，改为类簇标签 2 或者其他，但是无论给予何种的

类簇标签，所代表的的结果仍是原来的类簇，换句话说，类簇里面的类不变，如果可以为它起名，赋予独一无二的代号，就能清楚的分辨结果。那么这个名字或者代号就是我们要寻找的 P_i，给予一个唯一对应的新的类簇标签，使其在多次聚类结果中，仍能代表自己，里面的类也不会改变。

下面将举例说明如何定义类簇标签 P_i。这里我们假设对 22 个数据聚成 3 类，聚类了三次，结果见图 1。

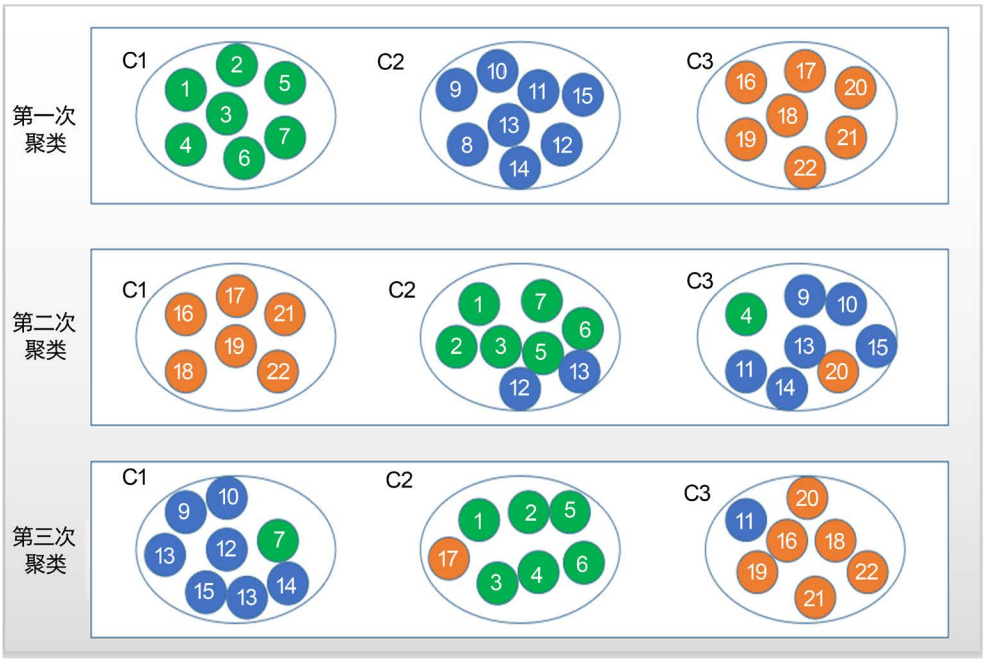

Figure 1. Schematic diagram of clustering results
图 1. 聚类结果示意图

从图中我们可以清楚的看到第一次聚类的 C1、第二次聚类的 C2 和第三次聚类的 C2 应该就是不同聚类的同一类簇。因为他们之间所含共同元素相比于其他两个最多。有了这样的设想，我们就可以进行下面的表格统计。在 3 次聚类结果中，对于每个数据都有一个数据列，如数据 1 的数据列为“122”，代表第一次聚类在 C1，第二次聚类在 C2，第三次在 C2。以此类推，我们可以得出一系列的数据类，因为我们所处理的数据是局部集中的，所以必然会有一部分数据一直被聚类在同一类簇中，因此必然会出现重叠次数较多的数据列，并记录下来，如下列表 1 所示。对于每个类簇而言，就已经将重叠数据最多类簇匹配起来，我们将最多的的三个数据列类簇定义为 P_1，类似的定义 P_2、P_3 等多个类簇。

Table 1. Representation of clustering results
表 1. 聚类结果表示

	三次聚类的类别			出现的次数
P_1	1	2	2	5
P_2	2	3	1	6
P_3	3	1	3	5

到这里，就完成了匹配原则，并且定义了类簇标签 P_1、P_2、P_3 等，完成了一开始的任务，即找到一

个与类簇唯一对应的标签，使其在多次聚类结果中，仍能代表自己，里面的类也不会改变。匹配原则伪代码如下：

算法 1：匹配原则算法：

算法：计算每一种标签匹配重复的个数
输入：N 组聚类结果
输出：每一种标签匹配重复的个数
步骤：
1. 将 N 组聚类结果作为纵列的矩阵组成 N 列的纵向矩阵；
2. 建立类矩阵，并将第一行放入类矩阵，并在最后一列记录重复个数为 1，类数标位 1；
3. for I = 2:k
{
4. 每一行的行矩阵与类矩阵中每一行进行对比。如果相同，那么，将类矩阵中相同的记录个数+1，并调整这一行的位置，使最大重复个数的在最后。如果不与类矩阵重复的，则将这一行放在第一行，并记录重复个数为 1；
5. 在做第 4 步的同时，为新出现的行标号，号码数为上一次新出现的类数+1；并且为也给原始的 50 列的聚类结果加上编号，变号为对应的类矩中对应对象的编号；
}
6. 输出结果。

2.3. 投票原则

根据上面的工作，我们已经将每一次聚类都进行的匹配，如上述的例子中，第一次聚类中 C_5 (标签为 5 的类簇)与第二次聚类中 C_1 与第三次的 C_2 和之后的其他类簇匹配，也就是说，虽然他们每次聚类后的标签名字都不一样，但是他们其实代表着不同聚类中同一个类。

有了这样一个概念后，我们会发现，有一些样本永远在同一个类，我们既可以基本肯定这个数据一定是属于这个类。而对于剩余的数据，由于聚类算法存在不稳定性，我们给了一定的允许误差。由此，我们定义了一个投票机制。

首先，我们先给了一个隶属度的概念。函数定义如下：

$$f\left(x_i\right)=\frac{m_j}{N} \tag{4}$$

其中，x_i 表示第 i 个样本，N 表示总共聚类的次数，m_j 表示 N 次中属于 C_j 类簇的次数。

然后，计算每一数据和每一类簇的隶属度，我们规定函数：

$$g\left(x_i\right)=\begin{cases} x_i \in C\left(m_j\right), & f\left(x_i\right)\geq 0.9 \\ x_i \in F\left(m_j\right), & 0<f\left(x_i\right)<0.9 \\ x_i \notin C_j, & f\left(x_i\right)=0 \end{cases} \tag{5}$$

最后，根据 $g\left(x_i\right)$ 函数，我们就可以确定每一个样本对象的类簇归属。

3. 聚类结果评价指标

根据是否利用真实样本的分布信息，聚类有效性评价指标分为外部指标(即通过比较聚类结果与真实分布的匹配程度对聚类结果进行评价)和内部指标(即通过评价聚类结果优劣来发现数据集内部结果和分布状态)。本文采用 ACC [14]和 DBI [15]两种评价指标。

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。这个方法就是根据预测的结果与真实值做对比，当此值越高说明聚类结果越好。

$$\mathrm{ACC}=\frac{1}{N}\sum_{i=1}^{k}M_i \tag{6}$$

其中 N 表示所有已经被确定类别的对象的个数，M_i 表示正确划分到第 i 个类的数据对象个数，k 表示聚类数。本文利用核心域的对象来计算三支聚类结果的 ACC 值。

3.2. Davies-Bouldin Index

Davies-Bouldin index 通过计算每个类簇最大相似度的均值，是一种评估聚类算法优劣的内部聚类评价指标。使用下列公式进行计算

$$\mathrm{DBI}=\frac{1}{N}\sum_{i=1}^{N}\max_{i\neq j}\left(\frac{\overline{S_i}+\overline{S_j}}{\left\|w_i-w_j\right\|_2}\right) \tag{7}$$

其中 $\left\|w_i-w_j\right\|_2$ 为簇类 i 与簇类 j 的距离，$\overline{S_i}$ 计算的是类内样本到簇质心的平均距离，计算公式为：

$$\overline{S_i}=\left(\frac{1}{T_i}\sum_{j=1}^{T_i}\left|x_j-A_j\right|^p\right)^{\frac{1}{p}} \tag{8}$$

x_j 代表类簇 i 中第 j 个样本点，A_i 是类簇 i 的质心，T_i 是类簇 i 中样本的个数，p 在通常情况下取值为 2。

4. 实验数据分析

为了验证算法的有效性，本文选用五组标准的 UCI 数据集，如表 2 所示。

Table 2. Data used in experiments
表 2. 实验中用到的数据

数据集	样本个数	样本维度	簇类
Wine	178	13	3
Blood	748	4	2
Seeds	210	7	3
Thyroid	215	5	3
Heartstatlog	270	13	2

首先通过 k-means 算法分别对这五组标准的 UCI 数据集进行 50 次二支聚类，其次将这 50 次的结果进行类匹配，然后以隶属度为 0.9 为界，把样本依次划分为相应类簇的核心域成员和边界域成员。最后通过聚类结果的评价准则 ACC、DBI、分别对二支聚类和三支聚类的结果进行评价。观察表 3，我们可以看出在原始 k-means 算法的基础上，进行三支聚类，可以有效提高算法的准确度，降低聚簇间的相似度。

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

数据集	聚类算法	ACC	DBI
Wine	k-means	0.702	0.777
	three-way k-means	0.704	0.369
Blood	k-means	0.723	0.579
	three-way k-means	0.743	0.436
Seeds	k-means	0.891	0.754
	three-way k-means	0.902	0.723
Thyroid	k-means	0.861	0.975
	three-way k-means	0.928	0.642
Heartstatlog	k-means	0.589	0.988
	three-way k-means	0.592	0.975

5. 结语

本文基于 k-means 算法对数据集进行多次聚类，然后利用标签匹配、投票规则等方式研究三支决策聚类方法。该方法更适用于存在不同差异但仍然有共同特点的数据，采用了类簇标签匹配求交集的方式对一部分数据进行归类和总结分析。相对于其他算法例如，均值偏移聚类算法、DBSCAN 聚类算法、GMMs、层次聚类算法，本文提出的聚类集成算法是有效的，实验结果表明该算法精确度得到了提高。但仍然存在的缺点有：聚类集成单一使用 k-means 算法，随机选取 k 个初始类聚类质心，然后对聚类中心进行迭代直到满足结束条件。因为随机选取聚类中心存在不确定因素，故聚类结果会存在差异。未来需要改进并提高的方面是：本文所提出的三支决策聚类模型有待进一步优化，且可以不断补充并尝试多种不同的聚类方法，最终能提高三支决策聚类的聚类精度从而获得较好的聚类结果。

基金项目

国家自然科学基金(Nos. 61503160)；江苏省高校自然科学基金(No. 15KJB110004)。

参考文献

[1] Li, J.H., Huang, C.C., Qi, J.J., Qia, Y.H. and Liu, W.Q. (2017) Three-Way Cognitive Concept Learning via Multi-Granularity. *Information Sciences*, **378**, 244-263. https://doi.org/10.1016/j.ins.2016.04.051

[2] Lingras, P. and Yan, R. (2004) Interval Clustering Using Fuzzy and Rough Set Theory. *IEEE Annual Meeting of the Fuzzy Information*, Banff, 27-30 June 2004, 780-784. https://doi.org/10.1109/NAFIPS.2004.1337401

[3] Liu, D., Yao, Y.Y. and Li, T.R. (2011) Three-Way Investment Decisions with Decision-Theoretic Rough Sets. *International Journal of Computational Intelligence Systems*, **4**, 66-74. https://doi.org/10.1080/18756891.2011.9727764

[4] Li, Y., Zhang, C. and Swanb, J.R. (2000) An Information Filtering on the Web and Its Application in Job Agent. *Knowledge-Based Systems*, **13**, 285-296. https://doi.org/10.1016/S0950-7051(00)00088-5

[5] Yao, Y.Y. (2010) Three-Way Decisions with Probabilistic Rough Sets. *Information Sciences*, **180**, 341-353. https://doi.org/10.1016/j.ins.2009.09.021

[6] 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

[7] Yao, Y.Y. (2012) An Outline of a Theory of Three-Way Decisions. In: Yao, J., Yang, Y., *et al*., Eds., *Rough Sets and Current Trends in Computing*, Springer, Berlin, Heidelberg, Vol. 7413, 1-17. https://doi.org/10.1007/978-3-642-32115-3_1

[8] Yu, H., Chu, S.S. and Yang, D.C. (2010) Autonomous Knowledge-Oriented Clustering Using Decision-Theoretic

Rough Set Theory. *Rough Set and Knowledge Technology*, In: Yu, J., Greco, S., Lingras, P., Wang, G. and Skowron, A., Eds., Springer, Berlin, Heidelberg, Vol. 6401, 687-694. https://doi.org/10.1007/978-3-642-16248-0_93

[9] Yu, H., Liu, Z.G. and Wang, G.Y. (2014) An Automatic Method to Determine the Number of Clusters Using Decision-Theoretic Rough Set. *International Journal of Approximate Reasoning*, **55**, 101-115. https://doi.org/10.1016/j.ijar.2013.03.018

[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., 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

[12] 于洪. 三支聚类分析[J]. 数码设计, 2016, 5(1): 31-35.

[13] Macqueen, J. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1967, 281-297.

[14] Schölkopf, B., Platt, J. and Hofmann, T. (2006) A Local Learning Approach for Clustering. *International Conference on Neural Information Processing Systems*, MIT Press, 2007, 1529-1536.

[15] Davies, D.L. and Bouldin, D.W. (1979) A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **1**, 224-227. https://doi.org/10.1109/TPAMI.1979.4766909