文章编号：2095-2163(2021)04-0173-03 中图分类号：TP311 文献标志码：A

海平面聚类算法

马 杰[1]，杨 磊[2]，徐 建[1]
(1 江苏师范大学 智慧教育学院(计算机科学与技术学院)，江苏 徐州 221116；2 中国矿业大学徐海学院 计算机系，江苏 徐州 221008)

摘 要：本文主要研究海平面聚类算法，通过与 AP 算法以及 MD 算法的比较和结合，采用适当的密度函数解决边缘点和归类丢失点的问题，有效地改进了其算法的功能和聚类效果。

关键词：海平面聚类算法；AP 算法；MD 算法

Sea level clustering algorithm

MA Jie[1], YANG Lei[2], XU Jian[1]
(1 School of Education Intelligent Technology (School of Computer Science and Technology) Jiangsu Normal University Xuzhou Jiangsu 221116, China; 2 Department of Computer Science, Xuhai College, China University of Mining & Technology, Xuzhou Jiangsu 221008, China)

【Abstract】 This paper mainly studies the sea level clustering algorithm. Through the comparison and combination with AP algorithm and MD algorithm, the appropriate density function is used to solve the problem of edge points and missing points, so that the function and clustering effect of the algorithm can be effectively improved.

【Key words】 sea level clustering algorithm; AP algorithm; MD algorithm

0 引 言

本文算法不是一个独立的聚类算法，是用来辅助其它聚类算法更好、更有效地聚类的辅助算法。与其它聚类算法结合使用，能有效地改善聚类算法的聚类效果。

1 问题的提出

有些算法聚类的结果与自然分类有出入，有些算法对某些情况不能正确的分类。比如：Affinity Propagation(AP)聚类算法，是基于数据点间的“信息传递”的一种聚类算法。算法的基本思想是：将全部样本看作网络节点，通过网络中各条边的消息传递 计算出各样本的聚类中心。聚类过程中，共有两种消息在各节点间传递，分别是吸引度(responsibility)和归属度(availability)。通过在点之间不断地传递信息，最终选出代表元以完成聚类。AP 算法通过迭代过程不断更新每一个点的吸引度和归属度值，直到产生 m 个高质量的 Exemplar(类似于质心)，同时将其余的数据点分配到相应的聚类中。其特点如下：

(1)不需要制定最终聚类个数。

(2)将已有数据点作为最终的聚类中心，而不是新生成聚类中心。

(3)模型对数据的初始值不敏感，多次执行 AP 聚类算法，得到的结果是完全一样的，即不需要进行随机选取初值步骤。

(4)对初始相似度矩阵数据的对称性没有要求。

(5)与 k 中心聚类方法相比，其结果的平方差误差较小，相比于 K-means 算法，鲁棒性强、准确度较高，但算法复杂度高、运算消耗时间多。

在实际的使用中，AP 有两个重要参数：preference(定义聚类数量)和 damping factor(控制算法的收敛效果)。

聚类就是个不断迭代的过程，迭代的过程主要是更新两个矩阵：

吸引度矩阵 R：$[r(i,k)]N \times N$

归属度矩阵 A：$[a(i,k)]N \times N$

$$r(i,k) = s(i,k) - \max_{k' \neq k}(a(i,k') + s(i,k'))$$

$$a(i,k') = \begin{cases} \min\{0, r(k,k) + \sum_{i' \notin (i,k)} \max(0, r(i',k))\}, & i \neq k \\ \sum_{i' \neq k} \max(0, r(i',k)), & i = k \end{cases}$$

作者简介：马 杰(1979-)，女，硕士，讲师，主要研究方向：机器学习、计算机网络、算法优化；杨 磊(1979-)，男，博士，副教授，主要研究方向：机器学习、计算机网络、算法优化；徐 建(1961-)，男，学士，讲师，主要研究方向：人工智能。
通讯作者：杨 磊 Email：29164753@qq.com

收稿日期：2021-01-20

$$r(i,k) \leftarrow s(i,k) - \max_{k's.t.k' \neq k} \{a(i,k') + s(i,k')\}$$

$$a(i,k) \leftarrow \min\{0, r(k,k)\} + \sum_{i's.t.\ i' \notin \{i,k\}} \max\{0, r(i',k)\}\}$$

在不断交替更新 a 和 r 值,达到一定的次数或收敛后,选取使得 $r(i,k) + a(i,k)$ 最大的那个 k 作为 i 的代表元。其中 $s(i,k)$ 表示 similarity,可以翻译为相似度或度量。是指点 k 作为点 i 的聚类中心的相似度,一般使用欧氏距离来计算。相似度值越大说明点与点的距离越近,这在几乎所有的聚类分析中都是最基础的量。

AP 算法(参见参考文献[1])是一个很好的聚类算法。但当有大类靠近小类时,往往会把大类的一些边缘点错分给小类。如对图 1 中的数据,其 AP 算法的聚类结果如图 2 所示。显然,没有分成左边两个小类,右边一个大类。而是小类占了大类的几个点。另外,还有一些算法(本文选定一种密度算法,本文称 MD 算法)对有些数据不能正确的分开。如图 3 的数据,右边两类中间有两行点连接在一起,很多聚类算法就无法将这两类分开。

图 1 AP 算法数据准备

Fig. 1 AP algorithm data preparation

图 2 AP 算法分类结果

Fig. 2 AP algorithm classification results

2 问题分析

由上述分析可以看出,问题都出在类的边缘点上。AP 算法的问题是大类的几个边缘点离大类的中心点过远,而离靠近其小类中心点更近。另外一些算法无法将图 3 右边两个类分开,是因为靠近两个类的共同边缘的点连接在了一起。

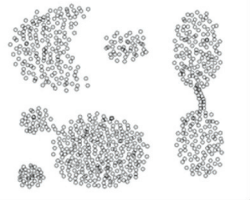

图 3 MD 算法数据准备

Fig. 3 MD algorithm data preparation

3 解决方法

如果能把这些出问题的边缘点先 0 拿掉(拿掉的点最后还要归类到分好的类中)再进行分类,就不会有上面的问题了。那么,一个问题是解决如何区分边缘点,其二是如何将拿掉的点归类。

首先,对每个点定义一个密度函数,使得类的边缘点的密度小,越靠近中心的点密度越大,这样就解决了第一个问题。再定义每个点的归属点为离此点最近的密度大于自己的点,这样第二个问题就解决了。判断边缘点时,不是直接用密度函数的密度值判断,而是用传导归属点数(既 A 点到其归属点 B 点,B 点再到其归属点 C 点,等等,一直传导下去所经历的点叫 A 点的传导归属点,这个过程叫传导归属。而传导归属数是所有能传导归属到此点的点的个数),传导归属点数越小越边缘,反之越中心。引进一个参数 k,传导归属数小于其则为边缘点。先剔除边缘点,然后根据某聚类算法聚类,最后将边缘点传导归属到已分好的类中。

本文定义密度函数:

(1)此点密度为,此点到所有点的距离的倒数之和。

(2) 数据的个数为 n, 每一点为其它各点打分。离此点最远的点得 1 分,次远点得 2 分,以此类推。最近的点得 $n - 1$ 分。定义每个点得密度为此点得分的总和。

第一个密度函数要求数据先要剔除相同的数据点。

4 应用效果

本算法与 AP 算法结合(以下密度函数均选择第一种),采用聚类图 1 的数据,选择 K 为 2,边缘点如图 4 所示(方形空心为边缘点),聚类结果如图 5 所示。本文算法与 MD 算法结合,采用聚类图 3 的数据,选择 k 为 3,边缘点如图 6 所示(圆形空心为边缘点),聚类结果如图 7 所示。

图 4 AP 算法结合海平面算法的边缘图

Fig. 4 Edge map of AP algorithm combined with sea level algorithm

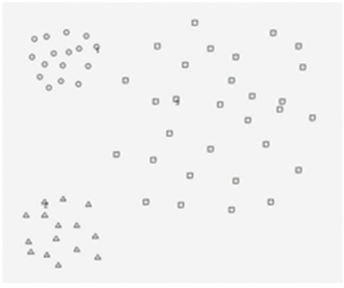

图 5　AP 算法结合海平面算法结果

Fig. 5　Results of AP algorithm combined with sea level algorithm

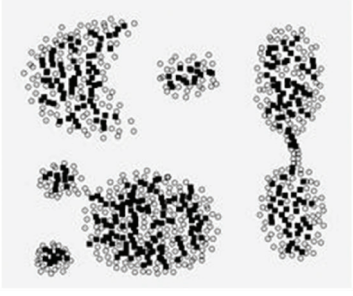

图 6　MD 算法结合海平面算法的边缘图

Fig. 6　Edge map of MD algorithm combined with sea level algorithm

图 7　MD 算法结合海平面算法的结果图

Fig. 7　Result chart of MD algorithm combined with sea level algorithm

5　结束语

本算法之所以叫海平面聚类算法，是因为 k 参数相当于设置海平面，边缘点都淹没在海水里，只对陆地进行聚类，因此得名。本算法与其它聚类算法结合可以明显改善聚类结果，经实验证明，本算法是有效的。

参考文献

[1] BRENDAN J. Frey and Delbert Dueck, Clustering by Passing Messages Between Data Points[J]. Science, 2017, 315(5814): 972-976.

.(上接第 172 页)

4　结束语

本文利用三种机器学习算法对铝及其合金晶粒尺寸进行预测。分析发现：(1)不同的机器学习模型用于晶粒尺寸的预测，其预测结果有较大差异，其中 RF 模型表现最佳，R^2 为 0.79，*RMSE* 为 7.04；(2)由于输入样本时是随机的，同一个机器学习模型中的预测结果 R^2 不稳定，不同的样本会导致模型的好坏，也可能是样本数量不足导致的。文中研究结果为进一步研究铝及其合金的晶粒尺寸提供有益的参考，有利于研究人员对晶粒细化的相关研究。

参考文献

[1] 米晓希，汤爱涛，朱雨晨，等. 机器学习技术在材料科学研究中的应用进展[J/OL]. 材料导报，2021(15)：1-18.

[2] 胡建军，曹卓，但雅波，等. 基于特征选择和机器学习的材料弹性性能预测[J]. 华南理工大学学报(自然科学版)，2019，47(5)：48-55.

[3] B H Z A, B H F A, B X H A, et al. Dramatically Enhanced Combination of Ultimate Tensile Strength and Electric Conductivity of Alloys via Machine Learning Screening[J]. Acta Materialia, 2020, 200：803-810.

[4] SHEN C, WANG C, WEI X, et al. Physical metallurgy-guided machine learning and artificial intelligent design of ultrahigh-strength stainless steel[J]. Acta Materialia, 2019, 179: 201-214.

[5] 王恩兆. 形核剂对铝合金晶粒细化极限行为的研究[D]. 济南：山东大学，2015.

[6] GENERAL CHAIR-KRISHNAPURAM B, GENERAL CHAIR-SHAH M, PROGRAM CHAIR-SMOLA A, et al. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]// Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM, 2016: 785-794.

[7] BREIMAN L. Random Forests[J]. Machine Learning, 2001.

[8] FREUND Y, SCHAPIRE R E. A Decision - Theoretic Generalization of On - Line Learning and an Application to Boosting. Journal of Computer and System Sciences, 1997, 55(1), 119-139.

[9] 晁代义，黄同璊，赫微，等. Al-Ti-B 细化剂对半连续铸造 7050 铝合金组织及性能的影响[J]. 特种铸造及有色合金，2020，40(11)：1256-1259.

[10] SWAMI A, JAIN R. Scikit-learn: Machine Learning in Python[J]. Journal of Machine Learning Research, 2013, 12(10): 2825-2830.