

文章编号: 2096-1618(2022)04-0396-05

结合模拟退火和多分配策略的密度峰值聚类算法

周俊, 蒋瑜, 马振明, 陈宏松
(成都信息工程大学软件工程学院, 四川 成都 610225)

摘要:针对密度峰值聚类算法在截断距离选取存在主观性依赖和非簇中心点的分配策略易出错的问题,提出一种结合模拟退火和多分配策略的密度峰值聚类算法(SA-DPC)。首先,利用模拟退火的启发式搜索找到全局最优的截断距离,设计以标准互信息(NMI)为目标函数的参数寻优模型;然后,从簇中心点开始以广度优先搜索的方式进行密度拓展;最后,找出锥形簇最近邻点依次分配。8个人工合成数据集的实验结果表明,改进的算法降低了聚类效果对截断距离的敏感性,且改进算法的ACC、ARI和AMI与原算法相比,分别最高提升了约35%、90%、80%。

关键词:密度峰值;启发式优化;广度优先搜索;密度拓展;簇最近邻

中图分类号:TP391

文献标志码:A

doi:10.16836/j.cnki.jcuit.2022.04.006

0 引言

在大数据时代,如何从大量低价值的数据中获得有益的信息一直是重点关注的问题。聚类分析是机器学习领域中一种重要的数据分析方法,旨在将数据按照一定的规则划分为不同的簇,使同簇的数据点相似性较高,簇间的数据集相似性较低^[1-2]。

Alex Rodriguez等^[3]提出了一种通过快速搜索和发现密度峰值的聚类算法(DPC)。该算法的优点是能够对任意非球形的数据集有较好的聚类效果,可描述数据分布,无需迭代即可得到聚类结果。

近年来,DPC算法在实践应用上取得了不错的进展,如在风能^[4]、交通安全^[5]、地震学^[6]等领域。但DPC算法仍存在问题,诸多学者针对其存在的不足进行了改进。如在截断距离选取困难方面,Chen等^[7]提出使用基于域密度的决策图来自动确定截断距离。在非簇中心点分配方面,Wang等^[8]提出从簇的角度搜寻未被分配的样本点。赵嘉等^[9]提出一种样本互相临近度概念,应用到非簇中心点的分配中。

从以上文献得到启发,本文通过研究模拟退火算法在截断距离寻优能力和非簇中心点的分配策略对聚类效果的影响,提出一种结合模拟退火算法和多分配策略的密度峰值聚类算法(SA-DPC)。首先利用模拟退火算法的启发式搜索能力寻找全局最优的截断距离,建立以NMI指标为目标函数的参数优化模型。然后提出非簇中心点的两步分配方式:密度拓展和簇近邻优先分配,前者将截断距离作为阈值,以广度优先搜索方式进行密度拓展;后者根据锥形簇的最近邻点依次进行分配。两种分配方式相得益彰。通过在多个人工合成数据集^[10]

的聚类结果表明,本文改进算法的评价指标比原算法有较大提升。

本文主要创新点:结合模拟退火算法自动选取全局最优截断距离,避免人为主观选取;提出密度拓展和簇最近邻点优先分配相结合的两步分配方式对非簇中心点进行分配,降低了聚类结果对截断距离选取的敏感性,提升了聚类效果。

1 SA-DPC 算法

1.1 截断距离启发式优化策略

图1给出DPC算法在Spiral数据集上取不同截断距离时的聚类结果,截断距离选取的范围仅从1.2%~1.7%,聚类结果却发生了较大变化。截断距离的选取对聚类效果具有至关重要的影响,因此对截断距离进行优化具有重要意义。

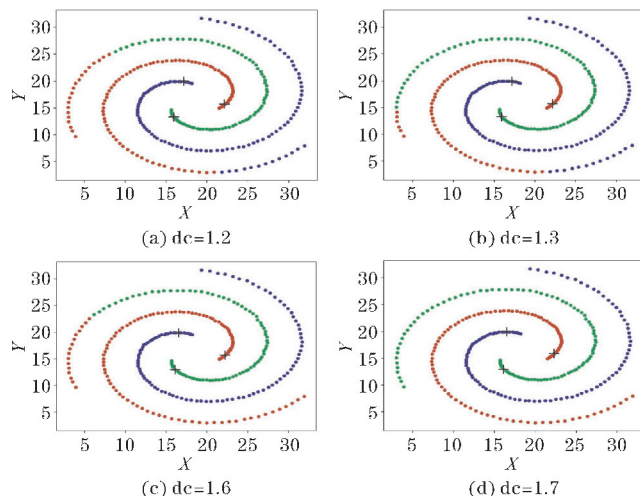


图1 不同dc在Spiral上的表现

模拟退火算法(simulated annealing algorithm, SA)是起源于自然界固体退火原理的一种启发式搜索算法,优点是能够有效跳出局部最优解进而找到全局最优解^[11]。

在启发式搜索过程中,建立以截断距离为自变量, NMI 指标为目标函数的参数优化模型。即建立截断距离和 NMI 指标的函数映射关系,给定一个截断距离值,就能得到一个唯一确定的 NMI 指标值, NMI 取值为 $[0,1]$,取值愈趋近 1,表明聚类效果愈好,截断距离的选取也愈佳。

经过 SA 算法的多次迭代搜索,找到能使 NMI 指标最大时对应的截断距离作为该数据集全局最优的截断距离,至此完成启发式截断距离优化过程。

1.2 两步分配策略

DPC 算法将非簇中心点分配给距其最近的密度峰所在簇,虽然分配策略相对简便,但极易出错。

图 2 是 DPC 算法在 Target 数据集上的聚类效果,该数据集分为 2 个类簇,中间靶心部分为一簇,其余靶心外圆环为一簇。黑色十字标记的样本点为类簇中心点。由图 2 可见,两个类簇中心点划分正确,靶心的样本点均分配正确。但外圆环部分则被分割成两个簇,外圆环仅右边一小部分样本点划分正确,其余部分因为到靶心簇中心点距离小于到外圆环簇的距离,从而被错误分配。

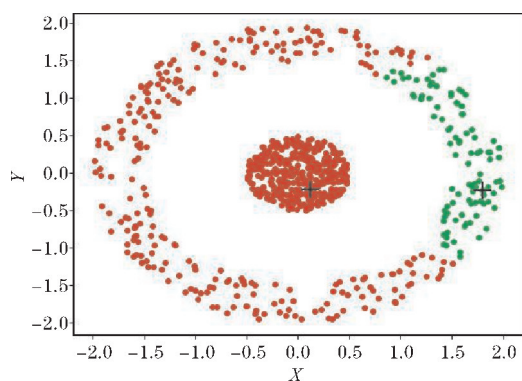


图 2 DPC 在 Target 上的表现

同样,不难发现图 1 中,在选取 4 种不同的截断距离时,类簇中心点皆分布在不同的类簇中,且位置偏移较小,但是聚类效果差异较大。这同样是 DPC 的分配方式存在缺陷造成的,在环形和流行数据集上表现特别明显。

不同的是,在图 1 中可以通过调整截断距离的取值,从而得到图 1(d)的正确聚类结果。而在 Target 数据集上,对截断距离调整均无法得到正确的聚类结果。这表明了 DPC 算法分配方式的缺陷有时可通过调整参

数来避免,但并不总是有效,有时也会带来聚类结果对截断距离选取敏感性强的问题。

从以上分析来看,对 DPC 算法分配策略的改进是必要的,因此给出以下 3 个与分配策略相关的概念。

密度拓展方法:簇中心为起点,以截断距离为阈值,找到满足样本点到簇中心的距离小于阈值的所有点,这些点均归属于当前簇;再以广度优先搜索方式搜索这些点并对满足阈值的点进行分配,重复该步骤直至没有符合条件的样本点出现。

锥形簇:经过密度拓展这一过程被分配的样本点的集合称为锥形簇。

本文所提出的初次分配策略,即锥形簇的形成算法具体描述如下。

算法 1 锥形簇形成算法

输入: data-样本点, CP-类簇中心点集合, dc-截断距离

输出: V-锥形簇集合, UV-未分配样本点集合

1. 从 CP 中选取一中心点 C_i , 并做类簇标记。
2. 建立一个空列表 V。
3. 将 data 中到 C_i 的距离小于 dc 的点加入 V 中, 标记为 C_i 同簇。
4. 取 V 中一点 V_i 。
5. 将 data 中未被标记且到 V_i 的样本点加入 V 中, 进行标记。
6. 直至 Visit 遍历完成, 完成该类簇锥形簇的形成; 否则转至 2 执行。
7. 直至 CP 遍历完成, 完成所有簇的锥形簇的形成, 输出 V 和 UV; 否则转至 1 执行。

簇最近邻点: 按照式 (1) 找到距离锥形簇边界最近的点, 这种点在再次分配过程中, 被优先进行分配。

$$CNP = \min_{j: p_j > p_i} (\text{dist}(m_i, n_j)) \quad (1)$$

其中, i 为未分配样本点的取值, j 为已分配样本点的取值。

再次分配策略是对锥形簇之外仍未被分配的样本点进行分配, 具体算法描述如下。

算法 2 再次分配策略算法

输入: V-锥形簇集合, U-未分配样本点集合

输出: Labels-聚类结果标签

1. 取 U 中一元素 U_i 。
2. 在 V 中找到 U_i 最近的样本点 V_i 。
3. 记录 U_i 所对应的距离值及样本点, 有新的最小距离值出现则替换掉当前距离值及样本点。
4. U 遍历完成, 当前记录的 U_i 即为 V_i 所在锥形簇的簇最近邻点, 将归为 V_i 所在簇。同时将 U_i 从集合 U 中去除。
5. 若 U 为空, 则再次分配算法完成, 输出 Labels; 否则, 转至 1 执行。

1.3 时间复杂度分析

DPC 算法的时间复杂度由距离矩阵、样本点局部密度和相对距离计算 3 个部分构成,各部分时间复杂度均为 $O(n^2)$,所以该算法的时间复杂度为 $O(n^2)$ 。

SA-DPC 算法的时间复杂度除了原算法 3 个部分外,还包括截断距离优化过程和两步分配策略 3 个部分。设模拟退火算法温度变化为 Δt ,迭代次数为 T ,截断距离与 NMI 指标模型维度为 1,所以本阶段时间复杂度为 $O(\Delta t \cdot T)$;设初次分配的样本点为 p ,密度拓展是以广度优先搜索方式进行,即 p 个样本点均被遍历一次,其中搜索满足阈值的样本点的复杂度为 $O(n)$,因此初次分配的时间复杂度为 $O(kpn)$;设再次分配的样本点个数为 $q=n-p$,对于每个未分配的样本点来说搜索最近且被分配的样本点的时间复杂度为 $O(q)$,所以再次分配过程的时间复杂度为 $O(pq)$ 。因此,SA-DPC 算法总的时间复杂度为

$$\begin{aligned} & (O(n^2)+O(kpn)+O(pq)) \cdot O(\Delta t \cdot T) \\ &= O(\Delta t \cdot T \cdot (n^2+kpn+pq)) \\ &< O(\Delta t \cdot T \cdot (n^2+kn^2+q^2)) \\ &< O(\Delta t \cdot T \cdot (n^2+kn^2+n^2)) \\ &= O(\Delta t \cdot T \cdot (k+2) \cdot n^2) \end{aligned}$$

其中, Δt 、 T 、 k 均为常数, $k < n$ (k 为类簇数), $p+q=n$,所以 SA-DPC 的时间复杂度量级为 $O(n^2)$,与 DPC 算法的时间复杂度量级相同。

2 仿真实验与结果分析

2.1 实验环境及数据集介绍

本文在 8 个人工合成数据集进行仿真实验,并与 3 种经典聚类算法 K-means^[12]、DBSCAN^[13] 和 AP^[14] 进行对比分析。其中,本文算法和 DPC 基于 python 语言实现,其他 3 种算法则通过调用 sklearn 的库函数实现。采用准确度指标 (ACC)、调整互信息 (AMI) 及调整兰德系数 (ARI) 3 个指标衡量聚类效果^[15]。3 个评价指标最大值皆为 1,指标值愈接近 1,表明聚类结果愈好。表 1 为本文仿真实验所选取的数据集的基本信息。

表 1 实验参数环境

数据集名称	样本规模/维度	类簇数	来源
Flame	240/2	2	人工数据集
Target	758/2	2	人工数据集
Pathbased	300/5	3	人工数据集
Compound	399/2	6	人工数据集
Spiral	312/2	3	人工数据集
Unbalanced	6500/2	7	人工数据集
Aggregation	788/2	7	人工数据集

2.2 分配策略对比

从图 3 SA-DPC 算法在 Target 数据集上的聚类效果看,SA-DPC 算法将样本点进行了正确的聚类。将图 3 和图 2 对比,两种算法在类簇中心点相同的情况下,SA-DPC 算法修正了 DPC 算法对样本点的错误分配,使非簇中心点聚类正确。

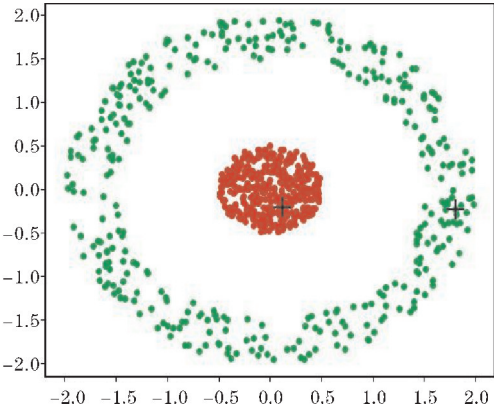


图 3 SA-DPC 算法在 Target 上的聚类效果

图 4 给出了 DPC 算法和 SA-DPC 算法在 Flame 数据集上截断距离同为 1.6 时的聚类效果。图 4(a) 在类簇中心点选择正确时,由于原分配方式的缺陷导致下边类簇的两侧被分配到上边的类簇中。图 4(b) 数据集被划分为上下两个锥形簇,这是分别从两个类簇中心进行密度拓展后得到的结果。初次分配后仍余 13 个黑色标记的样本点未被分配,原因是这些点距离锥形簇的最小距离小于截断距离阈值。经过再次分配,最后得到图 4(c) 的正确聚类结果。

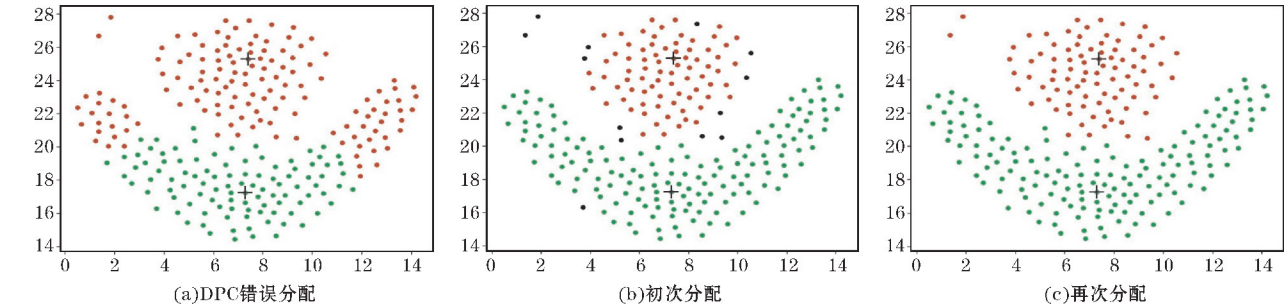


图 4 SA-DPC 与 DPC 的分配策略对比

图 5 是 DPC 和 SA-DPC 在 Spiral 数据集选取不同 dc 值所对应的 NMI 指标值变化的情况。DPC 算法的 NMI 指标与 dc 成正相关上升,当 dc 为 1.7, NMI 不再变化。而 SA-DPC 在 dc 取值不断变化的情况下, NMI 指标一直处于最高值 1, 表明改进后的分配策略, 能较好地降低聚类结果对 dc 取值的敏感性。

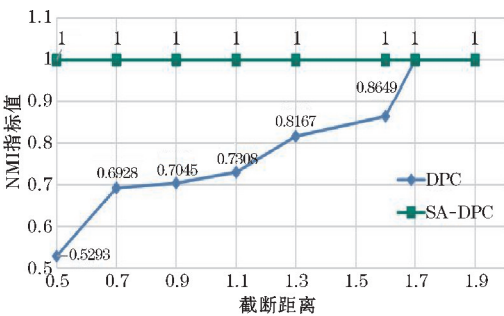


图 5 dc 对两种算法的影响

2.3 人工合成数据集实验结果

选用 4 种算法与本文算法在表 1 给出的数据集上进行仿真实验, 然后对聚类结果进行比较分析。表 2 给出了 5 种算法在 8 种人工数据集上详细的 3 种聚类评价指标数据, 其中加粗标注的数据为在同一数据集下 5 种算法中的最优值。

从算法角度来看, K-means 算法在 Unbalanced、R15 和 Aggregation 这 3 个非凸形数据集上表现较好, 对于环形和流行数据集表现较差。DBSCAN 算法在除了 Unbalanced 和 Pathbased 之外的数据集上均表现良好, 表明该算法能够对任意形状的数据集进行良好聚类, 但在处理类簇边界问题上较敏感导致聚类效果差。AP 算法在 R15 数据集上表现最好, 3 种评价指标值均在 99% 以上。该算法能够自动发现类簇个数, 无需提前指定。整体上看, SA-DPC 算法在人工合成数据集上的聚类效果要优于其他 4 种算法。

在 Spiral 和 Unbalanced 数据集上, SA-DPC 算法和 DPC 算法均能达到最优聚类效果; 在 Aggregation 数据集上 DPC 算法聚类效果较好, ACC 和 ARI 指标达到 0.91 以上; 在 R15 数据集上, K-means 算法和 AP 算法聚类效果最佳, 3 个指标值均在 0.99 以上, SA-DPC 算法和 DPC 算法较之略低, 其差值在 0.01 以内。综合评价指标数据, 则 DPC 算法在 Target(靶心)数据集表现最差, 仿真结果表明该算法在环形与球形结合的数据集上, 虽然能够正确识别到聚类中心, 但由于分配策略的缺陷导致最终聚类效果差。本文改进算法能够很好克服这一缺点, 改进算法的 ACC、ARI 和 AMI 与原算法相比分别最高提升了约 35%、90%、80%。

表 2 5 种算法在人工数据集的评价指标比较

Algorithm	Flame			Spiral			Unbalanced			Target		
	ACC	ARI	AMI	ACC	ARI	AMI	ACC	ARI	AMI	ACC	ARI	AMI
SA-DPC	1	1	1	1	1	1	1	1	1	1	1	1
DPC	1	1	1	1	1	1	1	1	1	0.653	0.092	0.199
K-means	0.874	0.453	0.397	0.343	-0.006	-0.006	1	1	1	0.714	0.182	0.282
DBSCAN	0.95	0.918	0.856	1	1	1	0.304	-0.002	0.006	1	1	1
AP	0.608	0.428	0.548	0.295	0.118	0.225	0.61	0.523	0.413	0.541	0.446	0.703

Algorithm	Pathbased			Compound			Aggregation			R15		
	ACC	ARI	AMI	ACC	ARI	AMI	ACC	ARI	AMI	ACC	ARI	AMI
SA-DPC	0.823	0.613	0.729	0.872	0.853	0.911	0.827	0.808	0.885	0.993	0.986	0.989
DPC	0.552	0.733	0.453	0.768	0.797	0.764	0.911	0.997	0.714	0.994	0.997	0.993
K-means	0.743	0.462	0.543	0.669	0.58	0.755	0.784	0.761	0.875	0.997	0.993	0.994
DBSCAN	0.783	0.637	0.684	0.97	0.964	0.932	0.827	0.809	0.888	0.92	0.912	0.96
AP	0.34	0.287	0.505	0.541	0.446	0.703	0.414	0.367	0.733	0.997	0.993	0.994

3 结束语

通过研究密度峰值聚类算法截断距离选取和非簇中心样本点的分配策略对聚类结果的影响, 提出了一种结合模拟退火和多分配策略的密度峰聚类算法。在

人工合成数据集上的仿真实验结果表明, 改进后的 SA-DPC 算法符合预期, 聚类指标基本上均优于 DPC 算法, 且能够对截断距离进行优化选取, 两步分配策略保证了非簇中心点的正确分配, 同时可降低聚类结果对参数的敏感性, 能够达到较好的聚类效果。如何实现 DPC 算法类簇个数的自动选取, 提高其在高维数据

集上的聚类效果将是下一步研究的重点方向。

参考文献:

- [1] 孙吉贵,刘杰,赵连宇. 聚类算法研究[J]. 软件学报,2008(1):48-61.
- [2] 纪守领,李进锋,杜天宇,等. 机器学习模型可解释性方法、应用与安全研究综述[J]. 计算机研究与发展,2019,56(10):2071-2096.
- [3] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191):1492-1496.
- [4] SHI H, YAN J, DING M, et al. An improved fuzzy c-means soft clustering based on density peak for wind power forecasting data processing[C]. Asia Energy and Electrical Engineering Symposium (AEEES), 2020.
- [5] 刘继新,董欣放,徐晨,等. 基于密度峰值的终端区航迹聚类与异常识别[J]. 交通运输工程学报, 2021, 21(5):214-226.
- [6] Vijay R K, Nanda S J. Seismicity analysis using space-time density peak clustering method[J]. Pattern Analysis and Applications, 2021, 24:181-201.
- [7] Chen J, Yu P S. A Domain Adaptive Density Clustering Algorithm for Data With Varying Density Distribution[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, (33)6:2310-2321.
- [8] Wang Y, Yang Y. Relative density-based clustering algorithm for identifying diverse density clusters effectively[J]. Neural Computing and Applications, 2021, 33:10141-10157.
- [9] 赵嘉,姚占峰,吕莉,等. 基于相互邻近度的密度峰值聚类算法[J]. 控制与决策, 2021, 36(3):543-552.
- [10] Fränti P, Sieranoja S. K-means properties on six clustering benchmark datasets[J]. Applied Intelligence, 2018, 48(12):4743-4759.
- [11] Gelatt M P, Vecchi S, Kirkpatrick C D. Optimization by Simulated Annealing[J]. Science, 1983, 220(4598):671-680.
- [12] McQueen J. Some methods for classification and analysis of multivariate observations[R]. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. Los Angeles: University of California, 1967:281-297.
- [13] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[R]. Proceedings of ACM SIGKDD'96, Portland, 1996:226-231.
- [14] Frey B J, Dueck D. Clustering by passing messages between data points[J]. Science, 2007, 315:972-976.
- [15] 杨燕,靳蕃, MOHAMED K. 聚类有效性评价综述[J]. 计算机应用研究, 2008(6):1630-1632.

Density Peak Clustering Algorithm Combining Simulated Annealing and Multiple Allocation Strategies

ZHOU Jun, JIANG Yu, MA Zhenming, CHEN Hongsong

(College of Software Engineering, Chengdu University of Information Technology, Chengdu 610225, China)

Abstract: In view of the subjective dependence of the truncation distance selection in the density peak clustering algorithm and the error-prone problem of the allocation strategy of non-cluster center points, a density peak clustering algorithm combining simulated annealing and multiple allocation strategies (SA-DPC) is proposed. Firstly, the heuristic search of simulated annealing is used to find the global optimal truncation distance, and a parameter optimization model with standard mutual information (NMI) as the objective function is designed. Then, starting from the center of the cluster, the density is expanded in a breadth-first search method. Finally, find out the nearest neighbors of the prototype cluster and assign them sequentially. Experimental results on 8 artificial synthetic data sets show that the improved algorithm reduces the sensitivity of the clustering effect to the cutoff distance, and the improved algorithm's ACC, ARI and AMI have increased by about 35%, 90% and 80% respectively compared with the original algorithm.

Keywords: density peak; heuristic optimization; breadth first search; density expansion; cluster nearest neighbor