Hans Journal of Data Mining 数据挖掘, 2020, 10(1), 39-46
Published Online January 2020 in Hans. http://www.hanspub.org/journal/hjdm
https://doi.org/10.12677/hjdm.2020.101004

Hans汉斯

Clustering Of Passenger Picked-Up Areas and Analysis of Hot Spots Based on Taxi Business Data

Hongyang Wang

School of Information Technology & Management, University of International Business and Economics, Beijing
Email: wonghy7@outlook.com

Received: Dec. 21st, 2019; accepted: Jan. 1st, 2020; published: Jan. 8th, 2020

Abstract

To optimize the taxi space resource dispatch, a clustering model using OPTICS algorithm is proposed for the clustering of taxi passenger carrying area under the condition of intensive road and small difference in data density. By comparing with the results of the traditional DBSCAN algorithm in experiment, it is found that the OPTICS algorithm can effectively eliminate the interference of parameter setting on the experimental results. It helps to the problem of traditional algorithm DBSCAN applied to such situation, and enhances the efficiency of taxi dispatch. This research has practical guiding effect on improving the load factor of taxi and reducing the time of idle load.

Keywords

Taxi, Passenger Picked-Up Areas, OPTICS

基于出租车运营数据的载客区域聚类及热点特征分析

王弘扬

对外经济贸易大学信息学院，北京
Email: wonghy7@outlook.com

收稿日期：2019年12月21日；录用日期：2020年1月1日；发布日期：2020年1月8日

摘　要

为优化出租车空间资源调度，针对路网密集、数据密度差异较小条件下的城市出租车载客区域聚类问题，

文章引用：王弘扬. 基于出租车运营数据的载客区域聚类及热点特征分析[J]. 数据挖掘, 2020, 10(1): 39-46.
DOI: 10.12677/hjdm.2020.101004

本文提出了一种采用OPTICS算法的聚类模型。通过实验与传统DBSCAN算法下的聚类结果进行对比，发现OPTICS算法更能有效地避免参数设置对实验结果的影响，解决传统DBSCAN算法在此类应用中聚类划分模糊的问题，达到出租车资源调度精细化的效果，对提升出租车载客率、降低空载时间具有现实指导作用。

关键词

出租车，载客区域，OPTICS

1. 引言

出租车市场供求不平衡问题是导致出租车司机收入、满意度下降的直接原因，间接导致司机调度距离过长，乘客超时等待取消订单等问题，直接制约出租车市场的发展，而出租车市场的供需不平衡问题往往具有短期、局部区域性等特点，如何更好地实现资源调度、满足出行需求是学术界和业界一直在探索的问题。

在互联网环境下，依托数据挖掘和大数据技术的发展，此类问题有着天然的应用场景和解决优势。郑林江等人[1]提出一种基于网格密度的聚类算法，将轨迹点投影在按网格划分的空间区域上，并根据轨迹点网格密度设定阈值，确认热点区域网格单元。Rami Ibrahim [2]在研究中发现与随机交换聚类算法的聚类结果在均方误差和聚类质量上要优于基于层次密度的空间聚类算法。桂智明[3]基于 MapReduce 方法将出租车行驶轨迹中的轨迹点进行聚类并匹配路网得到其行驶轨迹，减少了轨迹中无意义簇的生成，实现了更具优势的轨迹聚类。姬波[4]基于信息论来研究出租车的空载密集区域，从而达到提高载客率的效果。毕硕本[5]使用 Ripley'K 函数检验出租车上下客事件的空间聚集关系，分析出租车运营的周期性变化特征及时空分布规律，发现出租车运营具有明显的早晚高峰期，且工作日和非工作日也有较大差别。曲昭伟[6]提出轨迹线密度方法，并实际运用于成都市春熙路商圈，根据当地路网密度值划分路段热点，识别热点位置，并依据实际的出行数据对该方法进行实证分析，发现该方法对出行需求量高的路段区域具有更明显的识别效果。Tang [7]将 DBSCAN 应用于哈尔滨市不同区域的出租车上下客区域聚类，通过从 DPS 数据中提取的行程距离、时间、载客状态和非载客状态下的平均移动速度等影响因素来研究乘客的移动行为，并应用观测到的城市中心区 OD 矩阵建立了基于熵极大化方法的交通分布模型。

互联网条件下，出租车市场大量的运营数据有力地驱动本问题的解决。出租车需求在时间、空间上具有一定的相似性，而随着时间的推移，出租车需求在空间上不断地转移，形成一定潮汐效应的同时又有着实时差异。

为了探究城市出租车需求在时间上的精细变化，减少司机等待、空车时间，提升资源调度效率，本文将基于 DBSCAN 算法和 OPTICS 算法对出租车载客热点区域进行聚类划分并对比两种方法的聚类结果差异，证明 OPTICS 算法在应对簇间轮廓模糊的数据集时能更有效地划分簇结果，为具有不同路网类型的城市出租车载客热点预测提供新的参考。

2. 数据预处理

实验数据采自纽约市 2009 年~2015 年间的 100 万条出租车运营数据，每一条数据包含一次出租车行程单的 7 个属性，包括乘客上车时间、上下车经纬度、乘客人数、费用等信息。

首先删除包含缺失值或明显异常值的数据；由于纽约城临海，行程起终点 GPS 定位信息落于海中的数据量要大于内陆城市，本文选择将起终点地点落于水域的数据删除，得到 978,643 条数据。

经过初步的数据清洗，噪音数据有效减少，即便仍存在的小部分起终点距离过大的数据难以区别是否真实有效，因其数据量不足一百条，不足以对实验产生明确影响。对时间数据进行变量分类转换，区分月份、星期几、第几小时等离散化变量。

新增行程起终点距离属性列，起终点距离为：

$$\text{distance} = 2R \cdot \arcsin\sqrt{\text{haversin}\left(\frac{d}{R}\right)}$$

其中：

$$\text{haversin}\left(\frac{d}{R}\right) = \text{haversin}\left(\varphi_2 - \varphi_1\right) + \cos\left(\varphi_1\right) \cdot \cos\left(\varphi_2\right) \cdot \text{haversin}\left(\Delta\lambda\right)$$

$$\text{haversin}\left(\theta\right) = \sin^2\left(\frac{\theta}{2}\right)$$

3. 聚类建模

3.1. DBSCAN 算法介绍

DBSCAN 算法[8]是最为知名的基于密度的聚类算法，尤其广泛应用于空间密度聚类领域。

概念：

假设具有样本集 $D = \{x_1, x_2, x_3, \cdots, x_n\}$。

ε 邻域：对于任一点 $x_i \in D$，$U\left(x_i, \varepsilon\right)$ 称为 x_i 的邻域，邻域内所有的点与 x_i 的距离都不大于 ε，x_i 在 ε 邻域内的点的集合为 $N_\varepsilon\left(x_i\right)$。

$P_{\min}$：定义使得核心点成立的 ε 邻域内最低点数。

核心点：如果一个给定点 x_i 在其 ε 邻域内至少有 P_{min} 个样本点，那么就称该给定点 x_i 为核心点。

密度直达：对于核心点 x_i，如果存在样本点 x_i 满足 x_i 和 x_j 间的距离不大于 ε，那么称 x_i 密度直达 x_j。

密度可达：对于一系列样本点 $x_t, x_{t+1}, x_{t+2}, \cdots, x_{t+k}$，如果存在如下关系：$x_{t+m-1}$ 密度直达 x_{t+m}，那么 x_t 密度可达 x_{t+k}。

密度相连：在给定 ε 及 $P_{\min}$ 条件下，存在样本点 x_i, x_j 同时可由另一样本点 x_k 密度可达，则称样本点 x_i, x_j 密度相连。

DBSCAN 算法核心思想可以解释为：对于样本集合中任意一个样本点 x_i，根据 ε 及 $P_{\min}$ 条件判断该样本点是否为核心点，若点 x_i 是核心点，则依照密度直达的方式寻找其 ε 邻域内所有样本点集合 $N_\varepsilon\left(x_i\right) = \left\{x_j \mid x_j \in D \wedge x_j \neq x_i \wedge \text{distance}\left(x_i, x_j\right) < \varepsilon\right\}$，再判断 $N_\varepsilon\left(x_i\right)$ 内所有点是否为核心点，重复该步骤寻找所有由 x_i 出发密度相连的点构成一个簇，直至样本集 D 中所有的点都落入某个簇集合或非簇集合。

DBSCAN 作为一种基于空间密度地聚类算法，发现一个密度区域后以密度可达的方式拓展簇的范围，其聚簇轮廓往往是极不规则的，DBSCAN 在面对非凸集数据的时候体现了出色的聚类能力及抗噪能力。

3.2. OPTICS 算法介绍

OPTICS 算法[9]是一种基于 DBSCAN 的改进算法。由于 DBSCAN 在面对未知数据分布情形下对 ε 邻域及 P_{min} 值过于敏感，不同的参数选择对于其聚类结果产生较大的影响，也容易影响其聚类质量，在面对不断变化的密度空间数据时缺乏可控性。针对以上缺点，学者对其提出了改进型算法 OPTICS。

概念：

在 DBSCAN 相关概念的基础上，OPTICS 算法引入了两个新的定义。

核心距离：样本 $x \in D$，对于给定 ε 和 $P_{\min}$，如果 $U(x,\varepsilon)$ 内存在至少 $P_{\min}$ 个样本点，那么 $U(x,\varepsilon)$ 内与 x 距离相近的第 $P_{\min}$ 个点与 x 之间的距离就是 x 的核心距离。$N_{\varepsilon,P_{\min}}(x)$ 为样本 x 在其 ε 邻域内距离 x 第 $P_{\min}$ 近的点，其数学表达式如下：

$$\text{CoreDistance}(x)=\begin{cases}\text{undefined}, & \text{NumOf}\left(N_{\varepsilon}(x)\right)<P_{\min}\\ \text{distance}\left(N_{\varepsilon,P_{\min}}(x),x\right), & \text{else}\end{cases}$$

可达距离：样本 $x,y \in D$，对于给定 ε 和 $P_{\min}$，x,y 之间的可达距离为：

$$\text{ReachableDistance}(x)=\begin{cases}\text{undefined}, & \text{NumOf}\left(N_{\varepsilon}(x)\right)<P_{\min}\\ \max\left(\text{CoreDistance}(x),\text{Distance}(x,y)\right), & \text{else}\end{cases}$$

直接密度可达：对于样本 $x,y \in D$，若 x 为核心点，$\text{Distance}(x,y)<\varepsilon$，则称 x 到 y 直接密度可达。

计算过程：

1) 输入数据样本集合D，设置参数 ε 和 $P_{\min}$。

2) 初始化序列 A 和 B。

3) 判断集合D中是否存在未处理样本点，是则转入步骤 4，否则算法结束。

4) 选择一个D中未处理的核心点，将该核心点放入 A 序列，并将该核心点的直接密度可达点按可达距离升序排序放入 B 序列。

5) 判断 B 序列是否为空，是则转入步骤 3，否则转入步骤 6。

6) 选择 B 序列第一个点，将该点放入 A 序列，若该点为未处理核心点则将该点直接密度可达点放入 B 序列，并将 B 序列重新排序，否则转入步骤 5。

7) 重复步骤 3。

8) 算法结束。

4. 评价标准

Davies-Boulding 指数(DBI)

DBI 指数戴维森堡丁指数[10]，是一种针对聚类结果的无监督评价指标，由 Davies, D.L.，Bouldin, D.W. 提出。该方法首先定义了一个变量 S_i：

$$S_i=\left(\frac{1}{T_i}\sum_{j=1}^{T_i}\left|X_j-A_i\right|^p\right)^{\frac{1}{p}}$$

式中 T_i 代表第 i 个簇中的数据点个数，X_j 代表簇 i 中的第 j 个数据点，A_i 为簇 i 的簇核心，p 值用于调整距离计算方法；变量 S_i 衡量的是簇内样本点与簇核心的距离均值。

其次，DBI 指数定义了变量 $M_{i,j}$：

$$M_{i,j}=\left(\sum_{k=1}^{n}\left|a_{k,i}-a_{k,j}\right|^p\right)^{\frac{1}{p}}$$

$a_{k,i}$ 表示第 i 个簇的簇核心的第 k 个属性值，也就是说，$M_{i,j}$ 衡量的是第 i 个簇和第 j 个簇的簇核心间距。

随后，DBI 指数定义了变量 $R_{i,j}$：

$$R_{i,j} = \frac{s_i + s_j}{M_{i,j}}$$

$R_{i,j}$ 用于衡量第 i 个簇和第 j 个簇之间的相似度。

R_i 则为第 i 个簇与其他簇的相似度的最大值：

$$R_i = \max\left(R_{i,j}\right)$$

最终得到 DBI 指数：

$$\mathrm{DBI} = \frac{1}{n}\sum_{i=1}^{N} R_i$$

可见，DBI 指数非常良好地解释了聚类结果的离散程度，当同一簇内样本点离散程度较低而样本集合内各个簇核心间距较大时，DBI 指数较小，聚类结果簇集划分效果较明显。

5. 实验

实验环境：Windows10 x86 操作系统，Python 为 3.6 版本，scikit-learn 为 0.21.3 版本。

对本数据集打车起终点距离进行统计，统计结果如图 1 所示。

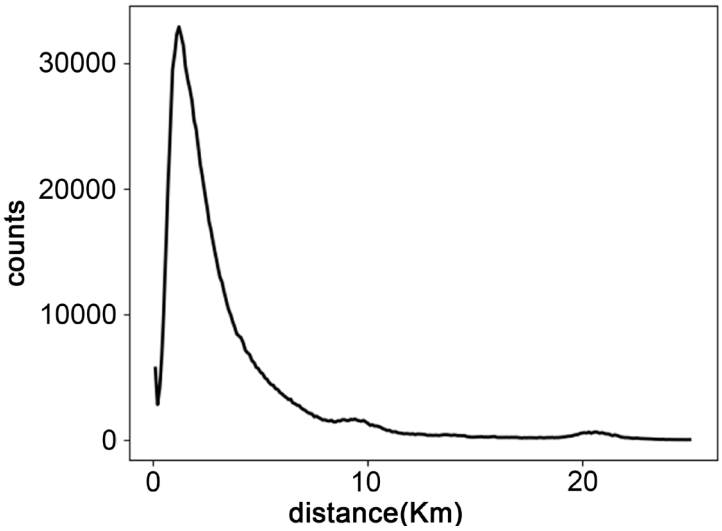

Figure 1. Distance frequency chart of the taxi journey
图 1. 打车起终点距离频数分布

数据表明，本数据集打车起终点距离均值为 3.32 千米，标准差为 3.73 千米，50%的行程起终点距离在 1.25 千米至 3.92 千米之间，图 1 横轴位于 10 千米和 20 千米处各有一个较为明显的突起，通过后续聚类结果发现原因为该城市航空公共交通基础设施客流聚集。可以看出，打车需求以中短程需求为主。

考虑到随时间推移城市交通设施结构的变化对数据的影响，本文选择 2013、2014 年周一至周五 9:00~10:00 期间产生的行程数据作为模型输入。

5.1. DBSCAN 算法实验

使用 DBSCAN 对其空间位置进行聚类，通过网格搜索的方式确定一个较为合理的聚类结果，选择在

ε 邻域取值 100 米，P_{min} 取值 14 条件下，其空间聚类结果如图 2 所示，灰色点为非簇样本，粉色点为所有样本数低于 100 的簇样本。

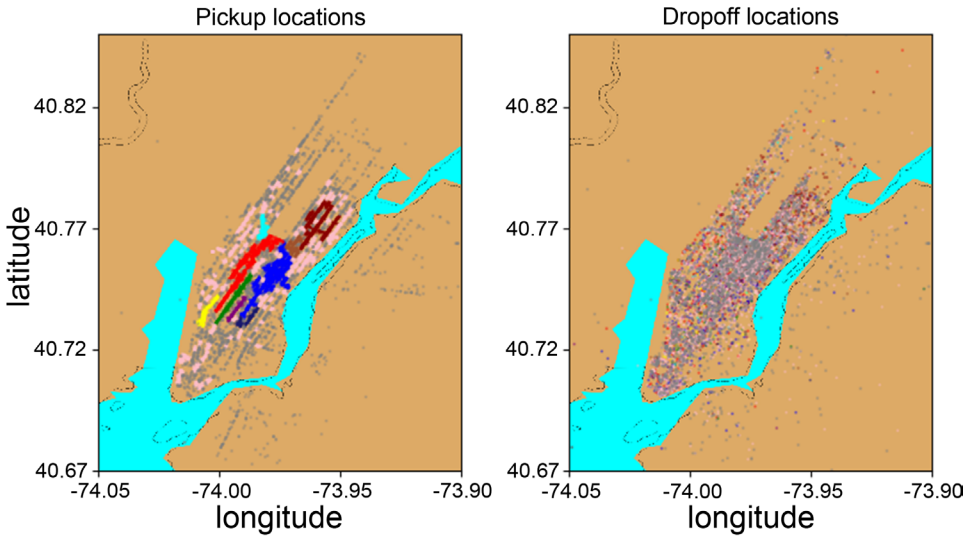

Figure 2. Result of DBSCAN clustering experiment
图 2. DBSCAN 算法空间聚类结果

5.2. OPTICS 算法实验

使用 OPTICS 对同一数据进行聚类，同样使用网格搜索的方式寻找局部较优参数，设置 P_{min} 为 5，max_eps 为 0.05 千米，空间聚类结果如图 3 所示，灰色点为非簇样本，共有 9 个样本数大于 100 的彩色簇集。

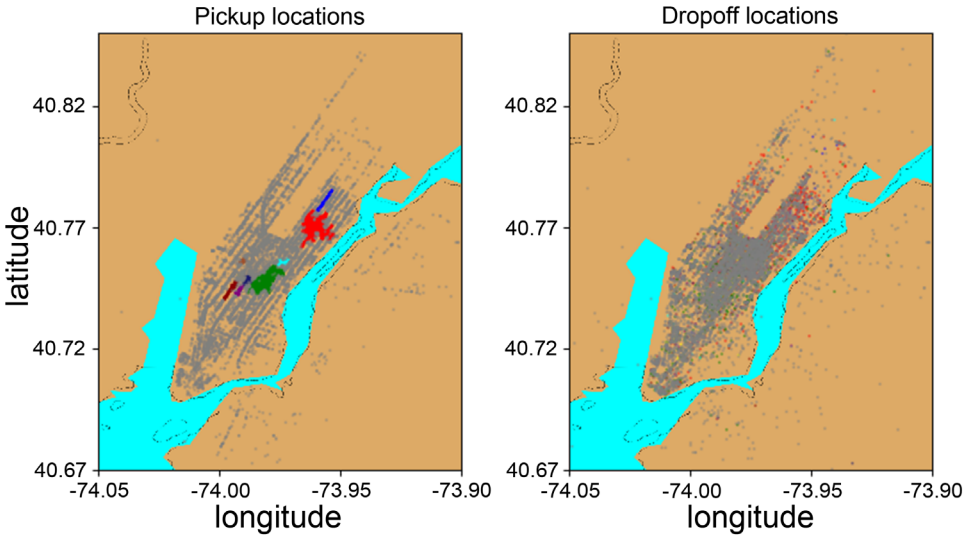

Figure 3. Result of OPTICS clustering experiment
图 3. DBSCAN 算法空间聚类结果

5.3. 实验结果统计与分析

两种方法得出的主要簇样本统计结果如表 1~2 所示。

Table 1. Statistics of the main clusters under DBSCAN
表 1. DBSCAN 聚类结果主要簇的样本点统计

簇编号	样本量	簇核心	起终点簇核心偏移量	起终点距离均值	起终点距离标准差
0	1494	(40.749796, −73.978731)	(−0.000434, −0.002933)	2.25	2.33
1	1364	(40.754429, −73.988863)	(−0.004354, 0.006218)	2.39	2.7
2	861	(40.771835, −73.959432)	(−0.008411, −0.012519)	2.28	1.8
3	380	(40.741781, −73.993604)	(0.006179, 0.007705)	2.04	1.6
4	260	(40.735644, −74.005358)	(0.008528, 0.017461)	2.58	2.66
5	231	(40.762835, −73.969031)	(−0.001989, −0.00257)	1.94	1.69
6	221	(40.769991, −73.982147)	(−0.008155, 0.004579)	2.02	1.46
7	162	(40.732825, −73.988756)	(0.008949, 0.000837)	2.01	1.05
8	136	(40.736977, −73.992913)	(0.003781, 0.003019)	2.2	1.92

Table 2. Statistics of the main clusters under OPTICS
表 2. OPTICS 聚类结果主要簇的样本点统计

簇编号	样本量	簇核心	起终点簇核心偏移量	起终点距离均值	起终点距离标准差
0	106	(40.781022, −73.956669)	(−0.014664, −0.01384)	2.46	1.57
1	1341	(40.769841, −73.960288)	(−0.008099, −0.011928)	2.31	1.93
2	134	(40.744907, −73.995032)	(−0.001333, 0.002858)	1.95	1.52
3	1148	(40.748236, −73.981069)	(−0.000374, −0.002332)	2.02	1.96
4	109	(40.750132, −73.991533)	(−0.003372, 0.006874)	2.22	2.23
5	106	(40.756173, −73.990491)	(−0.000316, 0.009939)	1.85	1.22
6	109	(40.755899, −73.97371)	(−0.000823, −0.002366)	2.69	2.94
7	168	(40.748347, −73.988695)	(0.003076, 0.005769)	1.89	1.81
8	100	(40.744145, −73.991906)	(0.00548, 0.003931)	1.92	0.97

Table 3. Comparison of clustering results under DBSCAN and OPTICS
表 3. DBSCAN 和 OPTICS 算法下聚类结果对比

算法	DBI	主要簇样本量
DBSCAN	60.05	5109
OPTICS	3.27	3321

根据实验结果的统计数据可知，OPTICS 算法下的实验结果具有更低的 DBI 指数，这表明各簇内样本点的离散程度低而簇间核心间距较大，具有明显的聚簇划分效果。本方法相较于常用的 DBSCAN 方法更利于出租车载客区域空间密度的划分，便于运营人员明确核心区域中心，将空载率高的区域车辆向高载客频率区域调度，减少出租车空载时间和乘客等待时间。

6. 结论

由于纽约市出租车载客区域更为密集，且簇间轮廓较为模糊，特定城市的载客热点聚类应当因地制宜采用不同的方法，本文在前人研究的基础上，分别采用 DBSCAN 和 OPTICS 两种算法对纽约市出租车

载客区域进行了聚类划分，多次调参取得实验结果差异如表 3 所示，DBI 差异明显，可见 OPTICS 算法在面对簇间轮廓模糊的数据集时可以很大程度上避免 ε 邻域及 $P_{\min}$ 值设置所带来的实验干扰，聚类效果强于 DBSCAN。

本研究有助于解决不同城市路网条件下出租车的资源调度问题，对大数据精细化运营有重要的现实意义。

参考文献

[1] 郑林江, 赵欣, 蒋朝辉, 邓建国, 夏冬, 刘卫宁. 基于出租车轨迹数据的城市热点出行区域挖掘[J]. 计算机应用与软件, 2018, 35(1): 1-8.

[2] Ibrahim, R. and Omair Shafiq, M. (2019) Detecting Taxi Movements Using Random Swap Clustering and Sequential Pattern Mining. *Journal of Big Data*, **6**, 1-26. https://doi.org/10.1186/s40537-019-0203-6

[3] 桂智明, 向宇, 李玉鉴. 基于出租车轨迹的并行城市热点区域发现[J]. 华中科技大学学报(自然科学版), 2012(S1): 187-190.

[4] 姬波, 叶阳东, 肖煜. 基于信息瓶颈方法的出租车空载聚集区聚类算法[J]. 小型微型计算机系统, 2013, 34(9): 2139-2143.

[5] 毕硕本, 万蕾, 杨树亮, 闫业超, Nkunzimana Athanase. 基于 GPS 数据的南京出租车上下客时间特征及热点时空分布[J]. 中国科技论文, 2018, 13(9): 1023-1028.

[6] 曲昭伟, 王鑫, 宋现敏, 夏英集, 袁咪莉. 基于出租车 GPS 大数据的城市热点出行路段识别方法[J]. 交通运输系统工程与信息, 2019, 19(2): 238-246.

[7] Tang, J.J., Liu, F., Wang, Y.H., *et al*. (2015) Uncovering Urban Human Mobility from Large Scale Taxi GPS Data. *Physica A*: *Statistical Mechanics and its Applications*, **438**, 140-153. https://doi.org/10.1016/j.physa.2015.06.032

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

[9] Ankerst, M., Breunig, M., Kriegel, H.P. and Sandler, J. (1999) OPTICS: Ordering Points to Identify the Clustering Structure. *Proceedings of the International Conference on Management of Data* (*SIGMOD*'99), Philadelphia, 1-3 June 1999, 49-60. https://doi.org/10.1145/304182.304187

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