Karafan
Quarterly Scientific Journal of National University of Skills
Spring 2025, Vol. 22, No. 1, p.36-59
Journal Homepage: https://karafan.tvu.ac.ir/?lang=en
https://karafan.nus.ac.ir/article_229118.html

Clustering based on Affinity Matrix First Search-Breadth Graph

Shahin pourbahrami*[1]

[1] Assistant Professor, Department of Computer Engineering, National University of Skills (NUS), Tehran, Iran

ARTICLE INFO

Received: 20.08.2024
Revised: 03.11.2024
Accepted: 24.11.2024

Keyword:
Clustering
nearest neighbor
first level search
Affinity matrix
Graph

***Corresponding Author:**
Shahin pourbahrami
Email: *shpourbahrami@tvu.ac.ir*

ABSTRACT

Density-based clustering algorithms are commonly used in machine learning and data mining due to their ability to identify clusters with different shapes and noisy objects. These algorithms are famous in data analysis and the use of their analysis output in industry and business. However, traditional clustering algorithms may have difficulty with datasets with different densities and overlapping neighboring clusters. To address these challenges, a new density-based clustering algorithm was proposed in the present research. In this algorithm, the dependency matrix and the first-level search graph are used to find the dense points and the connection between the points. The concept of the relevant space is introduced to define the local and global density, and a central point identification method is used to identify the cluster structures. This algorithm also uses an allocation strategy based on the relevant space for the remaining objects to achieve accurate clustering results. Experimental results on real datasets demonstrated the effectiveness of the proposed method in clustering performance.

E-ISSN: 2538-4430
ISSN: 2382-9796

EXTENDED ABSTRACT

Introduction

Density-based clustering algorithms are commonly used in machine learning and data mining due to their ability to identify clusters with different shapes and noisy objects. These algorithms are famous in data analysis and the use of their analysis output in industry and business. However, traditional clustering algorithms may have difficulty with datasets with different densities and overlapping neighboring clusters. To address these challenges, a new density-based clustering algorithm is proposed in this article. In this algorithm, the dependency matrix and the first-level search graph are used to find the dense points and the connection between the points. The concept of the relevant space is introduced to define the local and global density, and a central point identification method is used to identify the cluster structures. This algorithm also uses an allocation strategy based on the relevant space for the remaining objects to achieve accurate clustering results. Experimental results on real data sets show the effectiveness of the proposed method in clustering performance.

Clustering is a common type of unsupervised learning technique that plays an important role in data mining and machine learning [1]. Its main goal is to discover the inherent structure in a data set by grouping objects into clusters based on certain criteria. This process aims to minimize the differences in one cluster while maximizing the differences between different clusters. Clustering is a fundamental unsupervised learning approach that is used in various fields, including pattern recognition [2], image processing [3], bioinformatics [4], and information retrieval [5]. Currently, clustering is classified into different types based on different methods such as segmentation-based [6], model-based [7], hierarchy-based [8] and density-based clustering [9].

Methodology

The first step: formation of main sub-clusters.

The second step: creating a graph of the similarity of the reference point with its neighbors by the connection coefficient in the dependency matrix.

The third step: checking the connection of sub-clusters and non-cluster points in the graph.

Results and Discussion

Comparative experiments were designed to check the effectiveness and efficiency of the proposed algorithm in this section. ARI [31] and NMI [32] were used to evaluate the clustering results of different algorithms on real-world datasets. ARI and NMI values ranged from [-1,1] and [0,1], respectively, and the larger the value of these evaluation criteria, the better the clustering result.

The proposed algorithm was implemented in Google Colab with Python programming language, and its performance was evaluated on fourteen real-world datasets from the UCI Machine Learning Repository - Datasets site. Table 1 shows the details of the real-world dataset. There are two real-world data sources: one is the UCI repository containing biological datasets and the other is the KAGLE delta repository.

Table 1. Dataset on a real dataset.

DATASET	D	N	M	TYPE
Iris	4	150	3	REAL
Wine	13	178	3	REAL
Seed	7	210	3	REAL
Ecoil	8	336	8	REAL
WDBC	31	569	2	REAL
Dermatoloy	34	366	6	REAL
Frogs	22	7195	10	REAL
Heart	12	270	2	REAL
Yeast	8	1484	10	REAL
Libras	90	330	15	REAL
Segmentation	19	2100	7	REAL
Pen digits	16	7494	10	REAL
Nursery	8	12960	3	REAL
UJIIndoorLoc	520	21048	3	REAL

Table 2. The Clustering Results on UCI Datasets (ARI).

dataset	GCNN	NCAR	AFK	FKNN-DPC	NCARD	DPC	Our algorithm
Iris	0.90	0.۶۲	0.90	0.88	0.90	0.88	**0.۹۱**
Wine	**0.91**	0.71	0.۸۱	0.86	0.72	0.69	**0.91**
Seed	0.76	0.75	0.۷۷	0.76	0.۷۷	0.74	**0.78**
Ecoil	**0.75**	0.۵۷	0.۷۲	0.53	0.72	0.45	**0.75**
WDBC	0.82	0.۷۵	0.73	0.69	0.5۲	0.50	**0.83**
Dermatoloy	**0.84**	0.73	0.۸۵	0.79	0.۷۶	0.72	**0.84**
Segmentation	0.62	0.5۸	0.5۸	0.65	0.56	0.66	**0.67**
Heart	0.33	0.۲0	0.۲5	0.32	0.28	0.31	**0.37**
Yeast	0.26	0.۲۸	0.2۸	0.07	0.14	0.09	**0.30**
Libras	0.36	0.۳۸	0.۴۱	0.35	0.۳۹	0.35	**0.42**
Pen digits	0.67	0.۶۷	0.67	0.58	0.۶۶	0.67	**0.۶۸**
Frogs	0.81	0.۶۸	0.۷۴	0.65	0.۶۸	0.72	**0.82**
Nursery	0.40	0.2۶	0.۴4	0.35	0.24	0.51	**0.53**
UJIIndoorLoc	0.40	0.5۶	**0.۶۲**	0.57	0.۵3	0.43	**0.62**

Average	۰٫۶۳	۰٫۵۵	۰٫۶۲	۰٫۵۷	۰٫۵۶	۰٫۵۵	**۰٫۶۷**

Table 3. The Clustering Results on UCI Datasets (NMI).

dataset	GCNN	NCAR	AFK	FKNN-DPC	NCARD	DPC	Our algorithm
Iris	**0.93**	0.68	0.88	0.86	0.89	0.86	**0.93**
Wine	**0.89**	0.79	0.80	0.84	0.81	0.72	0.79
Seed	0.75	0.74	0.75	0.74	0.74	0.72	**0.76**
Ecoil	0.69	0.67	0.67	0.57	0.66	0.59	**0.70**
WDBC	**0.72**	0.69	0.62	0.64	0.69	0.48	**0.72**
Dermatoloy	**0.91**	0.86	0.90	0.86	0.89	0.82	**0.91**
Segmentation	0.75	0.72	0.76	0.76	0.75	0.75	**0.77**
Heart	0.33	0.38	**0.40**	0.28	0.42	0.24	0.33
Yeast	0.29	0.30	0.27	0.15	0.27	0.21	**0.32**
Libras	0.63	0.58	0.68	0.64	0.66	0.64	**0.69**
Pen digits	0.78	0.78	**0.79**	0.77	0.78	0.77	0.72
Frogs	0.69	0.69	0.69	0.60	0.69	0.61	**0.70**
Nursery	0.29	0.29	0.33	0.45	0.51	0.64	**0.66**
UJIIndoorLoc	0.55	0.58	**0.72**	0.70	0.60	0.58	**0.72**
Average	۰٫۶۵	۰٫۶۲	۰٫۶۶	۰٫۶۳	۰٫۶۴	۰٫۶۱	**۰٫۶۹**

Conclusions

In the present research, a new clustering algorithm was introduced, which finds clusters based on a dependency matrix and tree density search. A new neighborhood space was introduced to define the local and global density, and a cluster center point search method was used. The proposed algorithm provides a method based on the relevant space to allocate other unclustered data in the last step. The experimental results on the real data set showed the effective performance of the introduced clustering method.

کارافن
فصلنامه علمی دانشگاه ملی مهارت
بهار۱٤۰٤، دوره ۲۲، شماره ۱، ۵۹-۳٦
آدرس نشریه: https://karafan.nus.ac.ir/
https://karafan.nus.ac.ir/article_229118.html

خوشه‌بندی براساس ماتریس وابستگی و گراف جستجوی سطح اول

شهین پوربهرامی*[1]

1- استادیار گروه مهندسی کامپیوتر، دانشگاه ملی مهارت، تهران، ایران[1]

اطلاعات مقاله

دریافت مقاله: ۱۴۰۳/۰۵/۳۰
بازنگری مقاله: ۱۴۰۳/۰۸/۱۳
پذیرش مقاله: ۱۴۰۳/۰۹/۰۴

کلید واژگان:
خوشه بندی
نزدیکترین همسایه
جستجوی سطح اول
ماتریس وابستگی
گراف

***نویسنده مسئول:** شهین پوربهرامی
پست الکترونیکی:
shpourbahrami@tvu.ac.ir

چکیده

الگوریتم‌های خوشه‌بندی مبتنی بر چگالی به دلیل توانایی آنها در شناسایی خوشه‌هایی با اشکال مختلف و اشیاء نویز معمولاً در یادگیری ماشین و داده‌کاوی استفاده می‌شوند. این الگوریتم‌ها در تحلیل داده‌ها و کاربرد خروجی تحلیل آنها در صنعت و تجارت معروف هستند. با این حال، الگوریتم‌های سنتی خوشه‌بندی ممکن است در مجموعه داده‌هایی با چگالی‌های مختلف و خوشه‌های همسایه درهم‌تنیده مشکل داشته باشند. برای پرداختن به این چالش‌ها، یک الگوریتم خوشه‌بندی مبتنی بر چگالی جدید در این مقاله پیشنهاد شده است. در این الگوریتم ماتریس وابستگی و گراف جستجوی سطح اول برای یافتن نقاط پر چگال و ارتباط بین آنها استفاده شده است، مفهوم فضای مربوطه برای تعریف چگالی محلی و سراسری معرفی می‌گردد، و از یک روش شناسایی نقاط مرکزی برای شناسایی ساختارهای خوشه‌ای استفاده شده‌است. این الگوریتم همچنین از یک استراتژی تخصیص بر اساس فضای مربوطه برای اشیاء باقی مانده برای دستیابی به نتایج خوشه‌بندی دقیق استفاده می‌کند. نتایج تجربی بر روی مجموعه داده‌های واقعی، اثربخشی روش پیشنهادی را در عملکرد خوشه‌بندی نشان می‌دهد.

شاپای الکترونیکی: 2538-4430
شاپای چاپی: 2382-9796

۱. مقدمه

خوشه‌بندی یک نوع رایج تکنیک یادگیری بدون‌نظارت است که نقش مهمی در داده‌کاوی و یادگیری ماشین ایفا می‌کند [۱]. هدف اصلی آن کشف ساختار ذاتی در یک مجموعه داده با گروه‌بندی اشیاء در خوشه‌ها بر اساس معیارهای خاص است (به حداقل رساندن تفاوت‌ها در یک خوشه در حالی که تفاوت بین خوشه‌های مختلف را به حداکثر می‌رسانیم). خوشه‌بندی یک رویکرد اساسی یادگیری بدون‌نظارت است که در حوزه-های مختلف، از جمله تشخیص الگو [۲]، پردازش تصویر [۳]، و اینترنت اشیا [۴] [۵] بکار گرفته می‌شود. در حال حاضر، خوشه‌بندی بر اساس روش‌های مختلف به انواع مختلفی دسته‌بندی می‌شود: مانند مبتنی بر تقسیم‌بندی [۶]، مبتنی بر مدل [۷]، مبتنی بر سلسله مراتب [۸] و خوشه‌بندی مبتنی بر چگالی [۹].

شناخته شده‌ترین الگوریتم‌ها در این زمینه شامل خوشه‌بندی فضایی برنامه‌های کاربردی با نویز مبتنی بر چگالی[۱] [۱۰]، نقاط سفارش برای شناسایی ساختار خوشه‌بندی[۲] [۱۱] و الگوریتم میانگین شیفت[۳] است [۱۲]. الگوریتم سنتی خوشه‌بندی فضایی برنامه‌های کاربردی با نویز مبتنی بر چگالی خوشه‌هایی با اشکال مختلف را شناسایی می‌کند و نقاط پرت را بر اساس یک معیار متصل به چگالی حذف می‌کند. علاوه بر این، این روش خوشه‌بندی چگالی نیازی به تعیین تعداد خوشه‌ها ندارد و به‌طور خودکار شکل خوشه‌ها و تعداد خوشه‌ها را برای رسیدگی به چالش‌های مختلف خوشه‌بندی چگالی تشخیص می‌دهد. با این حال، تنظیم شعاع همسایگی ϵ و حداقل تعداد نمونه در خوشه‌بندی فضایی برنامه‌های کاربردی با نویز مبتنی بر چگالی ضروری است. علاوه بر این، مجموعه داده‌های مختلف ممکن است به تنظیمات پارامتر متفاوتی نیاز داشته باشند که منجر به چالش‌هایی در مدیریت نمونه‌های مرزی می‌شود. انتخاب پارامترهای نامناسب می‌تواند بر نتایج خوشه‌بندی تأثیر مخرب بگذارد [۱۲].

طبق گفته رودریگز و لایو (۲۰۱۴) در حوزه الگوریتم‌های خوشه‌بندی مبتنی بر چگالی، الگوریتم خوشه‌بندی پیک چگالی جستجوی سریع[۴] به عنوان یک الگوریتم برجسته و کاربردی است [۱۳]. مفهوم اصلی این الگوریتم شامل ارزیابی اهمیت نقاط نمونه بر اساس چگالی محلی و فاصله آن‌ها از سایر نمونه‌ها برای شناسایی پیک‌های چگالی است. در ابتدا، نمونه‌های با چگالی بالا که به اندازه کافی از سایر نقاط با چگالی بالا فاصله دارند، به عنوان مراکز خوشه‌بندی از نمودار تصمیم انتخاب می‌شوند. پس از آن، نمونه‌های باقی‌مانده با نزدیک‌ترین نمونه با چگالی بالاتر گروه‌بندی می‌شوند. با این حال، در کاربردهای عملی، الگوریتم خوشه‌بندی پیک چگالی در انتخاب مراکز خوشه‌ای و تخصیص دقیق نقاط با چالش‌هایی مواجه می‌شود. یکی از مسائل، تلاش الگوریتم برای تعیین دقیق مراکز خوشه در نمودارهای تصمیم‌گیری با

[1] Density-Based Spatial Clustering of Applications with Noise (DBSCAN)
[2] Ordering Points to Identify the Clustering Structure (OPTICS)
[3] Mean-Shift
[4] Density Peak Clustering (DPC)

چگالی‌های متفاوت است [۱۴]. علاوه بر این، استراتژی مرسوم برای انتخاب مراکز خوشه‌ای در الگوریتم خوشه‌بندی پیک چگالی ممکن است برای مجموعه داده‌های پیچیده مناسب نباشد [۱۵]. نگرانی دیگر آسیب‌پذیری نتایج الگوریتم خوشه‌بندی پیک چگالی در برابر مشکل «دومینو» است [۱۶]، که در آن یک نقطه با چگالی بالاتر نادرست می‌تواند منجر به طبقه‌بندی اشتباه نقاط همسایه با چگالی پایین‌تر آن در همان خوشه شود.

معایب الگوریتم‌های خوشه‌بندی مبتنی بر چگالی ذکر شده در بالا همچنان وجود دارد. یک الگوریتم خوشه-بندی جدید برای رسیدگی به مسائل مربوط به انتخاب مراکز خوشه و تخصیص نقاط نمونه معرفی شده است. این الگوریتم از یک ساختار ماتریس وابستگی و گراف جستجوی سطح اول استفاده می‌کند. این روش بر اساس شباهت خوشه‌بندی و چگالی اتصال گراف خوشه‌بندی انجام می‌دهد.

نوآوری‌های الگوریتم پیشنهادی عمدتاً به شرح زیر نشان داده شده است.

• یک تعریف جدید از شباهت بر اساس نزدیکترین همسایگان مرجع و ماتریس وابستگی و گراف جستجوی سطح اول پیشنهاد شده است.

• جستجوی کوتاه‌ترین مسیر برای نقاط دارای چگالی محلی مشابه که کوتاه‌ترین مسیر و اطلاعات ماتریس وابستگی بین دو نقطه داده را در نظر می‌گیرد تا بتواند اتصال بین نقاط را تضمین کند و همچنین به پیدا شدن دقیق‌تر سرخوشه کمک کند.

• به منظور کاهش زمان اجرای الگوریتم جستجوی سطح اول و کوتاه‌ترین مسیر پیشنهاد شده است. این استراتژی جستجو می‌تواند همسایگی اکثر نقاط داده را بدون جستجوی کل مجموعه داده‌ها پیدا کند، که زمان اجرا را تا حد زیادی کاهش می‌دهد.

ساختار باقیمانده مقاله به شرح زیر است: بخش ۲ ادبیات پیشین در مورد مفهوم اساسی الگوریتم خوشه‌بندی پیک چگالی و پیشرفت‌های آن را مورد بحث قرار می‌دهد. بخش ۳ توضیح مفصلی از الگوریتم پیشنهادی ارائه می‌دهد. بخش ۴ به تشریح یافته‌ها و مقایسه‌های تجربی می‌پردازد. در نهایت، بخش ۵ نتیجه‌گیری و مسیرهای بالقوه تحقیقات آینده را ارائه می‌دهد.

۲. کارهای مرتبط

۲.۱. ایده اصلی الگوریتم خوشه‌بندی پیک چگالی

الگوریتم خوشه‌بندی پیک چگالی[5]یک روش خوشه‌بندی کارآمد است که در زمینه‌های مختلف مورد استفاده قرار می‌گیرد [۱۷]. مفهوم اساسی آن شامل خوشه‌بندی با مشخص کردن اشیاء (نمونه) با چگالی اوج است، که نقاط داده‌ای با چگالی بالاتر در مقایسه با همسایگان خود هستند و معمولاً مراکز خوشه را نشان می‌دهند. فرآیند تایید اشیاء با چگالی اوج معمولاً شامل دو مرحله است: (۱) ایجاد یک ماتریس شباهت برای محاسبه چگالی محلی ρ و فاصله چگالی δ ۲) ایجاد یک نقشه تصمیم‌گیری بر اساس حاصل ضرب چگالی محلی و فاصله چگالی برای شناسایی مراکز خوشه. هنگامی که پیک‌های چگالی مشخص می‌شوند، نقاط غیرمرکزی به خوشه‌هایی که اجسام متراکم‌تر و نزدیک‌تر در آن قرار دارند، اختصاص داده می‌شوند.

ایده اصلی پشت الگوریتم خوشه‌بندی پیک چگالی یافتن مراکز خوشه‌ای با در نظر گرفتن چگالی محلی و فواصل نقاط نمونه است. در ابتدا، الگوریتم چگالی محلی هر نقطه نمونه را محاسبه می‌کند و سپس فاصله هر نقطه را تعیین می‌کند. سپس از این اطلاعات برای ایجاد یک نمودار تصمیم‌گیری برای شناسایی مراکز خوشه و توزیع خوشه استفاده می‌شود. با این حال، هنگام برخورد با خوشه‌هایی که دارای اشکال غیرکروی و چگالی‌های متفاوت هستند، انتخاب مراکز خوشه به طور دقیق از نمودار تصمیم می‌تواند بسیار چالش برانگیز باشد [۱۸].

در برخورد با مجموعه داده‌هایی که چگالی ناهموار و پراکنده هستند، انتخاب مراکز خوشه‌ای می‌تواند یک کار چالش برانگیز یا حتی غیرممکن باشد. این دشواری از این واقعیت ناشی می‌شود که روش خوشه‌بندی پیک چگالی برای محاسبه چگالی محلی و سراسری تنها بر فاصله اقلیدسی تکیه دارد. این محدودیت توانایی آن را برای ارزیابی دقیق شباهت واقعی بین نقاط در خوشه‌های غیرکروی و شناسایی مرکز خوشه درست از نمودار تصمیم را به چالش می‌کشد. در نتیجه، در مجموعه‌های داده با خوشه‌هایی با چگالی‌های متفاوت، الگوریتم خوشه‌بندی پیک چگالی ممکن است برای مشخص کردن مراکز خوشه‌ای مناسب مشکل داشته باشد که منجر به نتایج خوشه‌بندی نادرست می‌شود. بنابراین، نیاز به بررسی الگوریتم‌های خوشه‌بندی جدید وجود دارد که بتواند به طور موثر این مسئله خوشه‌بندی خاص را برطرف کند. استراتژی الگوریتم خوشه‌بندی پیک چگالی برای تخصیص نقاط نمونه شامل تخصیص هر نقطه نمونه غیرمرکزی به خوشه نزدیکترین همسایه با چگالی بالاتر است. در مطالعات این حوزه توصیه می‌شود که میانگین تعداد نزدیک-ترین نقاط نمونه همسایه باید حدود ۱٪ تا ۲٪ از کل نقاط نمونه باشد [۱۹]. هنگامی که تراکم نزدیکترین نقاط نمونه همسایه بسیار نزدیک است، تعیین اینکه کدام نقاط نمونه غیرمرکزی باید اختصاص داده شوند می‌تواند چالش برانگیز باشد، که منجر به تخصیص نادرست یا اتکا به نزدیک‌ترین همسایگان منتخب می‌شود. این می‌تواند منجر به مشکل "دومینو" شود، جایی که نقاط نمونه منطقه متراکم به اشتباه به یک

[5] Density Peak Clustering

خوشه اختصاص داده می‌شوند و باعث می شود نزدیکترین همسایگان آن‌ها نیز نادرست تخصیص داده شوند [۱۶].

۲,۲. بهبودهای الگوریتم خوشه‌بندی پیک چگالی

در سال‌های اخیر، تمرکز بهبود در الگوریتم خوشه‌بندی پیک چگالی بر دو حوزه اصلی بوده است: استراتژی تخصیص نقاط نمونه و استراتژی انتخاب مراکز خوشه‌ای. یکی از نمونه‌های الگوریتم خوشه‌بندی پیک چگالی پیشرفته[۶]FKNN-DPC است که رویکرد K-نزدیک‌ترین همسایه فازی را برای بهبود تخصیص نقاط نمونه ترکیب می‌کند. این روش در هنگام تخصیص نقاط نمونه، وابستگی فازی آنها را در نظر می‌گیرد [۱۶]. پیشرفت دیگر[۷]CDFTS-DPC است که یک روش فاصله رتبه مرتبه هسته و یک روش تبدیل و تغییر داده را برای عملکرد بهتر از الگوریتم خوشه‌بندی پیک چگالی اصلی ادغام می‌کند [۲۰].[۸]ADPC-KNN یک الگوریتم خوشه‌بندی پیک چگالی تطبیقی است که چگالی هر نقطه نمونه را محاسبه می‌کند و دسترسی چگالی بین زیر خوشه‌ها را بر اساس چگالی و فاصله محلی تعیین می‌کند [۲۱]. وانگ و همکاران یک الگوریتم خوشه‌بندی پیک چگالی چند مرکزی برای شناسایی خوشه‌هایی با پیک‌های چگالی چندگانه و تراکم مرکز خوشه کم معرفی کرد [۲۲]. محققان دیگر روش‌های مختلفی مانند استفاده از همسایگان طبیعی برای محاسبه مجدد پارامترهای الگوریتم خوشه‌بندی پیک چگالی، بهبود اندازه‌گیری چگالی و شناسایی مرکز خوشه و طراحی الگوریتم‌هایی برای مجموعه‌های داده با چگالی متغیر پیشنهاد کرده‌اند [۲۳]. با این حال، برخی از این رویکردها ممکن است دارای اشکالاتی مانند حساسیت به نویز، هزینه‌های زمانی بیش از حد، یا تولید پیک های چگالی محلی در مجموعه داده‌های متراکم باشند که منجر به افزایش مصرف زمان می‌شود.

یک روش خوشه‌بندی جدید که از همسایگی‌ها و هسته آپولونیوس استفاده می‌کند [۲۴]، معرفی شده است. این الگوریتم جدید شامل چهار مرحله اصلی است: ۱) تعیین همسایه‌های طبیعی، ۲) شناسایی نقاط هسته، ۳) ادغام هسته‌های خوشه و جداسازی آنها از دیگر خوشه‌ها، و ۴) استفاده از هسته آپولونیوس برای تخصیص نقاط باقی مانده به خوشه‌ها و شناسایی نقاط نویز. این روش با تخمین تراکم محلی از طریق همسایگی‌های طبیعی نقاط پر تراکم و طبقه‌بندی نقاط مرزی و خوشه براساس هسته آپولونیوس[۹]ساختار خوشه محلی و سراسری را به طور موثر آشکار می‌کند [۲۶].

---

[6] Robust Clustering by Detecting Density Peaks and Assigning Points based on Fuzzy weighted K-nearest neighbors

[7] CDF Transform-and-Shift: An effective way to deal with datasets of inhomogeneous cluster densities

[8] Adaptive Density Peak Clustering based on K-Nearest Neighbors with aggregating strategy

[9] Apollonius Function Kernel (AFK)

یک روش خوشه‌بندی جدید که بر اصول هندسی و مفهوم همسایگی‌های طبیعی برای ارزیابی چگالی محلی نقاط عمل می‌کند معرفی شده است [۲۵]. در ابتدا، الگوریتم خوشه‌های اولیه را با مشخص کردن نقاط متراکم با تعداد زیادی همسایه طبیعی شناسایی می‌کند. متعاقباً، مرکزهای خوشه‌ای بر اساس معیار چگالی طبیعی محلی تعیین می‌شوند. سپس یک تکنیک منحصربه‌فرد مبتنی بر همسایگی طبیعی برای شناسایی نقاط داده با تعداد کم همسایه‌های طبیعی، که به آنها نقاط ضعیف یا نویز گفته می‌شود، استفاده می‌شود. در نهایت، خوشه های نهایی با حذف این نقاط نویز به دست می‌آیند.

هدف الگوریتم ساخت همسایگی توسط منطقه آپولونیوس[1] شناسایی و ارائه یک الگوی هندسی کاربردی در داده‌ها از طریق داده‌کاوی است [۲۶; ۲۷]. با تجزیه و تحلیل الگوهای هندسی دقیق بر اساس روابط درون داده‌ها در فضای همسایگی، می‌توان با الگوریتم ساخت همسایگی توسط منطقه آپولونیوس روندهای رفتاری و شباهت‌های بین داده‌ها را کشف کرد. این الگوریتم هیچ دانش قبلی از مجموعه داده‌ها را ندارد. هدف مشخص کردن همسایگی دقیق با استفاده از دایره آپولونیوس است که به تعیین وضعیت همسایگی نقاط داده کمک می‌کند. اثربخشی ساختار آپولونیوس در ارزیابی شباهت‌های محلی بین مشاهدات، بعد جدیدی را به حوزه هندسه در داده‌کاوی معرفی کرده است.

روش خوشه‌بندی کوانتومی مبتنی بر دایره آپولونیوس که می‌تواند پهنای باند هسته را برای خوشه‌بندی بدون نیاز به دانش قبلی در مورد نقاط داده یا خوشه‌ها تعیین کند [۲۸]. خوشه‌بندی کوانتومی مبتنی بر دایره آپولونیوس جدیدترین رویکرد برای دستیابی به پهنای باند هسته تطبیقی با استفاده از ساخت همسایگی منطقه آپولونیوس است. تخمین تابع موج کوانتومی بر اساس نقاط داده در گروه همسایگی تشکیل شده توسط دایره های آپولونیوس بصورت بهینه بدست می‌آید و باعث افزایش دقت خوشه‌بندی کوانتومی می‌گردد.

الگوریتم ساخت همسایگی به نام ساخت همسایگی با تراکم منطقه آپولونیوس[2] معرفی شده‌است [۲۹]. این الگوریتم همسایگان نقاط داده را بر اساس ساختارهای هندسی و اطلاعات چگالی تعیین می‌کند. الگوریتم ساخت همسایگی با تراکم منطقه آپولونیوس برای خوشه‌بندی داده‌های با ابعاد بالا مؤثرتر است و می‌تواند داده‌های پرت محلی را شناسایی کند. با شناسایی نقاط داده مشابه با استفاده از دایره‌های آپولونیوس، الگوریتم چگالی و روابط بین نقاط را استخراج می‌کند و در نتیجه یک همسایگی دقیق و متمایز ایجاد می‌کند.

روش پیشنهادی

[1] Geometric-based Clustering method using Natural Neighbors (GCNN)
[1] Neighborhood Construction by Apollonius Region (NCAR)
[1] Neighborhood Construction with Apollonius Region Density (NCARD)

مرحله اول: تشکیل زیرخوشه‌های اصلی

تعریف ۱: K-نزدیکترین همسایه: در یک مجموعه داده X، برای هر نقطه داده x_i، K نزدیکترین همسایه‌ها به صورت زیر تعریف می‌شوند که اشیاء با کمترین فاصله تا x هستند.

(۱)

$$\mathrm{KNN}(x_i) = \left\{ x_j \,\middle|\, d(x_i, x_j) \le d(x_i, (x_i)_k) \right\}$$

تعریف ۲: نزدیکترین همسایگان متقابل: در یک مجموعه داده X، برای هر نقطه داده x_i، نزدیکترین همسایگان متقابل (x_j) هستند، اگر مجموعه‌ای از اشیاء که متقابلاً همسایه یکدیگر هستند. نزدیکترین همسایگان متقابل به صورت زیر تعریف می‌شود:

$$M\ \mathrm{NN}(x_i) = \left\{ x_j \,\middle|\, x_i \in KNN(x_j) \cap x_j \in KNN(x_i) \right\} \qquad (۲)$$

این روش فرآیند خوشه‌بندی را از هر شی x_i آغاز می‌کند. چگالی محلی شی مرجع x_i طبق معادله (۱) با چگالی محلی هر داده x_j در نزدیک‌ترین همسایگان مشترکشان مقایسه خواهد شد. نقاطی که دارای چگالی محلی مشابه با چگالی محلی x_i هستند به شی x_i (آغازگر خوشه) نسبت داده می‌شوند. خوشه بر اساس شباهتی از آنها طبق معادله ۲ بر اساس شباهت چگالی محلی هر جسم در نزدیک‌ترین همسایگان به رشد خود در مرحله بعد ادامه می‌دهد.

تعریف ۳: وزن نقطه x_i، x_j مجموع فواصل نزدیکترین همسایگان مشترک دو طرفه x_i، x_j است $(N_{x_j} \cap N_{x_i})$. این وزن نشان دهنده چگالی محلی نقطه است. نقطه ای که وزن کمتری دارد چگالی بیشتری دارد و بالعکس. N_{x_i} نزدیکترین همسایگان x_i است.

$$w(x_i, x_j) = \frac{\sum d(N_{x_i}, N_{x_j})}{\left| N_{x_i}, N_{x_j} \right|} \qquad for\ all\ \ N_{x_i} \in N_{x_j}\ \ and\ \ N_{x_j} \in N_{x_i} \qquad (۳)$$

مرحله دوم: ایجاد گراف شباهت نقطه مرجع با همسایگان توسط ضریب ارتباطات در ماتریس وابستگی

گام ۱: برای اینکه مسئله را تبدیل به گراف کنیم، پس از بدست آوردن وزن هر داده براساس تعداد و فاصله با همسایگان مشترک دو طرفه (برای استخراج ارتباط تمام داده‌ها با هم) دو ماتریس وابستگی ایجاد می-کنیم. در این ماتریس‌ها نقاط دارای همسایه‌های مشترک مقدارW برایشان در ماتریس اول و مقدار یک در ماتریس دوم وارد خواهد شد و نقاطی که هیچ همسایه مشترک دو طرفه ندارند مقدار صفر برایشان درج خواهد شد. در آخر مرحله دوم از داخل ماتریس وابستگیWها را برای هر نمونه استخراج و از کمترین مقدار یعنی پرچگالترین تا بیشترینWها مرتب می‌کنیم و در هر سطر و برای هر نمونه تعداد یک‌ها جمع می‌گردد تا در مرحله سوم پیمایش و بررسی را از پرچگال‌ترین نقاط شروع کنیم. اشتراک نمونه کم وزن استخراج شده و دارای تعداد یک بالا در هر سطر معرف یک نمونه پرچگال است. در آخر این مرحله پس از استخراج نقاط پرچگال همسایگان آنها از ماتریس استخراج و گراف آنها رسم می‌گردد.

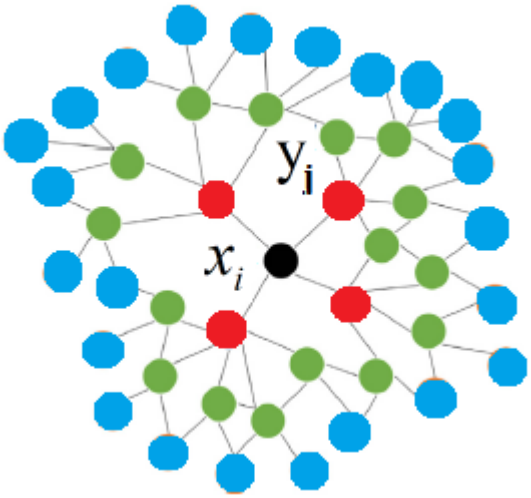
گام ۲: محاسبه شعاع همسایگی بین نقاط پرچگال (ریشه‌ها) و دورترین گره همسایه‌شان: این گره داده‌ی خواهد بود که جز همسایه‌ی همسایگان ریشه است و جز همسایگان سایر ریشه‌ها نیست و فاصله آن تا ریشه از فاصله بین دو گره ریشه کمتر است. دورترین همسایه ریشه‌ها با فرمول زیر محاسبه می‌گردد. FN_{R_i} دورترین نقطه برای ریشه‌ی R_i است. فاصله R_i تا دورترین نقطه از آن هم شعاع همسایگی آن است.

$$FN_{R_i} = \left\{ MKNN \cap KNN = \varnothing \ with\ other\ R \quad and \quad d\left(FN_{R_i}, R_i\right) < d\left(R_i, R_{i+1}\right) \right\} \tag{۴}$$

نکته: نقاط پرت و نویزی نقاطی هستند که در شعاع همسایگی هیچ ریشه‌ی نیستند و فاصله آنها تا ریشه بیشتر از فاصله بین دو ریشه است.

مرحله سوم: بررسی اتصال زیرخوشه‌ها و نقاط بدون خوشه در گراف

جستجوی کوتاه‌ترین مسیر برای نقاط دارای چگالی محلی مشابه درون شعاع همسایگی با روش جستجوی سطح اول و یافتن سرخوشه‌ها بصورت دقیق و ادغام خوشه‌ها با یافتن والد جایگزین.



شکل ۱، نقاطی که با رنگ‌های قرمز، سبز و آبی مشخص شده‌اند، به ترتیب نشان دهنده نقاط بازدید شده در سه تکرار هستند براساس جستجوی سطح اول (وزن یال‌ها) .

در تکرار اول ، این استراتژی از نقطه x_i برای بازدید از چهار نقطه قرمز شروع می‌شود و کوتاه ترین طول مسیر را از x_i تا چهار نقطه ثبت می‌کند. چگالی نقاط قرمز کمتر از x_i است (وزن x_i کمتر است، بنابراین چگالی آن نسبت به نقاط قرمز بیشتر است)، در نتیجه این استراتژی باید این نقاط را که مستقیما با نقاط قرمز در تکرار بعدی مرتبط هستند ، بازدید کند.

$$w_{x_i} < w_{children} \ of \ x_i \ (red \ \ points) \qquad (۵)$$

در تکرار دوم، این استراتژی از این نقاط سبز بازدید می‌کند و کوتاه ترین طول مسیر را از x_i به نقاط سبز ثبت می‌کند. چگالی نقطه y_j بالاتر از x_i است ، بنابراین والد x_i به طور موقت y_j ثبت می‌شود و فاصله نسبی x_i کوتاه‌ترین طول مسیر از x_i تا y_j است.

$$w_{y_i} < w_{x_i}, \quad y_i \quad parent \ x_i \qquad (۶)$$

درتکرار سوم، این استراتژی باید از نقاط آبی بازدید کند و کوتاه‌ترین طول مسیر را از x_i تا نقاط آبی ثبت کند. از آنجا که کوتاه ترین طول مسیر از x_i به هر نقطه آبی بزرگتر از کوتاه ترین طول مسیر از x_i

به y_j است، استراتژی پایان است و y_j والدین x_i است. باقی نقاط بدون خوشه به نزدیکترین سرخوشه و دارای بیشترین همسایگان مشترک نسبت داده خواهد شد.

نتایج تجربی

ارزیابی نتایج خوشه‌بندی براساس تابع فاصله و برچسب‌های تولید شده در تحلیل خوشه‌بندی انجام می‌شود. برای نشان دادن میزان شباهت بین دو شیوه برچسب‌گذاری می‌توان از "شاخص رَند" استفاده کرد. این شاخص توسط دانشمند آمار ویلیام رند[۱۳] در سال ۱۹۷۱ در مقاله‌ای با عنوان «معیارهای هدف برای ارزیابی روش‌های خوشه‌بندی[۱۴] طرفی شد[۱۵] ARI یک معیار متقارن اصلاح شده شاخص رَند است که از -۱ تا +۱ متغیر است. وقتی هر دو جامعه کاملاً متفاوت هستند، مقدار ۱- است و وقتی هر دو جامعه کاملاً مشابه هستند، مقدار ۱+ است. آزمایش‌های مقایسه‌ای برای بررسی اثربخشی و کارایی الگوریتم پیشنهادی در این بخش طراحی شده‌اند. [۳۰] ARI و [۳۱] NMI[۱۶] برای ارزیابی نتایج خوشه‌بندی الگوریتم‌های مختلف بر روی مجموعه داده‌های دنیای واقعی استفاده می‌شوند. اطلاعات متقابل نرمال شده یک معیار محبوب برای کیفیت خوشه‌بندی است که به عنوان اطلاعات متقابل بین انتساب‌های خوشه و یک برچسب-گذاری از قبل موجود از مجموعه داده که با میانگین حسابی حداکثر آنتروپی ممکن حاشیه‌های تجربی محاسبه شده‌است. مقادیر ARI و NMI دارای محدوده به ترتیب [-۱، ۱] و [۰، ۱] هستند، و هر چه مقدار این معیارهای ارزیابی بزرگتر باشد، نتیجه خوشه‌بندی بهتری به دست می‌آید.

Y: نتایج خوشه‌بندی.

C: برچسب‌های خوشه‌بندی واقعی.

n: تعداد نقاط داده.

n_{ij} : تعداد نقاط یکسان در هر دو خوشه c_i و خوشه y_i است.

$n_{i.}$ و $n_{.j}$: به ترتیب تعداد نقاط یکسان y_i و خوشه های c_i .

---

1 William Rand
2 Objective criteria for the evaluation of clustering methods
3 Adjusted Rand Index
4 Normalized Mutual Information

$$ARI(Y, C) = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2} / \binom{n}{2}}{1/2\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \sum_i \binom{n_{i.}}{2}\sum_j \binom{n_{.j}}{2} / \binom{n}{2}} \quad (۷)$$

$$NMI(Y, C) = \frac{MI(Y, C)}{\sqrt{H(Y)H(C)}} \quad (۸)$$

۴،۱. توضیحات مجموعه داده‌ها

الگوریتم پیشنهادی در گوگل کولب با زبان برنامه نویسی پایتون پیاده‌سازی شده‌است و عملکرد آن بر روی چهارده مجموعه داده دنیای واقعی از سایت (Datasets - UCI Machine Learning Repository) UCI [۱] ارزیابی می‌گردد. جدول ۱ جزئیات مربوط به مجموعه داده‌های دنیای واقعی را نشان می‌دهد. دو منبع داده‌های دنیای واقعی وجود دارد، یکی مخزن UCI که شامل مجموعه داده‌های بیولوژیکی و غیره است، و دیگری مخزن دلده KAGLE است. در این آزمایش‌های مقایسه‌ای، ابتدا اثربخشی الگوریتم پیشنهادی را با الگوریتم خوشه‌بندی پیک چگالی و پنج الگوریتم خوشه‌بندی پیک چگالی بهبود یافته بر روی مجموعه داده‌های مصنوعی، UCI آزمایش می‌کنیم. خلاصه‌ای از اطلاعات مجموعه داده‌ها در جدول ۱ قابل نمایش است که در آن N، D، M به ترتیب تعداد نقاط داده، ابعاد و تعداد خوشه‌ها را نشان می‌دهند. از اصل نرمال‌سازی حداقل حداکثر برای پردازش مجموعه داده استفاده می‌شود. برای هر الگوریتم به منظور اطمینان از قابلیت اطمینان آزمایش، با مجموعه‌ای از مقادیر پارامتر آزمایش می‌کنیم و بهترین نتیجه را انتخاب می‌کنیم. برای این الگوریتم‌های با پارامتر k، مقدار k از ۳ تا ۵ انتخاب می‌شود. برای سایر الگوریتم‌های با پارامتر dc، مقدار dc از ۰،۰۴ تا ۱ با گام ۰،۰۱ انتخاب می‌شود.

جدول۱. مجموعه داده‌ها.

Dataset	d	n	m	Type
Iris	4	150	3	Real
Wine	13	178	3	Real
Seed	7	210	3	Real
Ecoil	8	336	8	Real
WDBC	31	569	2	Real
Dermatoloy	34	366	6	Real
Frogs	22	7195	10	Real
Heart	12	270	2	Real
Yeast	8	1484	10	Real
Libras	90	330	15	Real
Segmentation	19	2100	7	Real

[1] https://archive.ics.uci.edu/datasets

Pen digits	16	7494	10	Real
Nursery	8	12960	3	Real
UJIIndoorLoc	520	21048	3	**Real**

جدول۲. شاخص دقت ARI در مجموعه داده‌های واقعی. UCI

dataset	GCNN	NCAR	AFK	FKNN-DPC	NCARD	DPC	Our algorithm
Iris	0.90	۲0.6	0.90	0.88	0.90	0.88	**0.91**
Wine	0.91	0.71	۱0.8	0.86	0.72	0.69	**0.91**
Seed	0.76	0.75	۷0.7	0.76	۷0.7	0.74	**0.78**
Ecoil	0.75	۵۷0.	۷۲0.	0.53	0.72	0.45	**0.75**
WDBC	0.82	۵0.7	0.73	0.69	۲0.5	0.50	**0.83**
Dermatoloy	0.84	0.73	۵0.8	0.79	۶0.7	0.72	**0.84**
Segmentation	0.62	۸0.5	۸0.5	0.65	0.56	0.66	**0.67**
Heart	0.33	0۲0.	5۲0.	0.32	0.28	0.31	**0.37**
Yeast	0.26	۲۸0.	۸0.2	0.07	0.14	0.09	**0.30**
Libras	0.36	۳۸0.	۱0.4	0.35	۹0.3	0.35	**0.42**
Pendigits	0.67	۷0.6	0.67	0.58	۶0.6	0.67	**۸0.6**
Frogs	0.81	۸0.6	۴0.7	0.65	۸0.6	0.72	**0.82**
Nursery	0.40	۶0.2	4۴0.	0.35	0.24	0.51	**0.53**
UJIIndoorLoc	0.40	۶0.5	۲0.6	0.57	3۵0.	0.43	**0.62**
Average	۰,۶۳	۰,۵۵	۰,۶۲	۰,۵۷	۰,۵۶	۰,۵۵	۰,۶۷

درجدول ۲ نتایج روش پیشنهادی با روش‌های نوین در حوزه خوشه‌بندی مقایسه شده است و براساس معیار ارزیابی ARI روش پیشنهادی روی چهارده مجموعه داده واقعی در اغلب داده‌ها دقت بالا یا برابری دارد. بر روی مجموعه داده‌های Wine ، Ecoil ، Dermatoloy دقت روش GCNN به ترتیب ۰,۹۱ ، ۰,۷۵، و ۰,۸۴ برابر با روش پیشنهادی است. بر روی داده UJIIndoorLoc روش AFK دقتی برابر با روش پیشنهادی با مقدار ۰,۶۲ قابل مشاهده‌است. میانگین دقت روش پیشنهادی با مقدار ۰,۶۷ درصد از میانگین بقیه روش‌ها بالاتر است. جستجوی کوتاه‌ترین مسیر برای نقاط دارای چگالی محلی و یافتن کوتاه‌ترین مسیر با اطلاعات درون ماتریس وابستگی می‌تواند اتصال و ارتباط بین نقاط را تضمین کند و همچنین در رسیدن به دقت بالا نسبت به سایر روش‌ها کمک کند.

جدول ۳. NMI در مجموعه داده‌های UCI .

dataset	GCNN	NCAR	AFK	FKNN-DPC	NCARD	DPC	Our algorithm
Iris	0.93	0.68	0.88	0.86	0.89	0.86	**0.93**
Wine	0.89	0.79	0.80	0.84	0.81	0.72	**0.79**
Seed	0.75	0.74	0.75	0.74	0.74	0.72	**0.76**
Ecoil	0.69	0.67	0.67	0.57	0.66	0.59	**0.70**
WDBC	0.72	0.69	0.62	0.64	0.69	0.48	**0.72**
Dermatoloy	0.91	0.86	0.90	0.86	0.89	0.82	**0.91**
Segmentation	0.75	0.72	0.76	0.76	0.75	0.75	**0.77**
Heart	0.33	0.38	0.40	0.28	0.42	0.24	**0.33**
Yeast	0.29	0.30	0.27	0.15	0.27	0.21	**0.32**
Libras	0.63	0.58	0.68	0.64	0.66	0.64	**0.69**
Pen digits	0.78	0.78	0.79	0.77	0.78	0.77	**0.72**
Frogs	0.69	0.69	0.69	0.60	0.69	0.61	**0.70**
Nursery	0.29	0.29	0.33	0.45	0.51	0.64	**0.66**

UJIIndoorLoc	0.55	0.58	0.72	0.70	0.60	0.58	**0.72**
Average	۰,۶۵	۰,۶۲	۰,۶۶	۰,۶۳	۰,۶۴	۰,۶۱	۰,۶۹

جدول ۳ دقت روش پیشنهادی با روش‌های نوین را براساس معیار ارزیابی NMI نشان می‌دهد. براساس معیار ارزیابی NMI روش پیشنهادی روی داده‌های واقعی در اغلب داده‌ها دقت بالای دارد. بر روی مجموعه داده Wine دقت روش GCNN بالاتر از روش پیشنهادی با تفاوت ۱۰ درصد است. بر روی مجموعه داده‌های WDBC و Dermatology دقت روش GCNN به ترتیب ۰.۷۲ ، ۰.۹۱ برابر با روش پیشنهادی است . دقت داده Heart با روش AFK بیشتر از روش پیشنهادی و نزدیک به ۰,۴۵ درصد است که روش پیشنهادی دارای دقت ۰,۳۳ درصد است و همچنین برای داده Pendigits این روش دقت بالای از سایر روش‌ها دارد. بر روی داده UJIIndoorLoc روش AFK دقتی برابر با روش پیشنهادی با مقدار ۰,۷۲ درصد دارد. میانگین دقت روش پیشنهادی براساس معیار ارزیابی NMI با مقدار ۰,۶۹ درصد از میانگین بقیه روش‌ها بالاتر است. جهت کاهش زمان اجرای روش پیشنهادی، جستجوی سطح اول و یافتن کوتاه‌ترین مسیر پیشنهاد شده است که این استراتژی می‌تواند همسایگی اکثر نقاط داده را بدون جستجوی کل مجموعه داده‌ها پیدا کند و زمان اجرا را تا حد زیادی کاهش دهد. ایراد روش پیشنهادی اعمال آن برروی کلان داده‌ها می‌باشد که جستجوی گراف ممکن است باعث پیچیدگی محاسباتی بالای گردد.

نتیجه‌گیری

در این مقاله یک الگوریتم خوشه‌بندی جدید معرفی می‌شود که بر اساس ماتریس وابستگی و جستجوی چگالی درخت، خوشه‌ها را پیدا می‌کند. فضای همسایگی جدید برای تعریف چگالی محلی و سراسری معرفی می‌گردد، همچنین یک روش جستجوی نقاط مرکزی خوشه‌ای استفاده شده است. الگوریتم پیشنهادی یک روش بر اساس فضای مربوطه برای تخصیص سایر داده‌های خوشه‌بندی نشده در مرحله آخر ارائه می‌دهد. نتایج تجربی بر روی مجموعه داده‌های واقعی، عملکرد موثر روش خوشه‌بندی معرفی شده را نشان می‌دهد. آزمایش‌های مقایسه‌ای کارایی الگوریتم پیشنهادی را در مقابل سایر الگوریتم‌های معروف و نوین با اعمال تغییرات با دقت بالایی نشان می‌دهد. برای بهبود این الگوریتم در آینده می‌توان از روش‌های استخراج همسایگی بدون پارامتر (تعیین تعداد همسایگی) مبتنی بر چگالی همانند گراف‌های هم انرژی در تشکیل ساختارهای اولیه خوشه‌ها استفاده نمود.
برای کارهای آتی ترکیب روش پیشنهادی با گراف‌های هم انرژی مورد نظر می‌باشد تا بر اساس آن بتوان روی کلان داده‌ها نیز خوشه‌بندی را با دقت بالا انجام داد.

منابع

[1] Chen, Y., Hu, X., Fan, W., Shen, L., Zhang, Z., Liu, X., Du, J., Li, H., Chen, Y., & Li, H. (2020). *Fast density peak clustering for large scale data based on kNN*. Knowledge-Based Systems, 187, 104824, https://doi.org/10.1016/j.knosys.2019.06.032.

[2] Jan, Z., Ai-Ansari, N., Mousa, O., Abd-Alrazaq, A., Ahmed, A., Alam, T., & Househ, M. (2021). *The role of machine learning in diagnosing bipolar disorder: scoping review*. Journal of medical Internet research, 23(11), e29749, https://preprints.jmir.org/preprint/29749.

[3] Wen, J., Xuan, S., Li, Y., Peng, Q., & Gao, Q. (2020). *Image segmentation algorithm based on neutrosophic fuzzy clustering with non- local information*. IET Image Processing, 14(3), 576-584, https://doi.org/10.1049/iet-ipr.2018.5949.

[4] Zou, Q., Lin, G., Jiang, X., Liu, X., & Zeng, X. (2020). *Sequence clustering in bioinformatics: an empirical study*. Briefings in bioinformatics, 21(1), 1-10, https://doi.org/10.1093/bib/bby090.

[5] Tombros, A., Villa, R., & Van Rijsbergen, C. J. (2002). *The effectiveness of query-specific hierarchic clustering in information retrieval*. Information processing & management, 38(4), 559-582, https://doi.org/10.1016/S0306-4573(01)00048-6.

[6] Jain, A. K. (2008). Data clustering: 50 years beyond k-means. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, https://doi.org/10.1007/978-3-540-87479-9_3.

[7] Jin, H., Leung, K.-S., Wong, M.-L., & Xu, Z.-B. (2005). *Scalable model-based cluster analysis using clustering features*. Pattern Recognition, 38(5), 637-649, https://doi.org/10.1016/j.patcog.2004.07.012.

[8] Murtagh, F., & Contreras, P. (2012). *Algorithms for hierarchical clustering: an overview*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86-97, https://doi.org/10.1002/widm.53.

[9] Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). *Density- based clustering*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3), 231-240, https://doi.org/10.1002/widm.30.

[10] Kumar, K. M., & Reddy, A. R. M. (2016). *A fast DBSCAN clustering algorithm by accelerating neighbor searching using Groups method*. Pattern Recognition, 58, 39-48, https://doi.org/10.1016/j.patcog.2016.03.008.

[11] Tang, C., Wang, H., Wang, Z., Zeng, X., Yan, H., & Xiao, Y. (2021). *An improved OPTICS clustering algorithm for discovering clusters with uneven densities*. Intelligent Data Analysis, 25(6), 1453-1471, https://doi.org/10.3233/IDA-205497.

[12] Hu, L., & Chan, K. C. (2015). *A density-based clustering approach for identifying overlapping protein complexes with functional preferences*. BMC bioinformatics, 16, 1-16, https://doi.org/10.1186/s12859-015-0583-3.

[13] Rodriguez, A., & Laio, A. (2014). *Clustering by fast search and find of density peaks*. science, 344(6191), 1492-1496, DOI: 10.1126/science.1242072.

[14] Guo, W., Wang, W., Zhao, S., Niu, Y., Zhang, Z., & Liu, X. (2022). *Density peak clustering with connectivity estimation*. Knowledge-Based Systems, 243, 108501, https://doi.org/10.1016/j.knosys.2022.108501.

[15] Xu, X., Ding, S., & Shi, Z. (2018). *An improved density peaks clustering algorithm with fast finding cluster centers*. Knowledge-Based Systems, 158, 65-74, https://doi.org/10.1016/j.knosys.2018.05.034.

[16] Xie, J., Gao, H., Xie, W., Liu, X., & Grant, P. W. (2016). *Robust clustering by detecting density peaks and assigning points based on fuzzy weighted K-nearest neighbors*. Information Sciences, 354, 19-40, https://doi.org/10.1016/j.ins.2016.03.011.

[17] Xu, X., Ding, S., Wang, L., & Wang, Y. (2020). *A robust density peaks clustering algorithm with density-sensitive similarity*. Knowledge-Based Systems, 200, 106028, https://doi.org/10.1016/j.knosys.2020.106028.

[18] Hou, J., Zhang, A., & Qi, N. (2020). *Density peak clustering based on relative density relationship*. Pattern Recognition, 108, 107554, https://doi.org/10.1016/j.patcog.2020.107554.

[19] Tong, W., Liu, S., & Gao, X.-Z. (2021). *A density-peak-based clustering algorithm of automatically determining the number of clusters*. Neurocomputing, 458, 655-666, https://doi.org/10.1016/j.neucom.2020.03.125.

[20] Zhu, Y., Ting, K. M., Carman, M. J., & Angelova, M. (2021). *CDF Transform-and-Shift: An effective way to deal with datasets of inhomogeneous cluster densities*. Pattern Recognition, 117, 107977, https://doi.org/10.1016/j.patcog.2021.107977.

[21] Yaohui, L., Zhengming, M., & Fang, Y. (2017). *Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy*. Knowledge-Based Systems, 133, 208-220, https://doi.org/10.1016/j.knosys.2017.07.010.

[22] Wang, Y., Wang, D., Zhang, X., Pang, W., Miao, C., Tan, A.-H., & Zhou, Y. (2020). *McDPC: multi-center density peak clustering*. Neural Computing and Applications, 32, 13465-13478, https://doi.org/10.1007/s00521-020-04754-5.

[23] Ding, S., Du, W., Xu, X., Shi, T., Wang, Y., & Li, C. (2023). *An improved density peaks clustering algorithm based on natural neighbor with a merging strategy*. Information Sciences, 624, 252-276, https://doi.org/10.1016/j.ins.2022.12.078.

[24] Pourbahrami, S., Balafar, M. A., & Khanli, L. M. (2023). *ASVMK: A novel SVMs Kernel based on Apollonius function and density peak clustering*. Engineering Applications of Artificial Intelligence, 126, 106704, https://doi.org/10.1016/j.engappai.2023.106704.

[25] Pourbahrami, S., & Hashemzadeh, M. (2022). *A geometric-based clustering method using natural neighbors*. Information Sciences, 610, 694-706, https://doi.org/10.1016/j.ins.2022.08.047.

[26] Pourbahrami, S., Balafar, M. A., Khanli, L. M., & Kakarash, Z. A. (2020). *A survey of neighborhood construction algorithms for clustering and classifying data points*. Computer Science Review, 38, 100315, https://doi.org/10.1016/j.cosrev.2020.100315.

[27] Pourbahrami, S., Khanli, L. M., & Azimpour, S. (2019). *A novel and efficient data point neighborhood construction algorithm based on Apollonius circle*. Expert Systems with Applications, 115, 57-67, https://doi.org/10.1016/j.eswa.2018.07.066.

[28] Abdolmaleki, N., Khanli, L. M., Hashemzadeh, M., & Pourbahrami, S. (2022). *ACQC: Apollonius Circle- based Quantum Clustering*. Journal of Computational Science, 64, 101877, https://doi.org/10.1016/j.jocs.2022.101877.

[29] Pourbahrami, S., Khanli, L. M., & Azimpour, S. (2020). *Improving neighborhood construction with Apollonius region algorithm based on density for clustering*. Information Sciences, 522, 227-240, https://doi.org/10.1016/j.ins.2020.02.049.

[30] Sundqvist, M., Chiquet, J., & Rigaill, G. (2023). *Adjusting the adjusted Rand Index: A multinomial story*. Computational Statistics, 38(1), 327-347, https://doi.org/10.1007/s00180-022-01230-7.

[31] Amelio, A., & Pizzuti, C. (2017). *Correction for closeness: Adjusting normalized mutual information measure for clustering comparison*. Computational Intelligence, 33(3), 579-601, https://doi.org/10.1111/coin.12100.