خوشه‌بندی داده‌ها بر پایهٔ شناسایی کلید

احسان فضل ارثی* و مسعود کاظمی نوقابی
گروه مهندسی کامپیوتر، دانشگاه فردوسی مشهد، مشهد، ایران

چکیده

خوشه‌بندی یکی از عناصر اصلی سازنده در بینایی رایانه و یادگیری ماشین است. چالش اصلی، یافتن راهی مناسب برای پیداکردن زیرمجموعه‌ای از نمونه‌های شاخص و ساختارهای خوشه‌ای مرتبط با آنها، با درنظرگرفتن یک معیار فاصله دوبه‌دو، است. در این مقاله شیوه‌ای جدید برای خوشه‌بندی پیشنهاد می‌شود که به‌صورت تکرارشونده، عناصر کلیدی یک مجموعه داده‌ای را بر پایهٔ یک تابع هدف مناسب، پیدا مـی‌کند. آزمایش‌های تجربی متعدد بیان‌گر برتری روش پیشنهادشده نسبت به روش‌های موجود، هم از نظر بهینگی و هم از نظر مؤثربودن، است. علاوه‌براین، روش پیشنهادی برای خوشه‌بندی داده‌های با مقیاس بالا توسعه داده می‌شود؛ به‌صورتی‌که میلیون‌ها داده را در چند ثانیه می‌توان پردازش کرد.

واژگان کلیدی: خوشه بندی؛ شناسایی کلید؛ مقیاس بالا.

Data Clustering Based On Key Identification

Ehsan Fazl-Ersi*, Masoud Kazemi Nooghabi
Department of Computer Engineering,
Ferdowsi University of Mashhad, Mashhad, Iran
Email: fazlersi@um.ac.ir

Abstract

Clustering has been one of the main building blocks in the fields of machine learning and computer vision. Given a pair-wise distance measure, it is challenging to find a proper way to identify a subset of representative exemplars and its associated cluster structures. Recent trend on big data analysis poses a more demanding requirement on new clustering algorithm to be both scalable and accurate. A recent advance in graph-based clustering extends its ability to millions of data points by massive utility of engineering endeavor and parallel optimization. However, most other existing clustering algorithms, though promising in theory, are limited in the scalability issue.

In this paper, a novel clustering method is proposed that is both accurate and scalable. Based on a simple criteria, "key" items that are representative of the whole data set are iteratively selected and thus form associated cluster structures. Taking input of pairwise distance measure between data instances, the proposed method searches centers of clusters by identifying data items far away from selected keys, but representative of unselected data items. Inspired by hierarchical clustering, small clusters are iteratively merged until a desired number of clusters are obtained. To solve the scalability problem, a novel tracking table technique is designed to reduce the time complexity which is capable of clustering millions of data points within a few minutes.

* نویسندهٔ عهده‌دار مکاتبات

* Corresponding author

To assess the performance of the proposed method, several experiments are conducted. The first experiment tests the ability of our algorithm on different manifold structures and various number of clusters. It is observed that our clustering algorithm outperforms existing alternatives in capturing different shapes of data distributions. In the second experiment, the scalability of our algorithm to large scale data points is assessed by clustering up to one million data points with dimensions of up to 100. It is shown that, even with one million data points, the proposed method only takes a few minutes to perform clustering. The third experiment is conducted on the ORL database, which consists of 400 face images of 40 individuals. The proposed clustering method outperforms the compared alternatives in this experiment as well. In the final experiment, shape clustering is performed on the MPEG-7 dataset, which contains 1400 silhouette images from 70 classes, 20 different shapes for each class. The goal here is to cluster the data items (here the binary shapes) into 70 clusters, so that each cluster only includes shapes that belong to one class. The proposed method outperforms other alternative clustering algorithms on this dataset as well.

Extensive empirical experiments demonstrate the superiority of the proposed method over existing alternatives, in terms of both effectiveness and efficiency. Furthermore, our algorithm is capable of large-scale data clustering where millions of data points can be clustered in a few seconds.

Keywords: **Clustering; Key Identification; Large Scale**

۱- مقدّمه

در بسیاری از شاخه‌های هوش مصنوعی، خوشه‌بندی[1] یکی از عناصر سازندهٔ اصلی است. خوشه‌بندی، فرآیندی است که در آن مجموعه‌ای از داده‌ها به گروه‌های متفاوتی (که خوشه[2] نامیده می‌شوند) افراز می‌شوند؛ به‌طوری‌که داده‌های موجود در یک گروه در ویژگی‌ها یا خصوصیت‌هایی اشتراک دارند که ممکن است در داده‌های دیگر گروه‌ها موجود نباشند. خوشه‌بندی می‌بایست ساختار ذاتی موجود در داده‌ها را شناسایی کند تا منجر به درک بهتری از داده‌ها شود. یک روش خوشه‌بندی خوب باید نسبت به تغییرات برون و درون‌طبقه‌ای[3] مقاوم باشد؛ بدین معنـــا، داده‌هایی که به یک طبقه تعلق دارند، باید فاصلهٔ کمی با یکدیگر داشته باشند؛ درنتیجه در خوشه‌های مشابهی گروه‌بندی شوند، و داده‌هایی که به طبقه‌های متفاوتی تعلق دارند، باید فاصله زیادی با یکدیگر داشته باشند و بنابراین در خوشه‌های متفاوتی قرار گیرند.

خوشه‌بندی کاربردهای زیادی ازجمله در تقطیع تصاویر[4] [4]، شناسایی الگو[5] [11]، تحلیل سند[6] [22] و تحلیل بازار[7] [23] دارد. یک نمونهٔ شاخص آن، تشخیص محصولات مشابه در بین حجم عظیمی از محصولات است که در پیشنهاد محصول[8] و یا سامانه‌های بازیابی[9] مورد استفاده قرار می‌گیرد [4].

خوشه‌بندی فرآیندی چالشی است؛ زیرا داده‌ها به‌طورمعمول شامل نوفه و داده‌های پرت می‌باشند که شناسایی و حذف آنها می‌تواند مشکل باشد [9]. یکی از رویکردهای اصلی برای حل این مسالهٔ خوشه‌بندی بر پایه اتصال[10] است که به‌عنوان ورودی، فاصله یا شباهت‌های دوبه‌دو بین نقاط داده‌ای را دریافت می‌کند و براساس این اصل اساسی پی‌ریزی شده است که داده‌هایی که فاصله کمی (شباهت زیاد) دارند به‌طورمعمول در یک خوشه قرار می‌گیرند. یک نمونه شاخص از این رویکرد، خوشه‌بندی سلسله‌مراتبی[11] [10] است که خوشه‌های متفاوتی را در سطوح مختلف فاصله‌ای تشکیل می‌دهد که منجر به یک نمایش دندروگرام[12] از داده می‌شود. یکی دیگر از روش‌های خوشه‌بندی معروف، انتشار وابستگی[13] [6] است که استنباط عبور پیام[14] را بر روی شباهت های دوبه‌دو بین داده‌ها انجام

[1] Clustering
[2] Cluster
[3] Intra- and inter-class variations
[4] Image segmentation
[5] Pattern discovery
[6] Document analysis

[7] Market research
[8] Product recommendation
[9] Retrieval systems
[10] Connectivity-based clustering
[11] Hierarchical clustering
[12] Dendrogram
[13] Affinity propagation
[14] Message-passing inference

می‌دهد. این روش قـــادر است، داده‌های شاخص را از مجموعه داده‌ها انتخاب و به‌صورت خودکار تعداد بهینه خوشه‌ها را تعیین کند. دیگر روش‌های خوشه‌بندی شامل روش‌های بر پایه مرکز[1] مانند کِی-میانگین[2] [21]،[8]، روش‌های بر پایه توزیع[3] مانند مدل مخلوط گوسی[4] [18]،[27]، روش‌های بر پایه گراف[5] مانند خوشه‌بندی طیفی[6] [24] و روش‌های مبتنی بر الگوریتم‌های تکاملی [1] هستند.

روند اخیر در تحلیل داده‌های عظیم[7] یک نیاز جدّی را بر الگوریتم‌های جدید خوشه‌بندی قرار می‌دهد که همان مقیاس‌پذیری و دقت بالا هستند. درنتیجه چند گونه از خوشه‌بندی کِی-میانگین مقیاس بالا به‌جهت سادگی آن پدیدار شدند [2]،[20]؛ با این وجود، این نسخه‌های سریع از کِی-میانگین از کاهش قابل ملاحظه‌ای از دقت به‌جهت افزایش سرعت رنج می‌برند. پیشرفت تازه‌ای در خوشه‌بندی بر پایه گراف قابلیت آن را به میلیون‌ها داده گسترش می‌دهد که در آن به میزان قابل توجهی از بهینه‌سازی موازی[8] استفاده می‌شود. بسیاری از روش‌های موجود خوشه‌بندی با این که در تئوری نویدبخش هستند، از منظر مقیاس‌پـذیری محدود می‌باشند.

در این مقاله یک روش خوشه‌بندی جدید پیشنهاد می‌شود که هم مقیاس‌پذیر و هم دقیق است. بر اساس یک معیار ساده، مرتباً عناصر «کلیدی» که نماینده کل مجموعه داده‌ها هستند، انتخاب می‌شوند و درنتیجهٔ ساختار خوشه‌ای مرتبط با آنها شکل می‌گیرند. با دریافت فاصله بین هر جفت از داده‌ها به‌عنوان ورودی، الگوریتم پیشنهادی در هر مرحله مرکز خوشه‌های جدید را با شناسایی داده‌هایی که بیشترین فاصله را از کلیدهای انتخاب‌شده تا آن مرحله دارند و نمایندهٔ داده‌هایی که انتخاب نشده‌اند، می‌باشند، تعیین می‌کند. با الهام از خوشه‌بندی سلسله‌مراتبی، مرتباً خوشه‌های کوچک با یکدیگر ادغام می‌شوند تا تعداد مطلوبی از خوشه‌ها به‌دست آیند. برای حل مسئله مقیاس‌پذیری، یک شیوه جدول ردیابی[9] جدید پیشنهاد می‌شود که هم از پیچیدگی زمانی می‌کاهد و هم خود را با میلیون‌ها داده در بازهٔ زمانی بسیار کوتاه تطبیق می‌دهد.

[1] Centroid-based
[2] K-means
[3] Distribution-based
[4] Guassian Mixture Models (GMM)
[5] Graph-based
[6] Spectral clustering
[7] Big data analysis
[8] Parallel optimization
[9] Tracking table

ساختار ادامه مقاله بدین قرار است: در بخش بعدی، جزئیات روش خوشه‌بندی پیشنهادی معرفی می‌شود؛ سپس نتایج مقدماتی چند آزمایش خوشه‌بندی ارائه و نکاتی در رابطه با مقایسه الگوریتم خوشه‌بندی پیشنهادی با روش‌های برتر موجود ذکر می‌شوند. و پایان این مقاله، دست‌آوردهای روش خوشه‌بندی پیشنهادی بیان می‌گردند.

۲- روش

۲-۱- ورودی و ساختار

الگوریتم خوشه‌بندی پیشنهادی به‌عنوان ورودی یک مجموعه داده که قرار است خوشه‌بندی شود و یک معیار توقف که فرآیند خوشه‌بندی را هرگاه یک شرط معین برقرار باشد، خاتمه می‌دهد، دریافت می‌کند. انتخاب‌های مختلف برای معیار توقف شامل تعداد مطلوب خوشه و یا یک حد آستانه برای فاصله یا شباهت است. جهت سادگی در توضیح روش، در ادامهٔ این مقاله فرض می‌شود که معیار توقف تعداد مطلوبی از خوشه‌ها است؛ هرچند دیگر معیارهای توقف نیز می‌توانند استفاده شوند. ازایـن‌پـس، تعداد مطلوب خوشه‌ها با C_{target} نمایش داده می‌شود. مجموعه داده‌ها جهت خوشه‌بندی، فهرستی از داده‌ها است که معرف اشیایی هستند که قرار است، خوشه‌بندی شوند؛ به‌طوری‌که یک شیء هر چیزی است که بتوان آن را به‌صورت دیجیتالی با یک فرمت معین نمایش داد؛ به‌نحوی‌که بتوانیم فاصلهٔ بین آنها را محاسبه کنیم. همچنین ازایـن‌پـس، تعداد داده‌های موجود در مجموعه داده را با N نمایش می‌دهیم. الگوریتم خوشه‌بندی پیشنهادی با دریافت $D_{original}$ که یک ماتریس فاصله $N \times N$ است که برای مجموعه داده‌ای محاسبه شده است، آغاز می‌شود. به‌صورت کلی، ماتریس فاصله برای فهرستی از m داده، یک ماتریس حقیقی و متقارن $m \times m$ است؛ به‌طوری‌که مؤلفهٔ سطر i ام و ستون j ام ماتریس، فاصلهٔ بین داده‌های i ام و j ام فهرست است:

$$D_{original}[i,j]=dist(i,j) \qquad (۱)$$

در اینجا، $dist(.,.)$ تابعی است که فاصله بین دو داده از مجموعه داده‌ها را بازمی‌گرداند؛ بنابراین، لازم است، داده‌ها متعلق به فضایی باشند که در آن فاصله به‌صورتی معنادار تعریف شده باشد. اگر داده‌های موجود در مجموعه داده‌ای بتوانند به‌صورت عناصر یک فضای برداری با ابعاد متناهی (مانند نقطه‌های روی یک صفحه دوبعدی استاندارد) بازنمایی

شوند؛ آن‌گاه مفهوم طبیعی فاصلهٔ اقلیدسی کفایت می‌کند. درحالتی‌که بازنمایی‌های دیگری استفاده شوند که دارای پیچیدگی بیشتری هستند، توابع فاصله‌ای دیگری می‌توانند به‌کار گرفته شوند. برای مثال، در برخی از بازنمایی‌ها، فاصله بر پایه تابع کسینوسی انتخاب مناسبی و برای برخی دیگر از بازنمایی‌ها، فاصله همینگ[1] می‌تواند مناسب باشد. یکی از مستقیم‌ترین روش‌های تولید $D_{original}$، محاسبه فاصلهٔ دوبه‌دوی تمامی عناصر موجود در مجموعه داده‌ای و به‌روزرسانی مؤلفهٔ متناظر آنها در ماتریس، با مقادیر به‌دست‌آمده است. بسته به طبیعت مجموعه داده‌ای و تابع فاصله، دیگر گزینه‌های پیاده‌سازی نیز ممکن است، وجود داشته باشند. با فرض دراختیارداشتن $D_{original}$، گام بعدی این است که برای هر عنصر مانند i در مجموعه داده‌ای، مجموعه $R_k(i)$ شامل k عنصر از نزدیک‌ترین همسایه‌های[2] i، شناسایی شود. عدد k در اینجا یک پارامتر ثابت الگوریتم خوشه‌بندی است و می‌تواند هر مقداری از صفر تا N-۱ را اختیار کند. بنابراین مجموعه $R_k(i)$ برای عنصر i در مجموعه داده‌ای، k عنصر از مجموعه داده‌ای به‌جز i هستند که فاصلهٔشان از i کمترین است. این همسایه‌ها با استفاده از مقادیر موجود در ماتریس فاصله یعنی $D_{original}$ می‌توانند شناسایی شوند. علاوه‌بر متغیرها و نشانه‌گذاری که در بالا معرفی شد، نشانه‌گذاری‌ها و تعاریفی که در ادامه می‌آیند نیز در طول مقاله استفاده خواهند شد. L فهرستی از N عدد است که بیان‌گر برچسب خوشه برای هر عنصر است. به‌عبارت‌دیگر برای هر عنصر i، $L[i]$ بیانگر خوشه‌ای است که i به آن تعلق دارد. در ابتدا $L=[1,2,3,\ldots,N]$ برقرار است که نشان‌دهندهٔ این حقیقت است که الگوریتم در آغاز، هر عنصر موجود در مجموعهٔ داده‌ای را به‌عنوان یک خوشه مجزا که حاوی تنها یک عنصر است، در نظر می‌گیرد. در طی اجرای الگوریتم، خوشه‌های کوچکتر با یکدیگر ادغام می‌شوند تا خوشه‌های بزرگتری را شکل دهند و مقادیر برچسب‌ها نیز به طبع به‌روزرسانی می‌شوند تا انعکاس‌دهندهٔ این فرآیند باشند؛ بنابراین، هر چند که تعداد مؤلفه‌های موجود در فهرست L برابر با N باقی خواهد ماند، ولی تعداد مؤلفه‌های منحصربه‌فرد در فهرست، متناظر با تعداد فعلی خوشه‌ها است و در حالت کلی، این مقدار در طول اجرای الگوریتم غیر صعودی خواهد بود. درنهایت g پارامتری ثابت است که تعیین می‌کند، چند تکرار نیاز است تا الگوریتم موفق به پیداکردن خوشه‌های نهایی شود. مقدار g بر روی دقت و پیچدگی زمانی الگوریتم تأثیر می‌گذارد و هر عدد حقیقی بزرگتر از ۱/۰ می‌تواند باشد.

[1] Hamming distance
[2] Nearest neighbors

۲-۲- انتخاب عناصر کلیدی

یک بخش مهم از روش خوشه‌بندی پیشنهادی، الگوریتمی جهت شناسایی تعداد معینی از عناصر کلیدی در مجموعه داده‌ای است. هر عنصر کلیدی یک داده یا یک خوشه میانی شامل تعدادی داده می‌تواند باشد. درنتیجه برخلاف برخی از روش‌های دیگر خوشه‌بندی، کلیدها یا مراکز خوشه‌ها عضو مجموعه داده‌ای هستند. با داشتن ماتریس فاصلهٔ D با اندازهٔ $m\times m$ که فاصلهٔ دوبه‌دوی m عنصر موجود در مجموعه داده‌ای را ذخیره کرده است و تعداد مطلوب عناصر کلیدی که آن را با c نمایش می‌دهیم، یک مجموعه از c عدد صحیح تولید می‌شوند که متناظر با اندیس‌های عناصری هستند که به‌عنوان کلید شناسایی می‌شوند. برای انتخاب عناصر کلیدی نیاز به یک یا چند تابع هدف مناسب است. در این روش، انتخاب عناصر کلیدی با شناسایی عنصری آغاز می‌شود که به‌صورت میانگین، کمترین فاصله را با هر یک از دیگر عناصر مجموعه داده‌ای دارد (تابع هدف ۱):

$$I_1=\arg\min_{1\le i\le m}\frac{1}{m}\sum_{j=1}^{m}D[i,j] \qquad (۲)$$

انتخاب عناصر کلیدی سپس به صورت تکراری با جستجو برای عنصر کلیدی بعدی، I_n، که بیشترین کمینه فاصله را با عناصری که تا به حال انتخاب شده‌اند، دارد (تابع هدف ۲)، ادامه می‌یابد:

$$I_n=\arg\max_{I_k\in K_n}\min_{I_j\in S_n}D[I_k,I_j] \qquad (۳)$$

در اینجا S_n مجموعه عناصری است که در تکرار n ام تا به حال انتخاب شده‌اند و K_n مجموعه عناصری است که در تکرار n ام هنوز انتخاب نشده‌اند. فرآیند انتخاب عناصر کلیدی، زمانی پایان می‌یابد که تعداد عناصر کلیدی انتخاب‌شده برابر با c شود.

۲-۳- خوشه‌بندی

با داشتن یک مجموعه داده‌ای، یک معیار توقف مانند C_{target} (تعداد مطلوب خوشه‌ها) و یک الگوریتم برای شناسایی عناصر کلیدی در یک مجموعه داده‌ای، الگوریتم خوشه‌بندی پیشنهادی با درنظرگرفتن مجموعه داده‌ای به‌عنوان

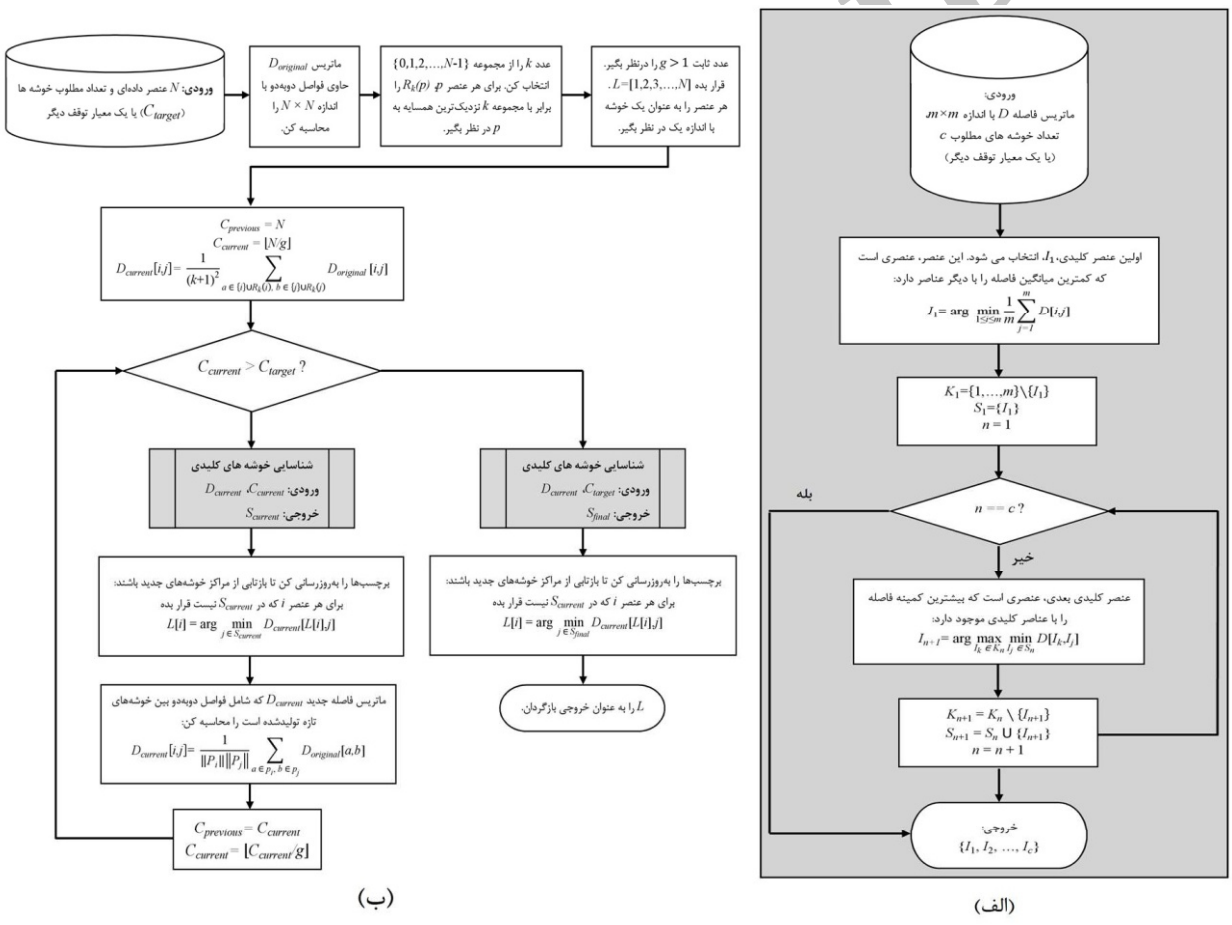
مجموعه‌ای از N خوشه که هر یک شامل تنها یک عنصر است، آغاز می‌شود. الگوریتم به‌صورت تکراری با ادغام‌کردن خوشه‌های کوچکتر با یکدیگر و تولید خوشه‌های بزرگتر ادامه می‌یابد. این ادغام‌ها به‌صورت مکرر اتفاق می‌افتند؛ بنابراین ممکن است چندین مرحله ادغام صورت بگیرد تا به‌طوردقیق C_{target} خوشه باقی بماند. در هر تکرار، $C_{previous}$ نمایش‌دهندهٔ تعداد خوشه‌ها در شروع آن تکرار است و $C_{current}$ تعداد مطلوب خوشه‌هایی است که در پایان آن تکرار به‌دست آمده است. این مقادیر در نخستین تکرار به‌صورت زیر مقداردهی اولیه می‌شوند:

$$C_{previous}=N,\ C_{current}=\lfloor N/g \rfloor \qquad (۴)$$

هم تعداد خوشه‌ها و هم فاصله بین خوشه‌ها به‌صورت پیوسته در طی اجرای الگوریتم، تغییر می‌کنند. یک ماتریس فاصله، $D_{current}$، فاصله بین خوشه‌های فعلی را در طول اجرای الگوریتم نگهداری می‌کند و به‌صورت زیر مقداردهی اولیه می‌شود:

$$D_{current}[i,j]=\frac{1}{(k+1)^2}\times \sum_{a\in\{i\}\cup R_k(i),\ b\in\{j\}\cup R_k(j)} D_{original}[a,b] \qquad (۵)$$

در اینجا، $R_k(i)$ و $R_k(j)$ به مجموعه‌های k عضوی از نزدیک‌ترین همسایه‌های عناصر i و j اشاره دارند که در بخش ۲-۱ معرفی شدند. درحقیقت رابطه بالا میزان فاصله بین دو خوشه اولیه i و j را برابر با میانگین فاصله دوبه‌دو بین تمامی k نزدیک‌ترین همسایگان i و j قرار می‌دهد. در طول اجرای الگوریتم، گام‌های زیر مرتب تا زمانی که $C_{current} > C_{target}$ برقرار است، تکرار می‌شوند:



(شکل-۱): (الف) روندنمای فرآیند انتخاب عناصر کلیدی را نشان می‌دهد؛ (ب) روندنمای کامل روش خوشه‌بندی پیشنهادی را نشان می‌دهد.

(Figure-1): (A) is the flowchart of the process of key item selection; (B) shows the full flowchart of our clustering method.

گام ۱) شناسایی عناصر کلیدی: الگوریتم انتخاب عناصر کلیدی (که در بخش ۲-۲ توصیف شد) بر روی ماتریس فاصله $D_{current}$ اعمال می‌شود تا به تعداد $C_{current}$ عنصر کلیدی شناسایی و سپس در مجموعه $S_{current}$ ذخیره شوند.

گام ۲) به‌روزرسانی برچسب‌های خوشه‌بندی: هر عنصر موجود که به‌عنوان عنصر کلیدی انتخاب نشده بود (چه تنها

یک عنصر داده‌ای باشد و چه یک خوشه میانی از عناصر داده‌ای باشد) با نزدیک‌ترین عنصر کلیدی به خود، که آن را از طریق ماتریس فاصله فعلی یعنی $D_{current}$ می‌توان شناسایی کرد، ادغام می‌شود. ادغام، با به‌روزرسانی برچسب‌های خوشه موجود در L برای عناصری که عضو خوشه ادغام‌شونده هستند، صورت می‌پذیرد. به‌طور دقیق‌تر، برای هر عنصر داده‌ای مانند i که عضو یک خوشه انتخاب‌شده نباشد (یعنی $L[i] \notin S_{current}$)، به‌روزرسانی زیر صورت می‌پذیرد:

$$L[i] = \arg \min_{j \in S_{current}} D_{current}[L[i],j] \qquad (۶)$$

در صورت لزوم، خوشه‌ها دوباره برچسب‌گذاری می‌شوند تا L شامل عناصری از مجموعهٔ $\{1,2,3,\ldots, C_{current}\}$ باشد.

گام ۳) به‌روزرسانی ماتریس فاصله: ماتریس فاصله $D_{current}$، جهت ذخیره فواصل بین $C_{current}$ خوشه جدید که درنتیجهٔ ادغام‌های صورت‌پذیرفته درگام قبلی به وجود آمده‌اند، به‌روزرسانی می‌شود. فرض کنید P_i نمایان‌گر مجموعه‌ای است که شامل همه عناصر عضو خوشه i به‌علاوه تمامی همسایگان آنها باشد؛ به عبارتی:

$$P_i=\{y : L[y] = i\} \cup \{ \bigcup_{j \in \{y: L[y] = i\}} R_k(j)\} \qquad (۷)$$

مؤلفهٔ سطر iام و ستون j ام ماتریس فاصله به‌روزرسانی‌شده با اندازهٔ $C_{current} \times C_{current}$ به‌صورت زیر محاسبه می‌شود:

$$D_{current}[i,j]= \frac{1}{\|P_i\|\|P_j\|} \times \sum_{a \in p_i, b \in p_j} D_{original}[a,b] \qquad (۸)$$

که در اینجا $\|P_i\|$ و $\|P_j\|$ به‌ترتیب بیان‌گر تعداد عناصر موجود در P_i و P_j می‌باشند. درحقیقت فاصله بین دو خوشه i و j برابر با میانگین فاصله دوبه‌دو بین تمامی عناصر موجود در P_i و P_j قرار داده می‌شود. بدین صورت تمامی نقاط یک خوشه و نزدیک‌ترین همسایگان آن‌ها نیز در اندازه‌گیری فاصله آن خوشه با دیگر خوشه‌ها تأثیر داده می‌شوند و درنتیجه فاصلهٔ اندازه‌گیری‌شده، دقیق‌تر خواهد بود.

گام ۴) به‌روزرسانی مقادیر $C_{previous}$ و $C_{current}$: این گام به شکل زیر صورت می‌پذیرد:

$$C_{previous}= C_{current},\ C_{current}= \lfloor C_{current}/g \rfloor \qquad (۹)$$

چهار گام بالا به‌صورت تکراری تا زمانی که $C_{current} \le C_{target}$ برقرار شود، اجرا می‌شوند. اگر مقدار $C_{current}$ در آخرین تکرار برابر با C_{target} نباشد، فرآیند چهار مرحله‌ای بالا برای یک بار دیگر با قراردادن $C_{current}= C_{target}$ اجرا می‌شود.

۳- خوشه‌بندی مقیاس بالا

در این بخش، روش خوشه‌بندی پیشنهادی توسعه داده می‌شود تا داده‌های با مقیاس بالا را نیز دربر بگیرد. در روش توصیف‌شده تا به اینجا، گلوگاه اصلی بخش انتخاب عناصر کلیدی است. لازمهٔ آن مرتب‌کردن فواصل N عنصر است که دست‌کم به $O(N \log N)$ زمان نیاز دارد (با فرض استفاده از بهترین الگوریتم مرتب‌سازی). ازآن‌جایی‌که چهار گام روش خوشه‌بندی $O(\log_g N)$ تکرار می‌شوند و در هر تکرار پیچیدگی زمانی برابر با $O(CN^2 \log N)$ می‌باشد، بنابراین، پیچیدگی زمانی کلی الگوریتم برابر با $O(CN^2 \log N \log_g N)$ است. این بار محاسباتی که از درجه دوم است، اعمال الگوریتم پیشنهادی بر روی میلیون‌ها داده را محدود می‌کند. با این حال، با درنظرگرفتن این که الگوریتم پیشنهادی تنها مقادیر کمینه بین فاصله عناصر کلیدی انتخاب‌شده و سایر عناصر انتخاب‌نشده را جستجو می‌کند، یک جدول به طول N را می‌توان تشکیل داد که برای هر عنصر مقادیر کمینه تا عناصر انتخاب‌شده تا اینجای کار را، در خود ذخیره کند. تنها زمانی این جدول نیاز به به‌روزرسانی دارد که عنصر کلیدی جدیدی انتخاب شده باشد و این به‌روزرسانی تنها محدود به عناصری است که فاصله آنها تا این عنصر کلیدی جدید، از مقادیر کمینه قبل از این تکرار، کمتر باشد. بدین صورت، به طرز چشم‌گیری پیچدگی زمانی را به $O(N)$ در بخش انتخاب عناصر کلیدی، می‌توان کاهش داد. بنابراین، پیچدگی زمانی کلی الگوریتم به $O(CN \log_g N)$ کاهش می‌یابد که برای خوشه‌بندی داده‌های مقیاس بالا مناسب است.

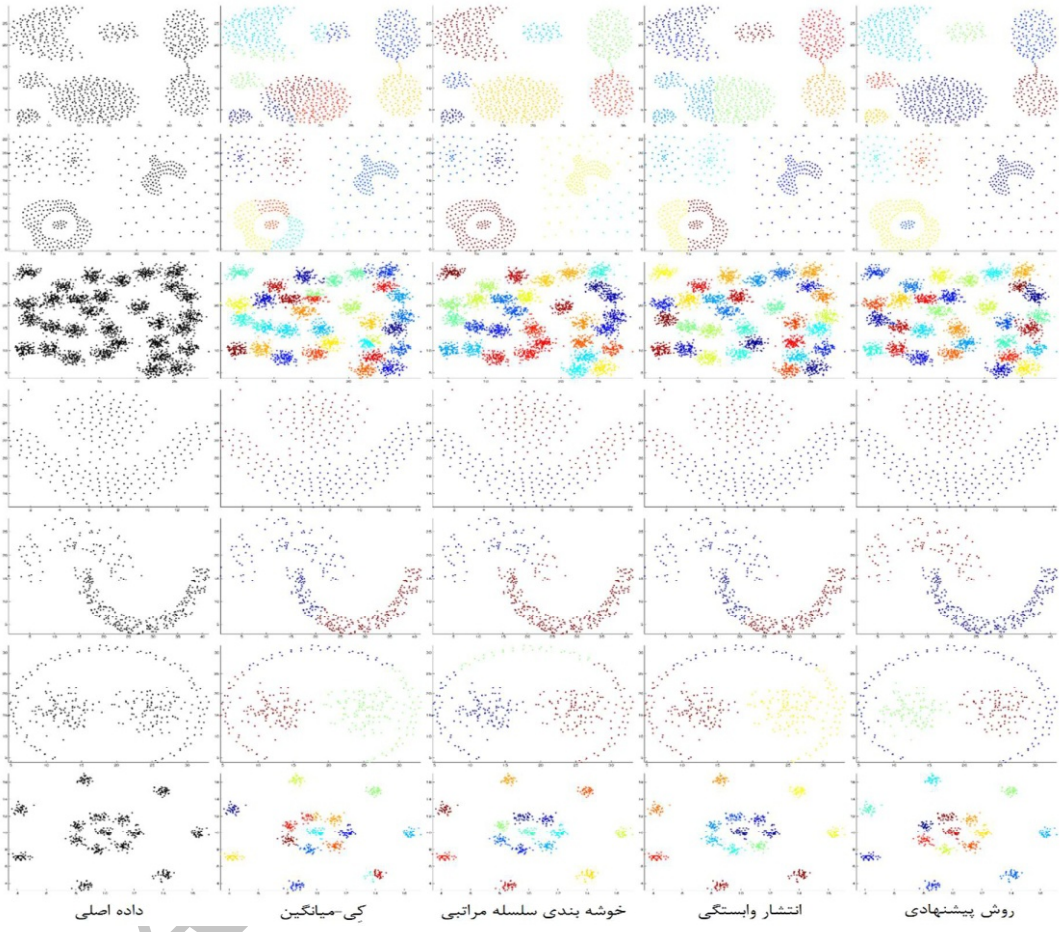
۴- آزمایش‌ها

جهت تأیید مؤثربودن الگوریتم پیشنهادی، آزمایش‌هایی بر روی سه مجموعه داده‌ای متفاوت انجام شده است. نخستین آنها مجموعه‌ای از نقاط داده‌ای دوبعدی ساختگی است که شکل‌های نقطه‌ای متفاوتی تولید می‌کنند. دومین آنها یک مجموعه داده‌ای معروف با نام ORL [19] است که به‌صورت عمومی در دسترس قرار دارد و شامل چهارصد عکس از چهره‌های چهل فرد مختلف است (۱۰ عکس از هر فرد).

سومین و آخرین آزمایش بر روی یک مجموعه داده‌ای عمومی دیگر به نام MPEG7 [12] انجام می‌شود که اغلب برای ارزیابی الگوریتم‌های خوشه‌بندی مورد استفاده قرار می‌گیرد و شامل ۱۴۰۰ شکل دودویی است. در هر آزمایش، دقت روش پیشنهادی با سه روش خوشه‌بندی رایج، یعنی کِی-میانگین، خوشه‌بندی سلسله‌مراتبی میانگین و انتشار وابستگی مقایسه می‌شود. همچنین، ازآن‌جاکه هدف مقایسه کیفیت خوشه‌بندی توسط روش‌های مختلف در شرایط یکسان است، تعداد خوشه‌های موردنیاز در هر آزمایش به‌صورت ازپیش‌تعیین‌شده و برابر با مقدار واقعی آن درنظر گرفته می‌شود.

۱-۴- شبیه‌سازی‌ها

آزمایش نخست بر روی یک مجموعه داده‌ای ساختگی از نقاط دوبعدی است تا توانایی الگوریتم پیشنهادی بر روی ساختارهای متفاوت منیفلد و تعداد مختلف خوشه آزمایش شود. به همین جهت، هفت شکل نقطه‌ای که در شکل (۲) نمایش داده شده‌اند، تولید شد.



(شکل-۲): شبیه سازی ها بر روی هفت شکل متفاوت. هر سطر روش پیشنهادی را با دیگر روش های موجود و پرکاربرد مقایسه می کند. شکل های مختلف شامل تعداد متفاوتی خوشه هستند.

(Figure-2): Simulations on seven different shapes. Every row shows the comparison of our algorithm and other existing popular clustering methods. Different shapes have different number of clusters.

در این آزمایش، به جای استفاده از معیار فاصلهٔ اقلیدسی، از یک معیار فاصله‌مانند استفاده می‌شود که اطلاعات منیفلد را نیز در نظر می‌گیرد [25]. استفاده از این معیار سبب می‌شود تا حدودی ضعف روش‌های خوشه‌بندی که فقط براساس فاصله عمل می‌کنند و درنتیجه نمی‌توانند برای مثال شکل‌های مقعر را به‌درستی خوشه‌بندی کنند، به‌خوبی جبران شود. برای ارزیابی کیفیت نتایج خوشه‌بندی مختلف، از یک شاخص رایج به نام اطلاعات مشترک هنجارسازی‌شده[1] [16] استفاده می‌شود. با داشتن یک مجموعه داده‌ای شامل N عنصر، برچسب‌های خوشه‌بندی به‌دست آمده که با L نمایش داده می‌شود (با تعداد P خوشه) و برچسب‌های خوشه‌بندی حقیقی که با Y نمایش داده می‌شود (با تعداد Q خوشه)، معیار اطلاعات مشترک هنجارسازی‌شده بدین صورت محاسبه می‌شود:

[1] Normalized Mutual Information (NMI)

$$NMI(L,Y) = \frac{I(L,Y)}{\big(H(L)+H(Y)\big)/2} \qquad (۱۰)$$

که در آن I اطلاعات مشترک[1] است و بدین‌صورت محاسبه می‌شود:

$$I(L,Y) = \sum_{p=1}^{P}\sum_{q=1}^{Q}\frac{|L_p \cap Y_q|}{N} \times \log\frac{|L_p \cap Y_q|}{|L_p|/N \times |Y_q|/N} \qquad (۱۱)$$

و H آنتروپی است که برای L و Y به‌صورت زیر محاسبه می‌شوند:

$$H(L) = -\sum_{p=1}^{P}\frac{|L_p|}{N}\log\frac{|L_p|}{N} \qquad (۱۲)$$

$$H(Y) = -\sum_{q=1}^{Q}\frac{|Y_q|}{N}\log\frac{|Y_q|}{N} \qquad (۱۳)$$

بیشترین مقدار ممکن برای NMI، برابر با یک است که بیان‌گر این است که خوشه‌بندی حاصل‌شده به‌طوردقیق معادل خوشه‌بندی حقیقی است.

همان‌طورکه در نتایج نشان داده شده در شکل (۲) مشاهده می‌شود الگوریتم خوشه‌بندی پیشنهادی نسبت به دیگر روش‌های موجود، در فراگیری شکل‌های مختلف توزیع داده بهتر عمل می‌کند.

۴-۲- توسعه برای مقیاس بالا

در این بخش مقیاس‌پذیری الگوریتم پیشنهادی به مجموعه‌های عظیم داده‌ای آزمایش می‌شود. دو آزمایش صورت‌گرفته عبارتند از: ۱) تعداد نقاط داده‌ای از یکصد تا یک میلیون و همچنین ابعاد داده از دو تا صد تغییر داده می‌شوند؛ درحالی‌که تعداد خوشه‌ها به‌صورت ثابت بیست در نظر گرفته می‌شود. ۲) تعداد نقاط داده‌ای از یکصد تا یک‌میلیون و همچنین تعداد خوشه‌ها از دو تا صد تغییر داده می‌شوند؛ درحالی‌که ابعاد داده به‌صورت ثابت بیست در نظر گرفته می‌شود. برای هرکدام از این آزمایش‌ها، به‌صورت تصادفی و مستقل ده بار اجرا صورت می‌گیرد و میانگین و انحراف از معیار زمان مورد نیاز برای خوشه‌بندی محاسبه و گزارش می‌شود. نتایج در شکل (۳) نشان داده شده‌اند. می‌توان مشاهده کرد که حتی با یک میلیون داده، الگوریتم پیشنهادی در بازهٔ زمانی بسیار کوتاه خوشه‌های موردنظر را پیدا می‌کند. لازم به ذکر است که برای این آزمایش‌ها، از نرم‌افزار متلب بر روی یک رایانه عادی و یک کد برنامه بهینه‌نشده استفاده شده است. افزایش سرعت بیشتری را با بهتر مهندسی‌کردن پیاده‌سازی الگوریتم می‌توان به‌دست آورد.

۴-۳- خوشه‌بندی چهره

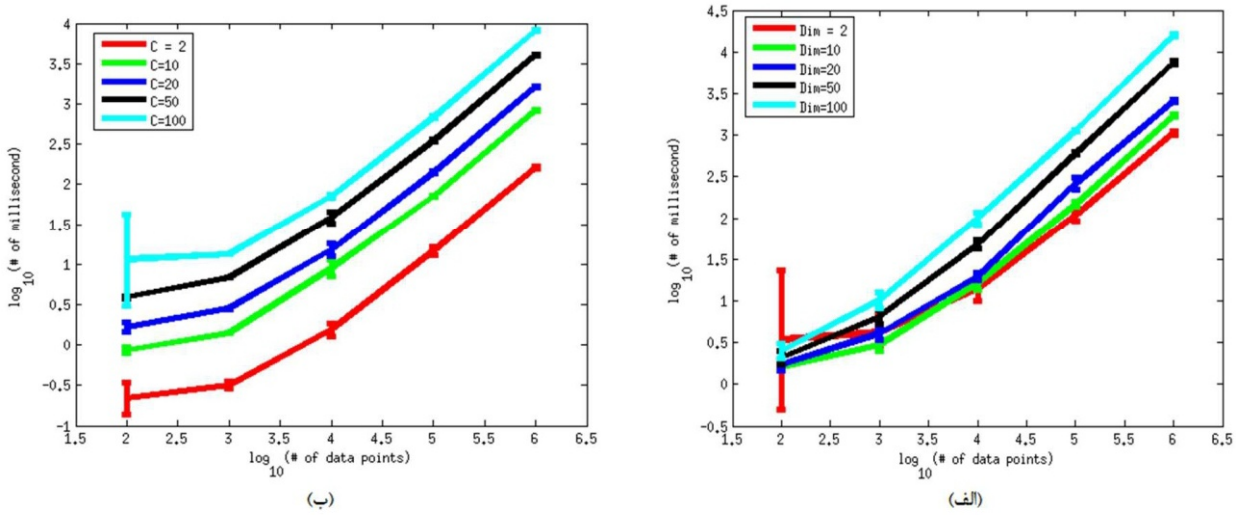
سومین آزمایش بر روی مجموعه داده‌ای ORL انجام می‌گیرد که شامل چهارصد تصویر از چهره چهل فرد است. هدف در اینجا خوشه‌بندی عناصر داده‌ای (که در اینجا همان تصاویر چهره هستند) به چهل خوشه است، به‌طوری‌که هر خوشه تنها شامل عکس‌هایی باشد که متعلق به یک شخص هستند. از مقادیر پیکسل‌ها برای بازنمایی عکس‌ها استفاده می‌شود و پس از اعمال PCA[2] ابعاد بازنمایی تصاویر به یکصد کاهش می‌یابد. سپس از معیار فاصلهٔ اقلیدوسی استفاده می‌شود تا فاصله بین هر دو چهره محاسبه شود و خوشه‌بندی بر روی ماتریس فاصله محاسبه‌شده اجرا می‌شود. برای ارزیابی کارآیی روش پیشنهادی، آن را با روش‌های مطرح و جدیدی همچون IEC[3] [15]، SEC[4] [14]، KCC[5] [26]، USELM[6] [7] و UDELM[7] [17] مقایسه می‌کنیم. از آن جایی که هر کدام از این روش‌ها مراحل پیش‌پردازش و بازنمایی متفاوتی از داده‌ها را استفاده می‌کنند، مقایسه مستقیم اعداد NMI گزارش‌شده امکان‌پذیر نیست. درنتیجه برای مقایسه بهتر، روش کِی-میانگین به‌عنوان روش پایه در نظر گرفته می‌شود و عملکرد هر روش بر اساس میزان ارتقای NMI آن روش نسبت به روش کِی-میانگین سنجیده می‌شود؛ به عبارت دیگر:

$$R_{method} = \frac{NMI_{method} - NMI_{K\text{-}means}}{1 - NMI_{K\text{-}means}} \qquad (۱۴)$$

که در این رابطه R_{metho} نشان‌دهندهٔ عملکرد روش خوشه‌بندی موردنظر در مقایسه با عملکرد روش کِی-میانگین است. همان‌طورکه در جدول (۱) می‌توان مشاهده کرد، روش خوشه‌بندی پیشنهادی نسبت به دیگر روش‌های مطرح خوشه‌بندی، بهتر عمل می‌کند. برای بهتر به‌تصویرکشیدن خوشه‌بندی بر روی این مجموعه داده‌ای، شکل (۴) نتایج خوشه‌بندی را هنگامی که یک مجموعه کوچک از یکصد تصویر چهره متعلق به ده فرد استفاده شده است، نشان می‌دهد.

[1] Mutual Information

[2] Principal Component Analysis
[3] Infinite Ensemble Clustering
[4] Spectral Ensemble Clustering
[5] K-means-based Consensus Clustering
[6] UnSupervised Extreme Learning Machine
[7] Unsupervised Discriminative Extreme Learning Machine



(شکل-۳): هزینه زمانی روش پیشنهادی بر روی یک مجموعه داده‌ای شبیه سازی شده با مقیاس بالا. (الف) تعداد نقاط داده‌ای از ۱۰۰ تا 10^6 و ابعاد داده از ۲ تا ۱۰۰ تغییر می کند در حالی که تعداد خوشه ها به طور ثابت برابر با ۲۰ هستند؛ (ب) تعداد نقاط داده‌ای از ۱۰۰ تا 10^6 و تعداد خوشه‌ها از ۲ تا ۱۰۰ تغییر می کند در حالی که ابعاد به طور ثابت برابر با ۲۰ هستند.

(Figure-3): Time cost of the proposed method on a large scale simulation dataset. (A) We range the number of data points from 100 to 10^6 (1 million) and also range the dimension of data from 2 to 100, while fixing the number of clusters to be 20; (B) We range the number of data points from 100 to 10^6 (1 million) and also range the number of clusters from 2 to 100, while fixing the number of dimensions to be 20.

(جدول-۱): نتیجه اعمال روش پیشنهادی بر روی مجموعه داده‌ای ORL، در مقایسه با دیگر روش‌های خوشه‌بندی.

(Table-1): The result of our proposed clustering method on the ORL dataset, in comparison to other methods

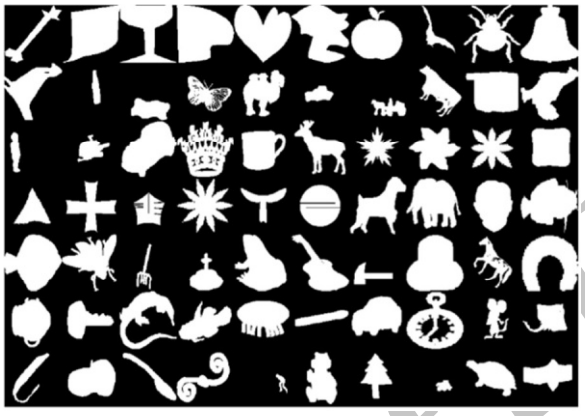
روش پیشنهادی	UDELM	USELM	KCC	SEC	IEC	روش
56.48	29.97	15.18	4.93	8.81	16.98	R_{method} (%)

(شکل-۴): نتایج بر روی مجموعه داده‌ای ORL. (الف) یک زیر مجموعه تصادفی از مجموعه داده‌ای ORL را نشان می دهد و (ب) نشان‌دهنده نتایج خوشه بندی بر روی داده های قسمت (الف) است.

(Figure-4): Results on ORL dataset. (A) is a random subset of the ORL dataset, and (B) is the corresponding clustering results using our algorithm.

۴-۴- خوشه‌بندی شکل

درنهایت چهارمین آزمایش بر روی مجموعه داده‌ای MPEG7 انجام می‌گیرد که شامل ۱۴۰۰ تصویر دودویی از هفتاد طبقه است که برای هر طبقه بیست شکل مختلف وجود دارد (به شکل (۵) برای دیدن نمونه‌ها توجه کنید). هدف در اینجا خوشه‌بندی عناصر داده‌ای (که در اینجا شکل‌های دودویی هستند) به هفتاد خوشه است، به‌طوری‌که هر خوشه تنها شامل شکل‌هایی باشد که به یک طبقه تعلق دارند. در این آزمایش از دو الگوریتم مختلف تطبیق شکل به نام‌های بافت شکل[1] یا SC [3] و فاصله داخلی[2] یا IDSC [13]، جهت تولید ماتریس فاصله، استفاده می‌شود. مشابه با آزمایش نخست، اطلاعات منیفلد نیز در محاسبه ماتریس فاصله منظور شده است. جدول (۲) خلاصه‌ای از نتایج به‌دست‌آمده در این آزمایش را گزارش می‌دهد. همان‌طورکه می‌توان دید، روش پیشنهادی بر روی این مجموعه داده‌ای نیز از سایر الگوریتم‌های خوشه‌بندی موجود بهتر عمل می‌کند.



(شکل-۵): شکل‌های نمونه از مجموعه داده‌ای MPEG7.

(Figure-5): Sample shapes from the MPEG-7 dataset.

(جدول-۲): نتیجه اعمال روش پیشنهادی بر روی مجموعه داده‌ای MPEG7، در مقایسه با دیگر روش‌های خوشه‌بندی. اعداد هر ستون بیانگر میزان کیفیت خوشه بندی بر حسب NMI می‌باشند.

(Table-2): The result of our proposed clustering method on the MPEG7 dataset, in comparison to other methods. The numbers represent the quality of clustering based on NMI.

معیار تخمین فاصله	خوشه بندی سلسله مراتبی	انتشار وابستگی	روش پیشنهادی
SC	0.8616	0.9073	**0.9606**
IDSC	0.8327	0.8974	**0.9493**
SC+IDSC	0.8792	0.9491	**0.9806**

۵- نتیجه‌گیری

یک روش خوشه‌بندی جدید ارائه شد که به‌صورت تکراری عناصر کلیدی شاخص را از مجموعه داده‌ها انتخاب می‌کند و خوشه‌های مربوطه را تشکیل می‌دهد. ایدهٔ اصلی روش پیشنهادی بسیار ساده است: فاصلهٔ بین مراکز خوشه‌های مختلف باید زیاد باشد؛ درحالی‌که فواصل درون یک خوشه باید کم باشند. این ایده به‌خوبی در انتخاب عناصر کلیدی که به‌صورت تکراری و با توجه به بیشینه فاصلهٔ بین هر دو داده انجام می‌گیرد، نمایان شده است. چهار ویژگی شاخص، روش پیشنهادی را از دیگر روش‌های موجود متمایز می‌سازد: ۱) این روش فرض می‌کند که تنها نزدیک ترین همسایه‌ها، که با توجه به معیار فاصله تعریف می‌شوند، معتبر هستند. این فرض در مقایسه با دیگر روش‌ها، که به‌طورمعمول فرض می‌کنند معیارهای کلی فاصله به‌خوبی تمامی فضای داده‌ای را می‌توانند توضیح دهند، فرضی به‌نسبه معتدل است. ۲) روش پیشنهادی از یک روش انتشار گراف[3] بهره می‌گیرد تا معیارهای فاصله را بهبود دهد که باعث می‌شود کیفیت

[1] Shape Contexts (SC)

[2] Inner Distance (IDSC)

[3] Graph diffusion method

نزدیک‌ترین همسایه‌ها به طور مضاعف تضمین شود. ۳) روش پیشنهادی نه‌تنها خوشه‌ها را معین می‌کند، بلکه داده‌های شاخص قابل اعتماد نیز تولید می‌کند. این موضوع در بسیاری از کاربردهای دنیای واقعی مانند طبقه‌بندی سلول‌ها در زیست‌شناسی سلولی حائز اهمیت است. ۴) این روش به‌سادگی قابل پیاده‌سازی و مقیاس‌پذیر به داده‌های با مقیاس بالا است.

هر چند که در توضیح، روش ساده به نظر می رسد، الگوریتم پیشنهادی نسبت به الگوریتم‌های برتر و رایج خوشه‌بندی موجود، چه از نظر دقت و چه از نظر سرعت، بهتر عمل می‌کند. علاوه‌براین، الگوریتم پیشنهادی توسعه داده شد تا برای خوشه‌بندی مقیاس بالا مناسب باشد. نتایج تجربی بر روی چندین مجموعه داده‌ای بیان‌گر برتری روش پیشنهادی است.

۶- مراجع 6-References

[۱] چاقری آرش، فیضی درخشی محمدرضا. *خوشه‌بندی خودکار داده‌ها با بهره‌گیری از الگوریتم رقابت استعماری بهبودیافته*. پردازش علائم و داده‌ها; ۱۴ (۲) :۱۶۹-۱۵۹; ۱۳۹۶

[1] Chaghari A, Feizi-Derakhshi M. Automatic Clustering Using Improved Imperialist Competitive Algorithm. JSDP; 14 (2) :159-169, 2017.

[2] Bahmani, B., Moseley, B., Vattani, A., Kumar, R., & Vassilvitskii, S. *Scalable k-means++*. Proceedings of the VLDB Endowment, 5(7), 622-633, 2012.

[3] Belongie, S., Malik, J., & Puzicha, J. *Shape matching and object recognition using shape contexts*. IEEE transactions on pattern analysis and machine intelligence, 24(4), 509-522, 2002.

[4] Comaniciu, D., & Meer, P. *Mean shift: A robust approach toward feature space analysis.* IEEE Transactions on pattern analysis and machine intelligence, 24(5), 603-619, 2002.

[5] El-Naqa, I., Yang, Y., Galatsanos, N. P., Nishikawa, R. M., & Wernick, M. N. *A similarity learning approach to content-based image retrieval: application to digital mammography.* IEEE transactions on medical imaging, 23(10), 1233-1244, 2004.

[6] Frey, B. J., & Dueck, D. *Clustering by passing messages between data points*. Science, 315(5814), 972-976, 2007.

[7] Huang, G., Song, S., Gupta, J. N., & Wu, C. *Semi-supervised and unsupervised extreme learning machines*. IEEE Transactions on Cybernetics, 44(12), 2405-2417, 2014.

[8] Jain, A. K., Murty, M. N., & Flynn, P. J. *Data clustering: a review*. ACM computing surveys (CSUR), 31(3), 264-323, 1999.

[9] Jain, A. K. *Data clustering: 50 years beyond K-means.* Pattern recognition letters, 31(8), 651-666, 2010.

[10] Johnson, S. C. *Hierarchical clustering schemes*. Psychometrika, 32(3), 241-254, 1967.

[11] Koyuturk, M., Grama, A., & Ramakrishnan, N. *Compression, clustering, and pattern discovery in very high-dimensional discrete-attribute data sets*. IEEE Transactions on Knowledge and Data Engineering, 17(4), 447-461, 2005.

[12] Latecki, L. J., Lakamper, R., & Eckhardt, T. *Shape descriptors for non-rigid shapes with a single closed contour*. In Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on (Vol. 1, pp. 424-429). IEEE, 2000.

[13] Ling, H., & Jacobs, D. W. *Shape classification using the inner-distance.* IEEE transactions on pattern analysis and machine intelligence, 29(2), 2007.

[14] Liu, H., Liu, T., Wu, J., Tao, D., & Fu, Y. *Spectral ensemble clustering*. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 715-724). ACM, 2015, August.

[15] Liu, H., Shao, M., Li, S., & Fu, Y. *Infinite ensemble for image clustering.* In Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, August.

[16] Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens, P. *Multimodality image registration by maximization of mutual information.* IEEE transactions on medical imaging, 16(2), 187-198, 1997.

[17] Peng, Y., Zheng, W. L., & Lu, B. L. *An unsupervised discriminative extreme learning machine and its applications to data clustering*. Neurocomputing, 174, 250-264, 2016.

[18] Reynolds, D. *Gaussian mixture models*. Encyclopedia of biometrics, 827-832, 2015.

[19] Samaria, F. S., & Harter, A. C. *Parameterisation of a stochastic model for human face identification*. In Applications of Computer Vision, 1994., Proceedings of the Second IEEE Workshop on (pp. 138-142). IEEE, 1994, December.

[20] Sculley, D. *Web-scale k-means clustering.* In Proceedings of the 19th international conference on World Wide Web (pp. 1177-1178). ACM, 2010, April.

[21] Soheily-Khah, S., Douzal-Chouakria, A., & Gaussier, E. *Generalized k-means-based clustering for temporal data under weighted and kernel time warp*. Pattern Recognition Letters, 75, 63-69, 2016.

[22] Steinbach, M., Karypis, G., & Kumar, V. *A comparison of document clustering techniques.* In KDD workshop on text mining (Vol. 400, No. 1, pp. 525-526), 2000, August.

[23] Tuma, M. N., Scholz, S. W., & Decker, R. *THE APPLICATION OF CLUSTER ANALYSIS IN MARKETING RESEARCH: A LITERATURE ANALYSIS.* B> Quest, 2009.

[24] Von Luxburg, U. *A tutorial on spectral clustering.* Statistics and computing, 17(4), 395-416, 2007.

[25] Wang, B., Mezlini, A. M., Demir, F., Fiume, M., Tu, Z., Brudno, M., & Goldenberg, A. *Similarity network fusion for aggregating data types on a genomic scale.* Nature methods, 11(3), 333-337, [26] Wu, J., Liu, H., Xiong, H., Cao, J., & Chen, J. *K-means-based consensus clustering: A unified view.* IEEE Transactions on Knowledge and Data Engineering, 27(1), 155-169, 2015.

[27] Zhang, Z., Pati, D., & Srivastava, A. *Bayesian clustering of shapes of curves.* Journal of Statistical Planning and Inference, 166, 171-186, 2015.

احسان فضل ارثی دکترای حرفه‌ای خود را از دانشگاه یورک کانادا در سال ۱۳۹۱ دریافت کرد. وی پس از گذراندان دوره پسا دکترا در دانشگاه اتاوا، در بخش صنعت کشور کانادا به‌عنوان مدیر تحقیق و توسعه یک شرکت سهامی عام مشغول به‌کار شد. ایشان در حال حاضر عضو هیأت علمی در دانشگاه فردوسی مشهد بوده و در زمینه‌های بینایی کامپیوتر و یادگیری ماشین پژوهش‌های خود را پیش می‌برد.

نشانی رایانامه ایشان عبارت است از:

fazlersi@um.ac.ir

مسعود کاظمی نوقابی دانشجوی مقطع کارشناسی رشتهٔ مهندسی کامپیوتر گرایش نرم افزار است. وی نزدیک به دو سال است که در آزمایشگاه پژوهشی بینایی کامپیوتر و رباتیک فعالیت دارد. زمینه‌های فعالیت وی یادگیری ماشین و بینایی کامپیوتر است.

نشانی رایانامه ایشان عبارت است از:

masoud.kazemi@mail.um.ac.ir