Study and Comparison of Partitioning Clustering Algorithms

بررسی و مقایسه الگوریتم‌های خوشه‌بندی تفکیکی

Faezeh Hosseininezhad, Afshin Salajegheh

Abstract — *Clustering is one of the main operations in data mining and its aim is to group similar objects in clusters. This technique seeks to discover structure of dataset by considering similarities or differences between data. Clustering algorithms can be divided into several categories including partitioning clustering algorithms, hierarchical algorithms and density based algorithms. In this paper we investigate some partitioning algorithms and consider them in term of some important parameters and finally compare them*[1].

Keywords — delay/fault-tolerant mobile sensor network, delivery delay, delivery probability, DFT-MSN, erasure coding, pervasive information gathering, queuing theory, replication, transmission overhead.

۱. چکیده

خوشه‌بندی یکی از اعمال اصلی در داده‌کاوی است و به معنای گروه‌بندی نمونه‌های مشابه در خوشه‌ها می‌باشد. این تکنیک به دانبال کشف ساختار در داده‌ها از طریق بررسی شباهت‌ها یا تفاوت‌های میان آنهاست. الگوریتم‌های خوشه‌بندی را می‌توان به چند دسته کلی تقسیم کرد که از جمله می‌توان الگوریتم‌های خوشه‌بندی تفکیکی، الگوریتم‌های سلسله‌مراتبی و الگوریتم‌های مبتنی بر چگالی را نام برد. در این مقاله به بررسی چند نمونه از الگوریتم‌های نوع تفکیکی پرداخته و سپس سعی می‌کنیم آنها را از لحاظ چند پارامتر مهم بررسی و با یکدیگر مقایسه کنیم.

کلمات کلیدی

الگوریتم های خوشه‌بندی تفکیکی، تابع هدف، یادگیری بدون نظارت

۲. مقدمه

امروزه با توجه به در اختیار داشتن حجم بالای اطلاعات در زمینه‌های مختلف و این حقیقت که ترکیبی از داده‌های مفید و غیرمفید در اختیار افراد قرار می‌گیرد، الزام استفاده از روش‌های خاص جهت استخراج اطلاعات مفید از داخل حجم انبوهی از داده‌ها به‌خوبی احساس می‌گردد. یکی از تکنیک‌هایی که طی سالیان اخیر جهت انجام این امر مطرح گردیده است تکنیک داده‌کاوی است. مدل‌ها و استراتژی‌های مختلف و متنوعی از فرآیند داده‌کاوی وجود دارد که از پرکاربردترین آنها می‌توان دسته‌بندی، خوشه‌بندی، رگرسیون و استخراج قوانین انجمنی را نام برد. به‌طور کلی الگوریتم‌های خوشه‌بندی را می‌توان به چند دسته از جمله روش‌های تفکیکی، روش‌های سلسله‌مراتبی و روش‌های مبتنی بر چگالی تقسیم کرد. نوع دیگر تقسیم‌بندی، الگوریتم‌ها را به دو دسته قطعی یا سخت و غیرقطعی یا فازی تقسیم می‌کند. در این مقاله ابتدا به توضیح مفاهیم اولیه تکنیک خوشه‌بندی پرداخته، سپس به بررسی چند نمونه از الگوریتم‌های تفکیکی و مقایسه آنها می‌پردازیم.

۳. خوشه‌بندی

خوشه، مجموعه‌ای از اشیا یا داده‌هاست که به یکدیگر شبیه‌اند و با اشیاء موجود در خوشه دیگر متفاوتند. خوشه‌بندی اساساً به معنای گروه-بندی نمونه‌های مشابه می‌باشد. یک خوشه‌بندی خوب خوشه‌هایی نتیجه می‌دهد که نمونه‌های داخل یک خوشه به هم نزدیک یا مشابه باشند و در عین حال شباهت بین خوشه‌ها حتی‌الامکان کم باشد. درواقع خوشه‌بندی یک کلاس‌بندی بدون نظارت است که درآن کلاس‌های از پیش تعریف شده‌ای وجود ندارد. به‌عبارتی خوشه‌بندی نوعی یادگیری بدون نظارت است. به همین دلیل، این تکنیک را می‌توان شکلی از یادگیری بوسیله داده‌ها یا مشاهدات(و نه یادگیری به‌وسیله مثال‌ها) دانست. اگرچه دسته-

[1] F. Hosseininezhad is a MSc student in Software Engineering, Islamic Azad University, Tehran Southern Branch, Tehran, Iran (email: st_f_hosseininezhad@azad.ac.ir)

A. Salajegheh is Assistant Professor of Department of Computer Engineering, Islamic Azad University, Tehran Southern Branch, Tehran, Iran (email: a_salajegheh@azad.ac.ir)

بندی یک ابزار مفید برای متمایز کردن گروه‌های داده است اما زمانی که برای مجموعه حجیمی از رکوردهای آموزشی به کار گرفته می‌شود، فرآیندی زمان‌بر و پرهزینه خواهد بود[۱]. باید توجه داشت که برخلاف دسته‌بندی، در خوشه‌بندی هیچ دسته‌ای از قبل وجود ندارد. در هر فرآیند خوشه‌بندی مراحلی طی می‌شود که عبارتند از:

۱. تهیه و ارائه ماتریس داده‌ها
۲. استاندارد کردن ماتریس داده‌ها
۳. محاسبه ماتریس مجاورت (فاصله یا مشابهت)
۴. اجرای روش خوشه‌بندی
۵. محاسبه معیار(های) اعتبار

یکی از مسائل مهم در فرآیند خوشه‌بندی تعیین معیاری برای محاسبه فاصله میان داده‌هاست. معیارهای مختلفی برای اندازه‌گیری فاصله بین اشیا وجود دارد که از معمول‌ترین و پرکاربردترین آنها می‌توان فاصله اقلیدسی را نام برد. فاصله اقلیدسی برای دو نقطه X ,Y در فضای n بعدی از رابطه (۱) بدست می‌آید.

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)} \qquad (۱)$$

۴. الگوریتم‌های خوشه‌بندی تفکیکی

این نوع خوشه‌بندی که خوشه‌بندی مبتنی بر تابع هدف نیز نامیده می‌شود n رکورد داده را به k دسته تقسیم می‌کند به‌طوری که هر دسته بیانگر یک خوشه است و k<=n. با توجه مقدار k(تعداد خوشه‌ها) یک الگوریتم خوشه‌بندی تفکیکی یک دسته‌بندی آغازی را کرده و با استفاده از تکنیک جابجایی تکراری تلاش می‌کند که با جابجایی اشیا از یک خوشه به خوشه دیگر، دسته‌بندی را بهبود دهد. این بهبود معمولاً با کمینه‌سازی یک تابع هدف(تابع هزینه، تابع خطا) تعریف شده میسر می‌شود. تاکنون الگوریتم‌های مختلف و متعددی از این نوع خوشه‌بندی توسعه داده شده‌اند که در ادامه به توضیح برخی از مهم‌ترین و پرکاربردترین آنها می‌پردازیم.

أ. الگوریتم k-means

این الگوریتم یکی از معروف‌ترین و ساده‌ترین الگوریتم‌ها است و علی‌رغم اینکه سال‌ها از ابداع آن می‌گذرد و پس از آن تعداد زیادی الگوریتم خوشه‌بندی توسعه داده‌شده‌اند، اما به‌دلیل مزایایی مثل سهولت پیاده‌سازی، سادگی و کارایی بالا هنوز هم به‌طور وسیعی مورد استفاده قرار می‌گیرد [۲]. الگوریتم با انتخاب k نقطه تصادفی به‌عنوان مراکز خوشه‌ها آغاز می‌شود سپس هر داده به خوشه متناظرش تخصیص داده می‌شود. در این الگوریتم میانگین داده‌های یک خوشه نماینده یک خوشه است. مراحل الگوریتم به‌صورت زیر است:

انتخاب تصادفی k نمونه برای مقدار دهی اولیه مراکز خوشه‌ها

برای هر نمونه نزدیکترین مرکز خوشه را با استفاده از فاصله اقلیدسی پیدا می کنیم .در پایان این مرحله تمام نمونه‌ها درk خوشه قرار دارند.

برای هر کدام از خوشه‌ها، مرکز ثقل جدید را محاسبه و مقدار آن‌را به‌روزرسانی می‌کنیم،

تکرار مراحل ۲و ۳ تا زمانی‌که الگوریتم خاتمه یابد.

الگوریتم تا زمانی ادامه می‌یابد که معیار مربع خطای تعریف شده به‌صورت رابطه (۲) حداقل شود. حاصل این عبارت مجموع فاصله اشیا از خوشه خودشان است.

$$E=\sum_{i=1}^{k}\sum_{p\in c_i}(p-m_i)^2 \qquad (۲)$$

که در آن m_i مرکز خوشه iام و k تعداد خوشه‌هاست.

همانطور که مشاهده شد، هدف الگوریتم k-means حداقل کردن فاصله بین اجزای یک خوشه و حداکثر کردن فاصله بین اجزای خوشه‌های مجزاست. اما این الگوریتم معایبی دارد که از جمله آن می‌توان موارد زیر را برشمرد:

حساس بودن نسبت به داده‌های دور از مرکز

وابسته بودن نتایج به انتخاب مراکز اولیه و انجام یک جستجوی محلی

ناتوانی در شناسایی داده نویز

در الگوریتم k-medoids که در ادامه توضیح داده می‌شود تلاش می‌شود که یکی از معایب الگوریتم k-means یعنی حساسیت به داده‌های دورافتاده برطرف شود.

ب. الگوریتم k-medoids

الگوریتم k-medoids الگوریتمی مبتنی بر شی می‌باشد و نماینده خوشه‌ها را از میان خود داده‌ها(و نه میانگین‌گیری از آنها) انتخاب می‌کند. در واقع medoid یک خوشه، مرکزی ترین عنصر یک خوشه است. هدف این متد کم کردن حساسیت نسبت به مقادیر بزرگ در مجموعه داده‌هاست[۱]. در این الگوریتم هر خوشه با یکی از داده‌های نزدیک به مرکز معرفی می‌شود. مراحل الگوریتم به‌صورت زیر است:

انتخاب تصادفی k شی به‌عنوان نماینده خوشه‌ها

برای هر نمونه نزدیکترین نماینده خوشه را پیدا می کنیم، در پایان این مرحله تمام نمونه ها درk خوشه قرار دارند

به‌طور تصادفی یک شی غیرمدوید را با یک مدوید جایگزین می‌کنیم

هزینه حاصل از تعویض شی غیرمدوید و شی مدوید را محاسبه می‌کنیم و در صورت منفی بودن هزینه، جابه‌جایی را انجام می‌دهیم.

(فاصله از مدوید قبلی- فاصله شی از مدوید جدید = هزینه)

مراحل ۲ تا ۴ را تا زمانی که تغییری رخ ندهد ادامه می‌دهیم

الگوریتم شرح داده شده PAM[۲] نامیده می‌شود که یکی از اولین الگوریتم‌های K-medoid است. در این روش بعد از انتخاب تصادفی k نماینده، الگوریتم در مراحل متوالی به دنبال یافتن انتخاب بهتر برای نماینده خوشه‌هاست. به این منظور همه جفت داده‌های ممکن که یکی از آنها به عنوان نماینده مطرح می‌شود، تحلیل می‌شوند. یکی از معایب این الگوریتم این است که به تعداد تکرار زیادی نیاز دارد و به همین دلیل برای خوشه‌بندی مجموعه داده‌های حجیم

[2] Partitioning Around Medoids

مناسب نیست. برای مجموعه داده‌های بزرگ از متد مبتنی بر نمونه-گیری CLARA[3] می‌توان استفاده کرد. این الگوریتم چندین نمونه تصادفی از پایگاه داده برمی‌دارد و الگوریتم PAM را روی هر نمونه اجرا کرده و آن نمونه را خوشه‌بندی می‌کند. بهترین خوشه‌بندی به‌عنوان خروجی ارائه می‌شود. سپس عناصر باقیمانده پایگاه داده را بر مبنای این خروجی به نزدیکترین خوشه اختصاص می‌دهد. درواقع در این الگوریتم یک قسمت کوچکی از داده‌های واقعی به عنوان نماینده داده‌ها انتخاب می‌شوند.

ج.الگوریتم k-modes

الگوریتم k-modes تعمیم داده شده الگوریتم k-means برای داده‌های اسمی است. ایده اصلی این روش این است که داده‌های اسمی را بر مبنای مد آن-ها به چند خوشه از پیش تعیین شده اختصاص دهد[۳]. اگر X و Y دو داده اسمی باشند که با m خصیصه تعریف شده‌اند، فاصله (عدم شباهت) این دو متغیر بر مبنای عدم تطابق خصیصه‌های متناظر از روابط (۳) و(۴) حاصل می‌شود.

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j) \qquad (۳)$$

$$\delta(x_j,y_j)=\begin{cases}0 & (x_j=y_j)\\ 1 & (x_j\neq y_j)\end{cases} \qquad (۴)$$

روند این الگوریتم مانند k-means است.

تابع هزینه برای این الگوریتم به‌صورت رابطه(۵) است:

$$E=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}d(x_{ij},q_{lj}) \qquad (۵)$$

n تعداد داده‌ها، m ابعاد داده، k تعداد خوشه‌ها و q مد یک خوشه را نشان می‌دهد. مانند الگوریتم k-means الگوریتم k-modes هم مستعد در ایجاد کمینه محلی است که بستگی به انتخاب مدهای اولیه دارد.

د. الگوریتم k-prototypes

این الگوریتم که الگوریتم k پیش‌الگو نیز نامیده می‌شود، بر پایه دو الگوریتم k-means و k-modes بنا نهاده شده و برای داده‌های ترکیبی استفاده می-شود[۳]. متدهایی مانند k-means به دلیل آنکه تابع هزینه‌ای مبتنی بر فاصله اقلیدسی را بهینه می‌کنند، محدود به داده‌های عددی هستند و این درحالی است که داده‌ها در دنیای واقعی ترکیبی از داده‌های اسمی و عددی هستند. در این الگوریتم یک پیش الگو مرکز یک خوشه است. اگر داده‌های X,Y با مشخصه‌های $A_1,A_2,\ldots,A_p,A_{p+1},\ldots,A_m$ تعریف شده‌باشند. فاصله بین آنها با استفاده از رابطه (۶) اندازه‌گیری می‌شود.

$$d(x,y)=\sum_{j=1}^{p}(x_j-y_j)^2+\gamma\sum_{j=p+1}^{m}\delta(x_j,y_j) \qquad (۶)$$

عبارت اول مربع فاصله اقلیدسی است که برای خصیصه‌های عددی به‌کار می-رود و دومی هم معیار فاصله برای خصیصه‌های اسمی است. ضریب γ به‌منظور حفظ تعادل تأثیر گذاری خصیصه‌ها استفاده می‌شود.

ه. الگوریتم FCM

الگوریتم‌های خوشه‌بندی قطعی، داده‌ها را به‌گونه‌ای افراز می‌کنند که هر داده دقیقاً به یک خوشه تخصیص داده شود. دربرخی موارد نمی‌توان هر داده را دقیقاً به یک خوشه تخصیص داد چراکه برخی داده‌ها بین خوشه‌ها قرار می‌گیرند. در این موارد روش‌های خوشه‌بندی فازی ابزارهایی مناسب‌تر برای نمایش ساختار واقعی این نوع داده‌ها هستند. در این تکنیک هر شی با درجه عضویتی به خوشه-ها تعلق می‌گیرد. بنابراین یک شی می‌تواند همزمان عضو دو یا چند خوشه باشد.

این الگوریتم که شکل توسعه یافته روش Kمیانگین است درصدد کمینه کردن تابع هدف رابطه (۷) است:

$$F_{FCM}=\sum_{k=1}^{k}\sum_{j=1}^{n}\mu_{j,k}^{\rho}.\left\|y_j-m^{(k)}\right\|^2 \qquad (۷)$$

که در آن $\mu_{j,k}$ درجه عضویت شی j در خوشه kو m مرکز خوشه k است.درجه عضویت، عددی بین ۰و۱ و مجموع درجه عضویت‌های یک شی در خوشه‌های مختلف برابر ۱است. پارامتر ρ میزان فازی بودن را در خوشه‌بندی کنترل می‌کند. هرچقدر این پارامتر کوچکتر باشد فازی بودن کمتر و هرچه به بی‌نهایت نزدیکتر باشد میزان فازی بودن بیشتر می‌شود. درجه عضویت و مرکز خوشه به‌ترتیب از روابط (۸)و (۹) به‌دست می‌آیند.

$$\mu_{j,k}=\frac{1}{\sum_{k'=1}^{k}\left(\frac{\left\|y_j-m^{(k)}\right\|}{\left\|y_j-m^{(k')}\right\|}\right)^{\frac{2}{\rho-1}}} \qquad (۸)$$

$$m^{(k)}=\frac{\sum_{j=1}^{N}\mu_{j,k}^{\rho}.y_j}{\sum_{j=1}^{k}\mu_{j,k}^{\rho}.},\quad k=1,\ldots,K. \qquad (۹)$$

۵.الگوریتم PSO

الگوریتم PSO[4] الهام گرفته از رفتار اجتماعی پرندگان است که توسط Kennedy و Eberhart در سال ۱۹۹۵ ارائه شد. این روش از الگوریتم‌های مبتنی بر هوش گروهی محسوب می‌شود. سیستم‌های مبتنی بر هوش گروهی معمولاً شامل جمعیتی از عامل‌ها هستند که به صورت محلی با یکدیگر و محیط اطرافشان در تعامل می‌باشند. نمونه‌های طبیعی از هوش گروهی شامل دسته‌های پرندگان، کلونی مورچگان، کلونی زنبورهای عسل و ... است. الگوریتمPSO به-دنبال کشف الگوهایی است که این قابلیت را به پرندگان می دهد تا به طور همزمان با هم پرواز نمایند و ناگهان تغییر جهت دهند و در گروه‌بندی جدیدی با شکل بهینه ترتیب یابند. تغییر موقعیت هر پرنده در فضای جستجو بر اساس

[3] Clustering LARge application

[4] Particle swarm optimization
[12] Fitness function

رفتار اجتماعی پرندگان و تمایل آنها به موفقیت در رقابت با دیگر پرندگان شکل می‌گیرد. به عبارت دیگر تغییرات در موقعیت پرندگان یا ذرات بر اساس تأثیر آزمایش خود یا دانش همسایگان آنها می‌باشد. این اثر در واقع نوعی از همزیستی اجتماعی محسوب می‌شود. معادلات سرعت و مکان هر ذره به صورت روابط (۱۰) و (۱۱) است:

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (۱۰)$$

$$v_i(t+1) = wv_i(t) + c_1r_1(x_i^{pb}(t) - x_i(t)) + c_2r_2(x^*(t) - x_i(t)) \quad (۱۱)$$

هر ذره دارای حافظه است و بهترین موقعیت خودش که تا کنون بدست آورده (xipd) و بهترین موقعیت همسایگانش (x*) را در هر تکرار در حافظه اش نگه می‌دارد و بر این اساس بردار سرعتش را تنظیم می کند.

تکنیک PSO عمدتاً برای حل مسائل بهینه‌سازی به‌کار می‌رود. در واقع مکان یک پرنده یک جواب برای مسئله محسوب می‌شود. زمانی که یک ذره به یک مکان جدید حرکت می‌کند، یک جواب متفاوت برای مسئله تولید می‌شود. این جواب با استفاده از یک تابع شایستگی[۵] ارزیابی می‌شود. برای به‌کارگیری این تکنیک در خوشه‌بندی، ابتدا باید فرآیند خوشه‌بندی را به‌عنوان یک مسئله بهینه-سازی مدل کنیم که هدف از اینچنین مدلی به‌دست آوردن مراکز خوشه‌هاست به‌طوری که یک تابع هدف بهینه شود.

همانطور که گفته شد، هدف الگوریتمی مانند k-means حداقل کردن فاصله بین اجزای یک خوشه و حداکثر کردن فاصله بین اجزای خوشه‌های مجزاست. با وجود سادگی، این الگوریتم یک مشکل دارد: نتایج به‌شدت وابسته به انتخاب مراکز اولیه است. و در واقع این الگوریتم یک جستجوی محلی را انجام می‌دهد. در جستجوی محلی جواب به‌دست آمده در یک مرحله معمولاً در مجاورت حل به‌دست آمده در مرحله قبل است. به‌علاوه چون k-means انتخاب تصادفی دارد روی یک مجموعه داده چند نتیجه می‌دهد. برای تعیین مراکز اولیه و تعیین تعداد خوشه‌ها چند الگوریتم توسعه داده شده‌اند که در ادامه به توضیح برخی از آنها می‌پردازیم.

أ. الگوریتم خوشه‌بندی مبتنی بر PSO

در این الگوریتم موقعیت هر ذره به‌صورت $x_i = (m^{(1)},\ldots,m^{(k)})$ نشان داده می‌شود. به‌طوری که $m^{(k)}$ مرکز خوشه k ام را نشان می‌دهد[۵، ۶]. شکل (۱) بیانگر این مطلب است. مراحل این الگوریتم به‌صورت زیر است:

گام اول: هر ذره به طور تصادفی k مرکز خوشه را از میان داده‌ها انتخاب می-کند.

گام دوم: به تعداد تکرارهای مناسب

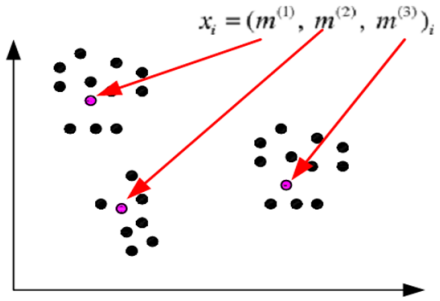
برای هر ذره عملیات زیر انجام می‌گیرد.

برای هر بردار داده فاصله اقلیدسی با تمام مراکز محاسبه می‌شود و داده به نزدیکترین مرکز خوشه منتسب می‌شود.

تابع شایستگی محاسبه می‌شود.

بهترین مکان کلی و بهترین مکان محلی برای هر ذره محاسبه می‌شود

مراکز خوشه‌ها با توجه به فرمول‌های PSO به روز می‌شوند.



شکل (۱) موقعیت هر ذره در خوشه‌بندی بر مبنای PSO

تابع شایستگی که در این‌گونه الگوریتم‌ها استفاده می‌شود یکی از شاخص‌های اعتبار خوشه‌بندی است مانند معیار تراکم(compactness measure) یا معیار تمایز(separation measure) و یا شاخص‌های ارزیابی دیگری که کیفیت خوشه‌های حاصل شده را می‌سنجند.

ب. الگوریتم خوشه‌بندی مرکب PSO

الگوریتم خوشه‌بندی بر مبنای PSO برای مجموعه داده‌های با ابعاد کم نسبت به الگوریتم k-means بهتر عمل می‌کند اما برای داده‌ها با ابعاد زیاد استفاده از الگوریتم PSO به تنهایی، منجر به تعداد تکرارهای زیاد و همگرایی کند می‌شود. به همین دلیل می‌توان این دو روش را ترکیب کرد.

در الگوریتم ترکیبی از روش PSO برای استنتاج مراکز اولیه در الگوریتم k-means استفاده می‌شود .در ابتدا با تعداد تکرارهای محدود به جوابی نزدیک به جواب بهینه همگرا می‌شویم. بعد از به‌دست آمدن مراکز الگوریتم توسط روش PSO ، الگوریتم k-means از آنها برای شروع یک فرآیند خوشه‌بندی بهینه استفاده می‌کند. در این الگوریتم مانند قبلی موقعیت هر ذره به‌صورت $x_i = (m^{(1)},\ldots,m^{(k)})$ نشان داده می‌شود. به‌طوری که $m^{(k)}$ مرکز خوشه k ام است. موقعیتی که همه ذرات به آن همگرا می‌شوند، مراکزی خواهد بود که ما برای شروع خوشه‌بندی k-means از آن استفاده می‌کنیم[۵، ۶].

۶. چالش‌ها و نیازمندی‌ها

تحلیل خوشه‌ای همواره با چالش‌ها و نیازهای بالقوه‌ای همراه است که از جمله آن می‌توان موارد زیر را برشمرد[۷، ۸]:

مقیاس‌پذیری: بسیاری از الگوریتم‌های خوشه‌بندی روی مجموعه داده‌ها با حجم کم(کمتر از چند صد داده) به‌خوبی عمل می‌کنند اما پایگاه داده‌های بزرگ ممکن است شامل میلیون‌ها داده باشند. خوشه‌بندی روی یک نمونه از یک مجموعه داده بزرگ ممکن است منجر به نتیجه نادرست شود. بنابراین نیاز به الگوریتم‌های مقیاس پذیر احساس می‌شود.

توانایی بررسی گونه‌های مختلف داده: بسیاری از الگوریتم‌ها برای داده‌های عددی ایجاد شده‌اند در صورتی که در بسیاری از کاربردها نیاز به خوشه‌بندی انواع دیگر داده مانند داده‌های باینری، اسمی، ترتیبی و یا حتی ترکیبی از آنها است.

ایجاد خوشه‌هایی با شکل دلخواه: تعداد زیادی از الگوریتم‌ها خوشه‌ها را بر اساس معیارهای فاصله مانند فاصله اقلیدسی یا منهتن، ایجاد می‌کنند. این نوع الگوریتم‌ها در نهایت به خوشه‌هایی کروی شکل و با چگالی یکسان می‌رسند این در حالی است که یک خوشه می‌تواند به هر شکلی باشد. بنابراین توسعه الگوریتم-

هایی که خوشه‌هایی با اشکال دلخواه و متفاوت شناسایی می‌کنند، اهمیت زیادی خواهد داشت. الگوریتم‌های تفکیکی معمولاً برای شناسایی خوشه‌های کروی شکل مناسبند.

توانایی بررسی داده‌های نویزی: بیشتر پایگاه داده‌ها در دنیای واقعی شامل داده‌ای آغشته به غلط، نا شناخته، داده‌های پرت یا داده‌های از دست رفته می‌باشند. بعضی از الگوریتم‌های خوشه‌بندی به اینگونه داده‌ها حساس هستند و ممکن است خوشه‌هایی با کیفیت
پایین نتیجه دهند.

بعد پذیری بالا: یک پایگاه یا انباره داده ممکن است دربردارنده داده‌های چندبعدی باشند. بسیاری از الگوریتم‌های خوشه‌بندی در بررسی داده‌ها با ابعاد کوچک، به عنوان مثال دو یا سه بعد، به‌خوبی عمل می‌کنند به همین دلیل کشف خوشه‌ها در فضای داده با ابعاد بالا یکی از چالش‌های خوشه‌بندی محسوب می‌شود.

حداقل کردن نیاز به تعیین پارامترهای وروری: بسیاری از الگوریتم‌ها نیازمند این هستند که کاربر پارامترهای معین مثل تعداد خوشه‌ها را به عنوان پارامتر ورودی ارائه دهد و این مسئله منجر به این خواهد شد که نتیجه خوشه‌بندی حساس به پارامترهای ورودی باشد.

تعیین کردن این پارامترها بخصوص در مجموعه‌هایی با داده‌های چندبعدی اغلب دشوار خواهد بود و نه تنها بار اضافی بر کاربر تحمیل می‌کند بلکه کنترل کیفیت فرآیند خوشه‌بندی را نیز دشوار خواهد کرد.

قابلیت تفسیر و قابلیت استفاده: کاربران انتظار دارند که نتایج خوشه‌بندی قابل تفسیر، قابل درک و قابل استفاده باشد. یعنی خوشه‌بندی ممکن است مقید به هدف یا کاربرد خاصی باشد. بنابراین بررسی این مسئله که چگونه یک هدف کاربردی ممکن است انتخاب شاخص‌ها و متدهای خوشه‌بندی را تحت تأثیر قرار دهد، اهمیت زیادی خواهد داشت.

۷.مقایسه الگوریتم‌ها

هرکدام از الگوریتم‌های شرح داده شده مزایا و معایبی دارند. بعضی از آنها به‌منظور رفع الگوریتم قبلی توسعه داده شده‌اند. نقصی که می‌توان برای الگوریتم k-means مثال زد حساسیت به داده‌های پرت و محدود بودن آن برای داده‌های عددی است. الگوریتم k-mediods پارامتر اول و الگوریتمی مانند k-modes پارامتر دوم را بهبود بخشیده‌است. اما الگوریتم‌های مذکور تنها برای داده‌های عددی یا اسمی مؤثر عمل می‌کنند و نیاز به الگوریتمی که احساس می‌شود که بتواند داده‌های ترکیبی را نیز تحت پوشش قرار دهد. الگوریتمی مانند k-prototypes به این منظور ایجاد شده‌است. از دیگر معایبی که می‌توان ذکر کرد وابسته بودن نتایج به انتخاب مراکز اولیه و انجام یک جستجوی محلی در برخی الگوریتم‌هاست که به‌منظور برطرف کردن آن الگوریتم‌های بهینه‌سازی مانند PSO توسعه داده‌شده‌اند.

جدول(۱) مقایسه الگوریتم‌های شرح داده شده را از لحاظ چند پارامتر مهم نشان می‌دهد. هر الگوریتم از نظر یک یا چند پارامتر بررسی شده‌است. خانه‌های خالی جدول بیانگر این است که در مورد آن الگوریتم خاص پارامتر مورد نظر، یا بررسی نشده و یا از اهمیت چندانی برخوردار نیست.

جدول(۱) بررسی پارامترهای مهم در الگوریتم‌های خوشه‌بندی تفکیکی

پارامتر / الگوریتم	حساسیت به نویز و داده‌های پرت	نوع داده متناسب	حجم داده‌ها	جستجوی محلی
k-means	زیاد	داده-های عددی	مناسب برای داده-های حجیم	انجام جستجوی محلی
k-mediod	کم	داده-های عددی	مناسب برای داده-ها با حجم کم	
CLARA		داده-های عددی	مناسب برای داده-های حجیم	
k-modes		داده-های اسمی		انجام جستجوی محلی
k-prototype		داده-های ترکیبی		
FCM		داده-های عددی		انجام جستجوی محلی
PSO-based clustering		داده-های عددی	مناسب برای داده-ها با حجم کم	انجام جستجوی سراسری
Hybrid PSO clustering		داده-های عددی	مناسب برای داده-های حجیم	انجام جستجوی سراسری

۸.نتیجه‌گیری

داده‌کاوی یک تکنولوژی نوظهور می‌باشد که همراه با توسعه تکنولوژی پایگاه داده‌ها، ایجاد و استفاده شده‌است. می‌توان گفت این تکنیک یک ابزار رایج برای تحلیل و استخراج اطلاعات از بین حجم زیادی از داده‌هاست که از طریق آن می‌توان بدون دخالت کاربران الگوهای مفیدی را برای اطلاعات پنهان به‌دست آورد. تکنیک خوشه‌بندی در زمینه‌های مختلف به‌منظور کشف الگوهای پنهان در داده‌ها به‌کار می‌رود. این تکنیک داده‌ها را بدون نیاز به برچسب‌زنی یک مجموعه داده آموزشی، به گروه‌های مجزایی تقسیم می‌کند و از این حیث نسبت به کلاس‌بندی داده‌ها، به‌خصوص برای مجموعه داده های حجیم، هزینه کمتری در بردارد. الگوریتم‌های خوشه‌بندی را می‌توان به چند دسته کلی تقسیم کرد که از

جمله می‌توان الگوریتم‌های خوشه‌بندی تفکیکی، الگوریتم‌های سلسله‌مراتبی و الگوریتم‌های مبتنی بر چگالی را نام برد. در این مقاله برخی از الگوریتم‌های مهم از دسته الگوریتم‌های تفکیکی را شرح داده و هرکدام از آنها را از لحاظ چند پارامتر مهم بررسی کرده و مزایا و معایب هرکدام را عنوان کردیم. الگوریتم‌های متعددی بر پایه این الگوریتم‌ها توسعه داده‌شده‌اند که هرکدام سعی در بهبود پارامتر(های) خاصی داشته‌اند.

REFERENCES

1 Han J, Kamber M. Data mining concept and techniques. 2006. Elsevier.

2 Anil KJ. Data clustering: 50 years beyond K-means. Pattern Recognition Letters. 2009.

3 Zhexue H. Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values.

4 Velmurugan T, Santhanam T. A survey of partition based clustering algorithms in data mining: An experimental approach. Information Technology Journal. 2011: 478-484.

5 Ahmadi A, Karry F, Kamel M. Flocking based approach for data clustering. 2009. Springer.

6 Xiaohui C, Potok TE, Palathingal P. Document clustering using particle swarm optimization. 2005. IEEE.

7 Pradeep R, Singh S. A survey of clustering technigues. International Journal of Computer Application. 2010; 7(12).

8 Agarwal P, Afshar AM, Ranjit B. Issues, Challenges and Tools of Clustering Algorithms. International Journal of Computer Science. 2011; 8(3).