مطالعات علوم کاربردی در مهندسی
دوره ۱۱، شماره ۱، بهار ۱۴۰۴
صفحات ۱۷۷-۱۶۲

Online ISSN: ٢٤٧٦-٤٥٠٧
Print ISSN: ٢٥٣٨-٢٠٥٥
www.irijournals.com

ترکیب الگوریتم‌های متاهیوریستیکی چیتا و ازدحام ذرات با هدف افزایش صحت خوشه‌بندی داده‌ها

مهسا بحرینی نژادی
کارشناسی ارشد مهندسی کامپیوتر- نرم‌افزار دانشگاه آزاد اسلامی واحد تنگستان (اهرم)

چکیده
امروزه، خوشه‌بندی نقش مهی را در اغلب زمینه‌های تحقیقاتی مانند مهندسی، پزشکی، زیست‌شناسی، داده‌کاوی و ... ایفا می‌نماید. در واقع خوشه‌بندی به معنای تقسیم‌بندی بدون نظارت می‌باشد؛ با استفاده از آن داده‌ها به دسته‌هایی که ازنظر پارامترهای موردعلاقه، شباهت بیشتری به یکدیگر دارند، تقسیم می‌گردند. امروزه به‌منظور افزایش صحت خوشه‌بندی، از روش‌های دومرحله‌ای استفاده می‌شود. به همین منظور، در این پژوهش برای خوشه‌بندی بهینه داده‌ها، از تکنیک دومرحله‌ای ازدحام ذرات معلق و چیتا استفاده شده است. در مرحله اول، الگوریتم ازدحام ذرات، تعداد خوشه‌های بهینه را به‌صورت اتوماتیک با استفاده از تابع برازش مجموع فاصله‌های اقلیدسی SED[1] استخراج می‌نماید. خوشه‌های به‌دست‌آمده از الگوریتم ازدحام ذرات، جهت بهینه‌سازی، به الگوریتم فرا ابتکاری چیتا ارسال و در مرحله دوم نیز از شاخص SED برای ارزیابی هزینه خوشه‌بندی استفاده شده است. به‌منظور ارزیابی روش پیشنهاد شده، آزمایش‌های مختلفی بر روی دیتاست های متنوعی انجام شده است. پس از ارزیابی نتایج، ثابت گردید که کارایی روش ارائه شده، مورد قبول می‌باشد.

کلمات کلیدی: داده‌کاوی، خوشه‌بندی، الگوریتم ازدحام ذرات، مجموع فاصله‌های اقلیدسی، الگوریتم فرا ابتکاری چیتا

[1] - Sum of Euclidean distances

دوره ۱۱، شماره ۱، بهار ۱۴۰۴، صفحات ۱۷۷-۱۶۲

مقدمه

کلاسترینگ یا خوشه‌بندی از جمله الگوریتم‌های قطعه بندی به حساب می آید. الگوریتم خوشه‌بندی، اطلاعاتی را که ویژگی های نزدیک به هم و مشابه دارند را در دسته های جداگانه که به آن خوشه گفته می‌شود قرار می دهد. به بیان دیگر خوشه‌بندی همان دسته بندی های ساده ای است که در کارهای روزانه انجام می دهیم. تحلیل خوشه یکی از فعالیت های مهم انسان می‌باشد. در کودکی ما با بهبود ناخود آگاه خوشه بندی در ذهنمان یاد می گیریم که چگونه بین سگ ها و گربه ها، یا حیوانات و گیاهان تفاوت قائل شویم. خوشه بندی اغلب به عنوان اولین گام و یکی از مهمترین روش های تحلیل داده‌ها بشمار می آید. خوشه بندی فرآیندی است که در آن اشیاء در گروه هایی از اشیاء مشابه دسته بندی می شوند. هر گروه یا خوشه شامل اشیائی است که شبیه به یکدیگرند و متفاوت از اشیاء گروه های دیگر می باشند.

مبانی نظری

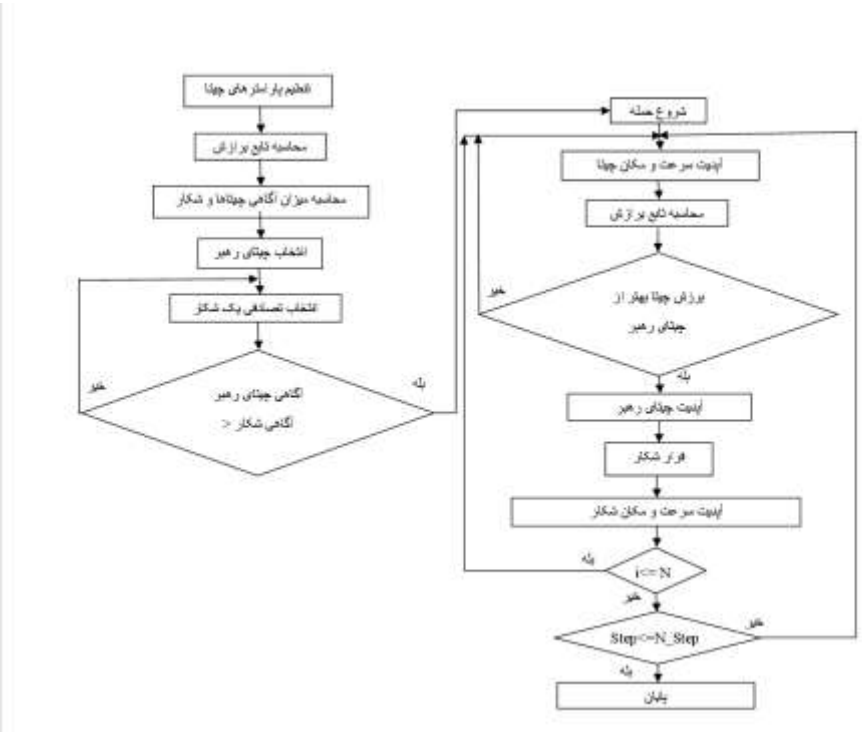
الگوریتم ازدحام ذرات

الگوریتم ازدحام ذرات یا Particle Swarm optimization که به اختصار PSO نامیده می‌شود، یکی از روش‌های بهینه‌سازی است. این الگوریتم در سال ۱۹۹۵ توسط آقایان کندی و ابرهات توسعه یافت.

این الگوریتم الهام گرفته از رفتار اجتماعی گونه های مختلف زیستی در زندگی گروهی آنان است. مانند پرندگان ، ماهی ها یا کلونی های مورچه ها. این الگوریتم تعامل بین اعضا را برای به اشتراک گذاری اطلاعات تقلید می کند. الگوریتم ازدحام ذرات کاربرد وسیعی در حوزه های متعددی در بهینه‌سازی و در ترکیب با دیگه الگوریتم‌های بهینه‌سازی داشته است.

الگوریتم چیتا

چیتاها شکارچیانی باهوش هستند که برای سرعت ساخته شده اند. سری کوچک با مقاومت ناچیز هوا در برابر آن، پاهای بلند و دمی عضلانی برای حفظ تعادل دارند. قلب و شش هایشان برای اکسیژن رسانی به عظلات هنگام دویدن، بزرگتر از حد معمول است. نیمی از توده عضلاتشان در اطراف ستون فقرات قرار گرفته است تا بتوانند مانند فنر منعطف باشند. این باعث می‌شود تا طول گام هایشان بلندتر شود و در رسیدن به حداکثر سرعت به آنها کمک می کند. برای چیتا فقط ۳ گام کافی است تا سرعتش از صفر به 65 km/h برسد و فقط ۳ ثانیه نیاز است تا سرعتش به ۱۱۰km/h برسد. چیتاها برای شکارکردن نیاز به حواس هماهنگ دارند. اگرچه چیتا بسیار سریع است، ولی باید قبل از اقدام به حمله، به اندازه کافی به شکار نزدیک شده باشد. با پیدا کردن شکار، کاملا متمرکز می‌شود. کار گروهی احتمال موفقیت را افزایش می دهد. وقتی یکی خسته می‌شود، دیگری کار تعقیب را ادامه می دهد. چیتاهای نر معمولا به صورت گروهی شکار می کنند. شکل ۱ مراحل الگوریتم چیتا را بازنمایی می کند. این الگوریتم در سال ۲۰۲۱ توسط Ghaedi و همکارانش ارائه گردیده است.



شکل ۱. مراحل الگوریتم چیتا

همانطور که مشاهده می کنید، پس از مقداردهی اولیه، چیتای رهبر انتخاب و عملیات حمله به یک شکار تصادفی آغاز می‌شود. در جریان حمله، چیتای رهبر و شکار، سرعت و تغییر مکان خود را متناسب با یکدیگر تغییر می دهند. در این عملیات، ممکن است چیتای رهبر تغییر کند. سرعت و مکان شکار نیز در حین فرار تغییر می کند.

روش تحقیق

ابتدا با مطالعه روش‌های پایه ای خوشه‌بندی و نیز تحلیل الگوریتم‌های متاهیوریستیکی، با ترکیب دو الگوریتم فرا ابتکاری ازدحام ذرات PSO و چیتا CHEETAH، روشی دومرحله‌ای ارائه می‌شود.

در مرحله اول، برای حدس اولیه خوشه ها و خوشه‌بندی داده‌ها، از الگوریتم فرا ابتکاری PSO برای خوشه‌بندی داده‌ها استفاده می‌شود. برای ارزیابی هزینه خوشه‌بندی، از شاخص مجموع فاصله‌های اقلیدسی استفاده می‌شود. دلیل استفاده از الگوریتم PSO در مرحله اول این است که با توجه به اینکه این الگوریتم ممکن است در بهینه محلی به دام بیفتد و نتیجه مطلوب نهایی حاصل نشود، چنانچه این اتفاق افتاد، الگوریتم چیتا این ضعف را پوشش دهد.

پس از حدس اولیه مراکز خوشه ها توسط الگوریتم PSO، در مرحله دوم از الگوریتم فرا ابتکاری چیتا برای خوشه‌بندی نهایی استفاده می‌شود. این الگوریتم در آزمایش‌های متعددی، عملکرد بهتری نسبت به سایر الگوریتم‌های فرا ابتکاری داشته است.

عملکرد روش پیشنهادی با استفاده از آزمایش‌های متعددی مورد ارزیابی قرار می گیرد و از دیتاست های مرجع برای آزمون استفاده می‌شود.

فرضیه ها

- با ترکیب الگوریتم‌های PSO و چیتا، کارایی خوشه‌بندی داده‌ها افزایش می یابد.
- صحت خوشه‌بندی روش‌های ترکیبی بیشتر از روش‌های سنتی و پایه می‌باشد.

اهداف تحقیق

اهداف علمی:

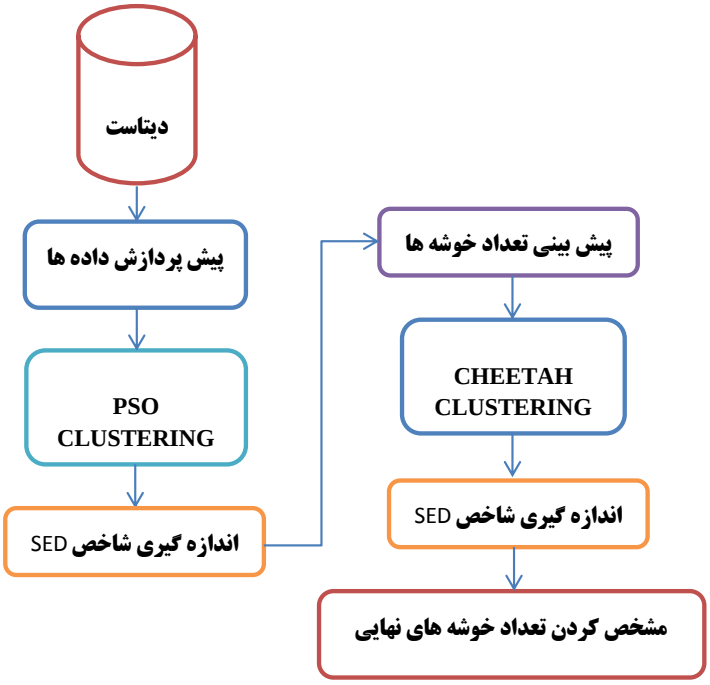
- افزایش کارایی خوشه‌بندی داده‌ها
- روش ترکیبی ارائه شده در تشخیص نویزها بسیار کارا می‌باشد.

اهداف کاربردی:

استفاده از روش پیشنهاد شده خوشه‌بندی در حوزه های مرتبط با خوشه‌بندی داده‌ها

روش پیشنهادی

مراحل چارچوب پیشنهادی در شکل ۲ نشان داده شده است. در اولین مرحله پیش از استفاده از داده‌ها، عملیات پیش پردازش داده‌ها انجام می‌شود. یکی از مهم ترین عملیات پیش پردازش داده‌ها، نرمالسازی آنهاست. نرمالسازی باعث می‌شود همه داده‌ها در یک محدوده مشخص قرار گیرند و خروجی نهایی صحت بالاتری داشته باشد. در این تحقیق از تکنیک نرمالسازی به روش MIN-MAX استفاده می‌شود.



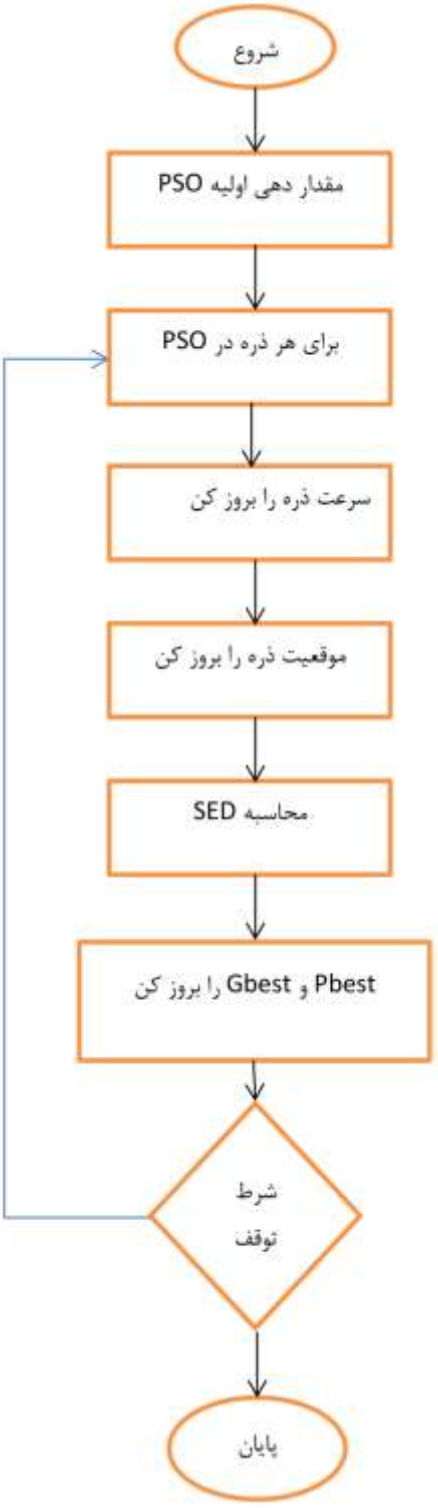
شکل ۲. مراحل روش پیشنهادی

به‌منظور نتیجه بهتر، ابتدا داده‌ها پیش پردازش می شوند. به‌منظور جایگزینی داده‌های ناموجود، از روش مقدار میانگین ویژگی و برای نرمالسازی نیز روش Min-Max استفاده می‌شود.

پس از نرمالسازی، از الگوریتم PSO برای خوشه‌بندی و استخراج خوشه‌های اولیه استفاده می‌شود. شکل ۳ مراحل خوشه‌بندی با الگوریتم PSO را نشان می دهد. همانطور که مشاهده می‌شود، پس از مقدار دهی اولیه ذرات، در یک حلقه، مراحل بروزرسانی سرعت و موقعیت ذرات مکررا اتفاق می افتد. برای هر ذره تابع SED فراخوانی و هزینه خوشه‌بندی محاسبه می‌شود. در صورتیکه

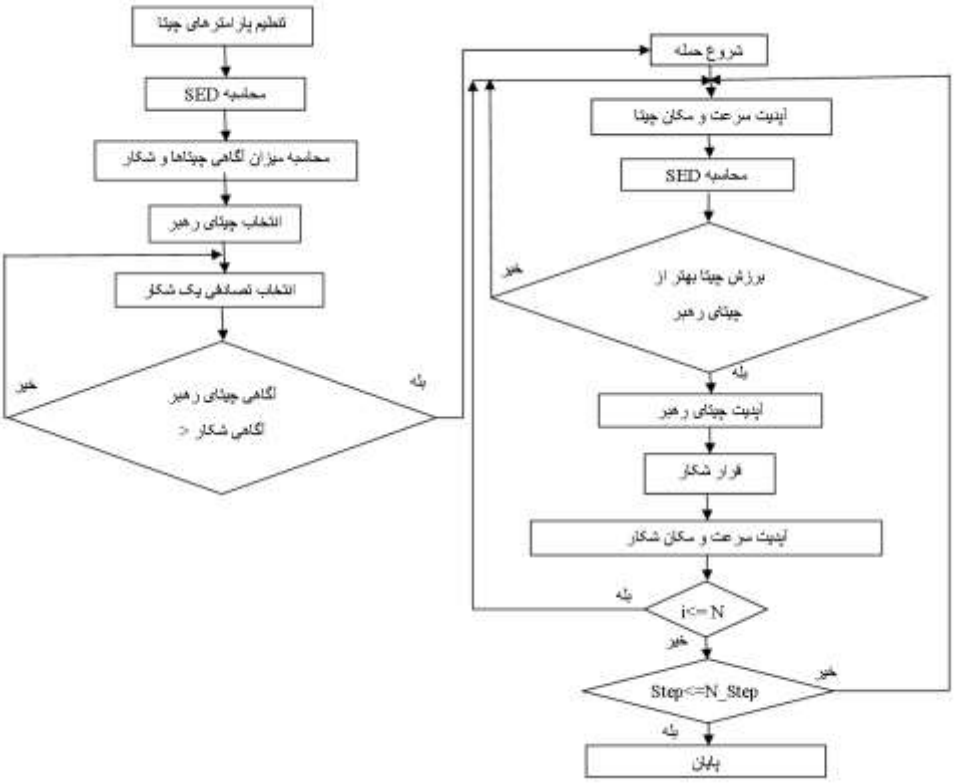
شرط توقف محقق گردد، الگوریتم PSO کار خوشه‌بندی را تمام می کند و تعداد خوشه ها و داده‌های متعلق به هر خوشه استخراج می‌شود.

تعداد خوشه‌های استخراج شده به‌منظور بررسی و خوشه‌بندی نهایی، به الگوریتم چیتا ارسال می‌شود. از آنجا که این الگوریتم دو تغییر مکان برای هم چیتاها و هم شکار دارد، عملکرد بسیار خوبی برای عملیات بهینه‌سازی دارد.



شکل۳.مراحل خوشه‌بندی با الگوریتم PSO

شکل ۴ مراحل خوشه‌بندی نهایی با استفاده از الگوریتم چیتا را به تصویر کشیده است.



شکل۴.مراحل خوشه‌بندی با الگوریتم چیتا

همانطور که در شکل ۴ مشاهده می کنید، در ۲ مرحله، مقدار SED محاسبه می‌شود. در واقع از آن برای محاسبه هزینه خوشه‌بندی الگوریتم چیتا استفاده می‌شود. هر چقدر این مقدار کمتر باشد، هزینه خوشه‌بندی کمتر و نتیجه مطلوب تر است.

الگوریتم چیتا با خوشه‌بندی دوباره داده‌ها، آن دسته از داده‌هایی که ممکن است به اشتباه در مرحله اول توسط الگوریتم PSO خوشه‌بندی شده باشند را دوباره خوشه‌بندی می‌نماید.

دیتاست

برای انجام آزمایشات و ارزیابی نتایج، از دیتاست های مرجع بدون برچسب Sales_Transactions_Dataset_Weekly و Facebook Live Sellers in Thailand و برچسب دار Vowel، Iris، Wine و Sonar استفاده شده است.

دیتاست[36]Sales_Transactions_Dataset_Weekly شامل مقادیر خرید هفتگی ۸۰۰ محصول در طول ۵۲ هفته است. مقادیر به‌صورت نرمال ارائه شده است. این دیتاست حاوی ۸۱۱ نمونه با ۵۳ ویژگی می‌باشد که در کل ۴-۱ نشان داده شده است.

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	811	**Area:**	N/A
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	53	**Date Donated**	2017-07-16
Associated Tasks:	Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	95913

شکل۵.اطلاعات مربوط به دیتاست Sales_Transactions_Dataset_Weekly

دیتاست Facebook Live Sellers in Thailand[37] شامل صفحات فیس بوک ۱۰ خرده فروش مد و لوازم آرایشی تایلندی ونیز پست هایی با ماهیت متفاوت (فیلم، عکس، وضعیت، و لینک) است. معیارهای تعامل شامل نظرات، اشتراک‌گذاری‌ها و واکنش‌ها هستند.

Data Set Characteristics:	Multivariate	**Number of Instances:**	7051	**Area:**	Business
Attribute Characteristics:	Integer	**Number of Attributes:**	12	**Date Donated**	2019-04-22
Associated Tasks:	Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	61155

شکل۶.اطلاعات مربوط به دیتاست Facebook Live Sellers in Thailand

این دیتاست شامل ۷۰۵۱ نمونه با ۱۲ ویژگی می‌باشد. تنوع تعامل مصرف کننده از طریق تکنیک PCA و برجسته کردن تغییرات ناشی از بکارگیری Facebook Live تجزیه و تحلیل می‌شود. مؤلفه فصلی از طریق مطالعه میانگین‌های معیارهای تجاری مختلف برای بازه‌های زمانی مختلف (ساعتی، روزانه و ماهانه) تحلیل می‌شود. در نهایت، پست‌های پرت آماری شناسایی می شوند که ازنظر رویکرد و فعالیت‌های فروش، بیشتر مورد تجزیه و تحلیل کیفی قرار می‌گیرند.

اطلاعات مربوط به دیتاست های برچسب دار Vowel، Iris، Wine و Sonar در جدول ۱ نشان داده شده است که از سایت UCI قابل دسترس می‌باشد.

جدول۱.دیتاست های مرجع برچسب دار(کلاس دار)

دیتاست	تعداد ویژگی	تعداد رکورد	تعداد کلاس
Vowel	۱۰	۵۲۸	۱۱
Iris	۴	۱۵۰	۳
Wine	13	175	3
Sonar	60	208	2

شبیه سازی و تحلیل نتایج

شبیه سازی با نرم‌افزار متلب۲۰۲۰ انجام شده است. از معیارهای Accuracy، Precision، Recall و F-score برای ارزیابی دیتاست های دارای برچسب و از شاخص SED برای ارزیابی دیتاست های بدون برچسب استفاده شده است. این معیارها از ماتریس درهمریختگی قابل محاسبه است. ماتریس در هم ریختگی مطابق جدول ۲ نشان داده می‌شود.

جدول۲.ماتریس درهمریختگی

Actual label \ Predicted label	Positive	Negative
Positive	TP	FN
Negative	FP	TN

$$n = TP + TN + FP + FN$$

معیار صحت

معیار صحت(Accuracy) ، بیان کننده تعداد «پیش‌بینی‌های صحیح انجام شده» توسط دسته‌بند، تقسیم بر، تعداد «کل پیش‌بینی‌های انجام شده» توسط همان دسته‌بند است.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

معیار Accuracy به تنهایی، معیار مناسبی برای ارزیابی عملکرد یک دسته‌بند نیست. زمانی که کلاس‌های موجود در داده‌ها نامتوازن (Imbalanced) باشند (یعنی، تعداد داده‌های متعلق به کلاس (برچسب) خاص از کلاس‌های دیگر بسیار بیشتر باشد)، ممکن است سیستم با پدیده خاصی به نام تناقض صحت (Paradox Accuracy) مواجه شود. در نتیجه این تناقض، مدل دسته‌بند به احتمال زیاد عملکرد بسیار خوبی در پیش‌بینی کلاس (برچسب) داده‌ها از خود نشان می‌دهد؛ زیرا، اکثریت داده‌ها تنها به یکی از کلاس‌ها تعلق دارند. در صورتی که چنین پدیده‌ای رخ دهد، بهتر است که معیارهای دیگری نظیر فراخوانی (Recall)ودقت (Precision) برای ارزیابی عملکرد سیستم در نظر گرفته شوند.

معیار یادآوری

معیاریادآوری(Recall) ، بیان کننده نسبت «تعداد داده‌ها ی درست دسته‌بندی شده» در یک کلاس خاص، به تعداد کل داده‌هایی است که باید در همان کلاس خاص دسته‌بندی شوند. مقدار بالا برای معیار یادآوری، بیانگر تعداد کم داده‌هایی است که به اشتباه، در آن کلاس خاص دسته‌بندی نشده‌اند. استفاده از این معیار، به تنهایی، برای ارزیابی عملکرد سیستم درست نیست و باید در کنار معیاردقت (Precision) مورد استفاده قرار بگیرد. زیرا، به راحتی می‌شود مدل‌های دسته‌بندی متنی طراحی کرد که یادآوری بالایی داشته باشند و این لزوما به معنای دقت (Precision) بالا نیست.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

معیار دقت

معیار دقت(Precision) ، نسبت تعداد «پیش‌بینی‌های صحیح انجام شده» برای نمونه‌های یک کلاس خاص، به تعداد «کل پیش‌بینی‌ها» برای نمونه‌های همان کلاس خاص را (این تعداد، مجموع تمامی پیش‌بینی‌های صحیح و پیش‌بینی‌های نادرست را شامل می‌شود) ارزیابی می‌کند. مقدار بالا برای معیار دقت، بیانگر تعداد کم داده‌هایی است که به اشتباه، در کلاس خاص دسته‌بندی شده‌اند. شایان توجه است که معیار دقت، فقط برای مواردی ارزیابی می‌شود که در آن‌ها، مدل دسته‌بندی تعلق یک نمونه به یک کلاس خاص را پیش‌بینی کرده باشد. برای برخی از فعالیت‌ها، مدل‌هایی نیاز است که سطح دقت آن‌ها بالا باشد؛ به عبارت دیگر، پاسخ‌ها تنها باید زمانی به کاربران ارسال شوند که مدل دسته‌بندی، با احتمال بالا، پیش‌بینی‌های درستی انجام داده باشد. در هنگام ارزیابی عملکرد یک مدل دسته‌بندی متن، بهتر است که از این معیار در کنار معیار یادآوری (Recall) استفاده شود.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

معیار امتیازF-score

این معیار، پارامترهای دقت (Precision) ویادآوری (Recall) را با هم ترکیب می‌کند تا مشخص شود یک مدل دسته‌بند تا چه حد عملکرد خوبی از خود نشان می‌دهد. به این معیار، میانگین متوازن (Harmonic Mean) دو معیار دقت (Precision)و یادآوری (Recall) نیز گفته می‌شود. این معیار، نسبت به معیار صحت(Accuracy) ، تصویر دقیق‌تری از نحوه عملکرد مدل دسته‌بند روی تمامی کلاس‌های موجود در داده‌ها ترسیم می‌کند.

$$\text{F} - \text{score} = 2 * \text{Precision} * \frac{\text{Recall}}{\text{Precision} + \text{Recall}}$$

آزمایشات

مرحله اول(الگوریتم PSO):

ابتدا خوشه‌بندی داده‌ها را با استفاده از الگوریتم PSO انجام می دهیم و شاخص SED را برای آن اندازه گیری می کنیم. جدول ۳ مقادیر شاخص SED را برای الگوریتم‌های فرا ابتکاری PSO، ACO، [38]CSA و [39]ICSA نشان می دهد. در

واقع برای ارزیابی کارایی الگوریتم PSO در مرحله اول، عملکرد آن را با شاخص SED با سایر الگوریتم ها مقایسه کردیم. همانطور که مشاهده می‌شود، برای هر دو دیتاست، شاخص SED برای الگوریتم PSO کمترین مقدار را نشان می دهد.

جدول۳.شاخص SED الگوریتم‌های مختلف بر روی دیتاست های بدون برچسب مرجع

دیتاست / الگوریتم	Sales_Transactions_Dataset_Weekly	Facebook Live Sellers in Thailand
ACO	0.345	۴۴,۸
PSO	۰,۱۲۵	۲۲,۱۴
CSA	0.287	۳۲,۷۸
ICSA	0.219	۲۸,۱۵

جدول ۴، تعداد خوشه‌های بهینه به‌دست‌آمده الگوریتم‌های مختلف برای دیتاست های مرجع بدون برچسب Sales_Transactions_Dataset_Weekly وFacebook Live Sellers in Thailand نشان می دهد.

جدول۴.تعداد خوشه‌های بهینه به‌دست‌آمده الگوریتم‌های مختلف برای دیتاست های مرجع بدون برچسب

دیتاست / الگوریتم	Sales_Transactions_Dataset_Weekly	Facebook Live Sellers in Thailand
ACO	۶۸	۲۲
PSO	۵۲	۱۵
CSA	۵۷	۱۹
ICSA	۵۵	۱۸

مقادیر SED مربوط به دیتاست های برچسب دار Vowel، Iris، Wine و Sonar در جدول ۵ نشان داده شده است.

جدول۵. مقادیر SED دیتاست های برچسب دار

دیتاست / الگوریتم	Sonar	Vowel	Iris	Wine
ACO	۱۱۹	۰,۶۷	۱۸,۶	۰,۵۶
PSO	۸۹	۰,۴۵	۱۱,۲	۰,۱۷
CSA	۲۲۱	۰,۸۹	۲۲,۹	۰,۴۱
ICSA	۱۱۷	۰,۷۱	۱۸,۸	۰,۳۵

جدول ۶ تعداد خوشه‌های بهینه به‌دست‌آمده توسط الگوریتم‌های مختلف را برای دیتاست های مرجع برچسب دار نشان می دهد.

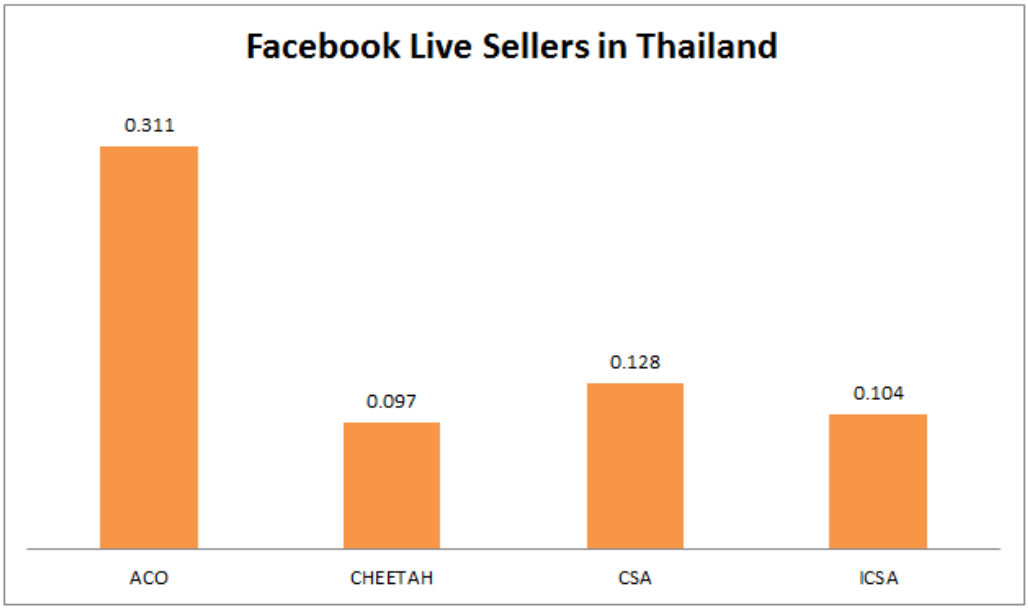
جدول ۶. تعداد خوشه‌های به‌دست‌آمده مربوط به دیتاست های برچسب دار

دیتاست / الگوریتم	Sonar	Vowel	Iris	Wine
ACO	۴	۱۴	۵	۶
PSO	۳	۱۲	۴	۵
CSA	۴	۱۵	۵	۶

ICSA	۳	۱۶	۴	۷

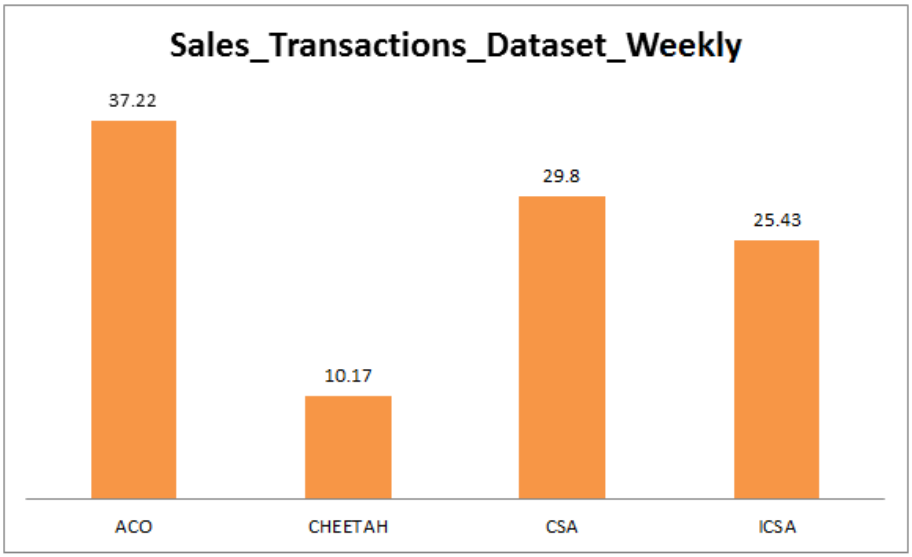
مرحله دوم(الگوریتم چیتا):

در مرحله دوم روش پیشنهادی از الگوریتم بهینه‌سازی چیتا برای استخراج خوشه‌های نهایی استفاده شده است. شکل۷، مقادیر معیار SED را برای الگوریتم‌های مختلف بر روی دیتاست مرجع بدون برچسب Facebook Live Sellers in Thailand نشان می دهد.



شکل۷.مقادیر SED الگوریتم‌های مختلف برروی دیتاست Facebook Live Sellers in Thailand

شکل ۸ نیز مقادیر معیار SED را برای الگوریتم‌های مختلف بر روی دیتاست مرجع بدون برچسب Sales_Transactions_Dataset_Weekly نشان می دهد.



شکل۸.مقادیر SED الگوریتم‌های مختلف برروی دیتاست Sales_Transactions_Dataset_Weekly

جدول ۷ تعداد خوشه‌های نهایی به‌دست‌آمده توسط الگوریتم‌های مختلف را برای دیتاست Facebook Live Sellers in Thailand و Sales_Transactions_Dataset_Weekly نشان می دهد.

جدول۷.خوشه‌های نهایی به‌دست‌آمده توسط الگوریتم‌های مختلف برای دیتاست های بدون برچسب

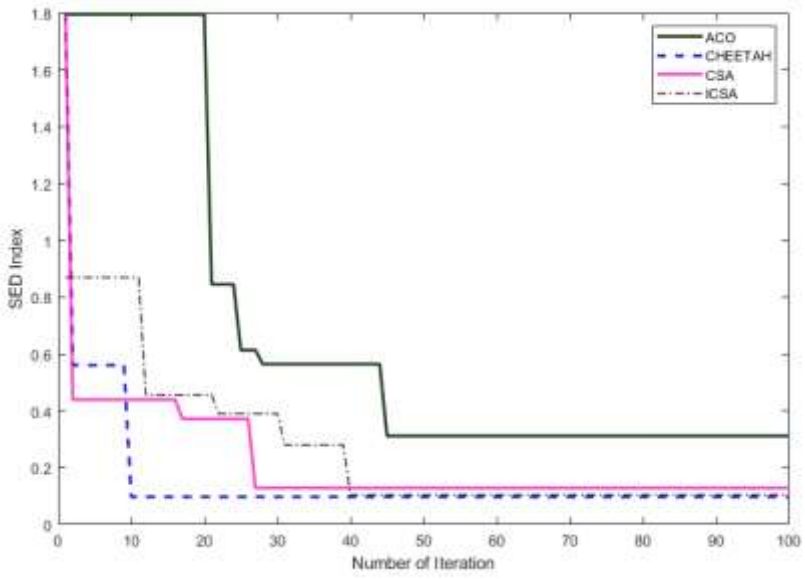
دیتاست / الگوریتم	Sales_Transactions_Dataset_Weekly	Facebook Live Sellers in Thailand
ACO	۶۴	۱۸
CHEETAH	45	۱۰
CSA	۵۴	۱۶
ICSA	۵۲	۱۵

جدول ۸ نیز تعداد خوشه‌های بهینه به‌دست‌آمده توسط الگویتم های مختلف را برای دیتاست های برچسب دار نشان می دهد.

جدول ۸. تعداد خوشه‌های بهینه به‌دست‌آمده توسط الگویتم های مختلف را برای دیتاست های برچسب دار

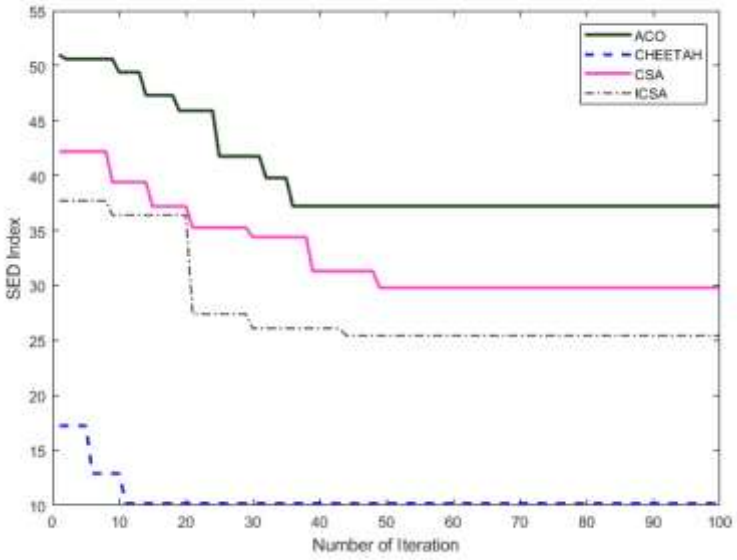
دیتاست / الگوریتم	Sonar	Vowel	Iris	Wine
ACO	۴	۱۲	۳	۵
CHEETAH	۲	۱۱	۳	۳
CSA	۳	۱۴	۵	۴
ICSA	۳	۱۵	۴	۵

در آزمایشی دیگر چگونگی همگرایی الگوریتم‌های مختلف را برای دیتاست های بدون برچسب نشان می دهیم. شکل روند همگرایی را برای الگوریتم‌های مختلف برروی دیتاست Facebook Live Sellers in Thailand نشان می دهد. با توجه به شکل ۹ روند همگرایی الگوریتم CHEETAH بسیار سریع می‌باشد.



شکل۹.چگونگی همگرایی الگوریتم‌های مختلف را برای دیتاست های بدون برچسب

شکل ۱۰ روند همگرایی الگوریتم CHEETAH را برای دیتاست Sales_Transactions_Dataset_Weekly نشـان مـی دهد. در این دیتاست هم روند همگرایی بسیار سریع می‌باشد و نسبت به بقیه الگوریتم ها سریع تر به جواب نهایی همگرا می‌شود.



شکل۱۰.روند همگرایی الگوریتم CHEETAH برای دیتاست Sales_Transactions_Dataset_Weekly

در آزمایشـی دیگـر، بـا اسـتفاده از روش ترکیبـی پیشـنهادی PSO-CHEETAH، معیارهـای Acuuracy، Precision، Recall و F-score برای دیتاست های برچسب دار Vowel، Iris، Wine و Sonar اندازه گیـری شـده اسـت. جـدول ۹ مقادیر این معیارها را نشان می دهد.

جدول۹.معیارهای کارایی برای دیتاست های برچسب دار با استفاده از روش پیشنهادی

دیتاست / معیار	Vowel	Iris	Wine	Sonar
Accuracy	۰,۹۸۸	۰,۹۸۹	۰,۹۸۳	۰,۹۸۲
Recall	۰,۹۸۴	۰,۹۸۱	۰,۹۸۶	۰,۹۷۴
Precision	۰,۹۸۵	۰,۹۸۴	۰,۹۹	۰,۹۸
F-score	۰,۹۸۴۴۹۹۷۵	۰,۹۸۲۵	۰,۹۸۸	۰,۹۷۶۹۹

در آزمایش آخر، معیار Acuuracy روش پیشنهادی با الگوریتم‌های خوشه‌بندی K-means و Dbscan مقایسه شده است. جدول ۱۰ این مقایسه را برای دیتاست های برچسب دار Vowel، Iris، Wine و Sonar نشان می دهد.

جدول۱-.مقایسه معیار Acuuracy روش پیشنهادی با الگوریتم‌های خوشه‌بندی K-means و Dbscan

الگوریتم / دیتاست	K-means	DBSCAN	PSO-CHEETAH
Vowel	90.17	89.12	۰,۹۸۸
Iris	۸۷,۳۹	۹۰,۲۲	۰,۹۸۹
Wine	۸۹,۶	۸۷,۷۴	۰,۹۸۳
Sonar	۸۸,۱	۹۰,۳۸	۰,۹۸۲

با توجه به نتایج به‌دست‌آمده مشاهده می گردد که روش ترکیبی ارائه شده با کارایی بالایی داده‌ها را خوشه‌بندی می کند.

نتیجه گیری

در این تحقیق به‌منظور خوشه‌بندی بهینه داده‌ها از یک روش ترکیبی استفاده گردید. در روش ارائـه شـده از ترکیـب دو الگـوریتم متاهیوریستیکی PSO و CHEETAH برای خوشه‌بندی استفاده گردید. روش پیشنهادی شامل دو مرحله خوشه‌بنـدی بـود کـه در مرحله اول از الگوریتم ازدحام ذرات PSO و در مرحله دوم از الگوریتم CHEETAH برای خوشه‌بندی نهایی استفاده گردیـد. از شاخص SED برای ارزیابی تابع برازش خوشه‌بندی استفاده گردید.

آزمایش‌های مختلفی انجام گرفت و نتایج مورد ارزیابی قرار گرفت. نتایج به‌دسـت‌آمـده نشـان داد کـه روش پیشـنهادی بـا کـارایی بالایی، داده‌ها را خوشه‌بندی می کند.

به عنوان کار آینده می توان از شاخص های مختلف خوشخ بندی برای تابع برازش استفاده کرد. همچنین به‌منظور ارزیابی بیشتر کارایی می توان تعداد خوشه ها را از پیش مشخص شده در نظر گرفت و نتایج به‌دست‌آمده را تحلیل نمود.

[1] A. L. N. Fred and A. K. Jain, “Data clustering using evidence accumulation,” in *2002 International Conference on Pattern Recognition*, 2002, vol. 4, pp. 276–280 vol.4, doi: 10.1109/ICPR.2002.1047450.

[2] A. Strehl and J. Ghosh, “Cluster Ensembles --- A Knowledge Reuse Framework for Combining Multiple Partitions,” *J. Mach. Learn. Res.*, vol. 3, pp. 583–617, 2002.

[3] H. Ralambondrainy, “A conceptual version of the K-means algorithm,” *Pattern Recognit. Lett.*, vol. 16, pp. 1147–1157, 1995.

[4] J. Bezdek, *Feature Selection for Binary Data: Medical Diagnosis with Fuzzy Sets*, vol. 45. 1976.

[5] I. Aljarah, M. Mafarja, A. A. Heidari, H. Faris, and S. Mirjalili, “Clustering analysis using a novel locality-informed grey wolf-inspired clustering approach,” *Knowl. Inf. Syst.*, vol. 62, no. 2, pp. 507–539, 2020, doi: 10.1007/s10115-019-01358-x.

[6] H. Menéndez, F. Barril Otero, and D. Camacho, “Medoid-based clustering using ant colony optimization,” *Swarm Intell.*, vol. 10, Jun. 2016, doi: 10.1007/s11721-016-0122-5.

[7] P. Das, D. Das, and S. Dey, “A Modified Bee Colony Optimization (MBCO) and it’s hybridization with k-means for an application to data clustering,” *Appl. Soft Comput.*, vol. 70, Jun. 2018, doi: 10.1016/j.asoc.2018.05.045.

[8] R. J. Kuo and F. E. Zulvia, “An improved differential evolution with cluster decomposition algorithm for automatic clustering,” *Soft Comput.*, vol. 23, no. 18, pp. 8957–8973, 2019, doi: 10.1007/s00500-018-3496-z.

[9] A. Tripathi, S. Kapil, and B. Manju, “Parallel Bat Algorithm-Based Clustering Using MapReduce,” 2018, pp. 73–82.

[10] O. Tarkhaneh, A. Isazadeh, and H. Jabbari Khamnei, “A new hybrid strategy for data clustering using cuckoo search based on Mantegna levy distribution, PSO and k-means,” *Int. J. Comput. Appl. Technol.*, vol. 58, p. 137, Jan. 2018, doi: 10.1504/IJCAT.2018.094576.

[11] H. Abdulwahab, N. Ahmad, A. Alsewari, and S. Salih, “An Enhanced Version of Black Hole Algorithm Via Levy Flight for Optimization and Data Clustering Problems,” *IEEE Access*, vol. PP, p. 1, Aug. 2019, doi: 10.1109/ACCESS.2019.2937021.

[12] K.-W. Huang, Z.-X. Wu, H.-W. Peng, M.-C. Tsai, Y.-C. Hung, and Y.-C. Lu, “Memetic Particle Gravitation Optimization Algorithm for Solving Clustering Problems,” *IEEE*

Access, vol. 7, pp. 80950–80968, 2019, doi: 10.1109/ACCESS.2019.2923979.

[13] H. Yu, Z. Chang, G. Wang, and X. Chen, "An efficient three-way clustering algorithm based on gravitational search," *Int. J. Mach. Learn. Cybern.*, vol. 11, May 2020, doi: 10.1007/s13042-019-00988-5.

[14] A. Kaur, S. Pal, and A. Singh, "Hybridization of Chaos and Flower Pollination Algorithm over K-Means for data clustering," *Appl. Soft Comput.*, vol. 97, p. 105523, May 2019, doi: 10.1016/j.asoc.2019.105523.

[15] J. Chen, X. Qi, L. Chen, F. Chen, and G. Cheng, "Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection," *Knowledge-Based Syst.*, vol. 203, p. 106167, Jun. 2020, doi: 10.1016/j.knosys.2020.106167.

[16] L. Abualigah, A. T. Khader, E. Hanandeh, and A. Gandomi, "A novel hybridization strategy for krill herd algorithm applied to clustering techniques," *Appl. Soft Comput.*, vol. 60, Jul. 2017, doi: 10.1016/j.asoc.2017.06.059.

[17] S. Majhi, "Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection," *Evol. Intell.*, vol. 14, Mar. 2021, doi: 10.1007/s12065-019-00260-3.

[18] G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: Ten years after.," *SIGKDD Explor.*, vol. 1, pp. 59–61, Jan. 2000, doi: 10.1145/846183.846197.

[19] J. Z. Huang, "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining," 1997.

[20] C.-H. Yun, K.-T. Chuang, and M.-S. Chen, "Adherence clustering: An efficient method for mining market-basket clusters," *Inf. Syst.*, vol. 31, pp. 170–186, May 2006, doi: 10.1016/j.is.2004.11.008.

[21] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS—Clustering Categorical Data Using Summaries," in *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999, pp. 73–83, doi: 10.1145/312129.312201.

[22] L. Bai and J. Liang, "The K-Modes Type Clustering plus between-Cluster Information for Categorical Data," *Neurocomput.*, vol. 133, pp. 111–121, 2014, doi: 10.1016/j.neucom.2013.11.024.

[23] L. Bai, J. Liang, C. Dang, and F. Cao, "A novel fuzzy clustering algorithm with between-cluster information for categorical data," *Fuzzy Sets Syst.*, vol. 215, pp. 55–73, Mar. 2013, doi: 10.1016/j.fss.2012.06.005.

[24] F. Cao, J. Liang, D. Li, and X. Zhao, "A weighting k-modes algorithm for subspace clustering of categorical data," *Neurocomputing*, vol. 108, pp. 23–30, 2013, doi: https://doi.org/10.1016/j.neucom.2012.11.009.

[25] A. Saha and S. Das, "Categorical fuzzy k-modes clustering with automated feature weight learning," *Neurocomputing*, vol. 166, pp. 422–435, 2015, doi: https://doi.org/10.1016/j.neucom.2015.03.037.

[26] R. S. Sangam and H. Om, "The k-modes Algorithm with Entropy Based Similarity Coefficient," *Procedia Comput. Sci.*, vol. 50, pp. 93–98, 2015, doi: https://doi.org/10.1016/j.procs.2015.04.066.

[27] J. ji, W. Pang, C. Zhou, X. Han, and Z. Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data," *Knowledge-Based Syst.*, vol. 30, pp. 129–135, Nov. 2013, doi: 10.1016/j.knosys.2012.01.006.

[28] A. Tyagi and S. Sharma, "Implementation Of ROCK Clustering Algorithm For The Optimization Of Query Searching Time," *Int. J. Comput. Sci. Eng.*, vol. 4, May 2012.

[29] Z. He, X. Xu, and S. Deng, "Squeezer: An efficient algorithm for clustering categorical data," *J. Comput. Sci. Technol.*, vol. 17, no. 5, pp. 611–624, 2002, doi: 10.1007/BF02948829.

[30] P. Andritsos, P. Tsaparas, R. Miller, and K. Sevcik, *LIMBO: Scalable clustering of categorical data*. 2004.

[31] D. Barbara, Y. Li, and J. Couto, *COOLCAT: An entropy-based algorithm for categorical clustering*. 2002.

[32] Y. Yang, X. Guan, and J. You, "CLOPE: A Fast and Effective Clustering Algorithm for Transactional Data," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Aug. 2002, doi: 10.1145/775047.775149.

[33] D. Wang, D. Tan, and L. Liu, "Particle swarm optimization algorithm: an overview," *Soft Comput.*, vol. 22, no. 2, pp. 387–408, 2018, doi: 10.1007/s00500-016-2474-6.

[34] H. Ghaedi, S. R. Kamel Tabbakh Farizani, and R. Gaemi, "A Novel Meta-heuristic Framework for Solving Power Theft Detection Problem: Cheetah Optimization Algorithm," *Int. J. Ind. Electron. Control Optim.*, vol. 5, no. 1, pp. 63–76, 2022, doi: 10.22111/ieco.2022.39528.1370.

[35] L. Kaufman and P. Rousseeuw, *Finding Groups in Data: An Introduction To Cluster Analysis*. 1990.

[36] "Sales_Transactions_Dataset_Weekly," 2017. https://archive.ics.uci.edu/ml/datasets/Sales_Transactions_Dataset_Weekly.

[37] "Facebook Live Sellers in Thailand," 2019. https://archive.ics.uci.edu/ml/datasets/Facebook+Live+Sellers+in+Thailand.

[38] A. Askarzadeh, "A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm," *Comput. Struct.*, vol. 169, pp. 1–12, 2016, doi: 10.1016/j.compstruc.2016.03.001.

[39] H. Ghaedi, S. R. K. Tabbakh, and R. Ghaemi, "Improving Electricity Theft Detection using Combination of Improved Crow Search Algorithm and Support Vector Machine," *Majlesi J. Electr. Eng.*, vol. 15, no. 4 SE-Articles, Dec. 2021, doi: https://doi.org/10.52547/mjee.15.4.63.