
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش مصنوعی

تحلیل روابط گروهی در جریان‌داده با استفاده از خوشه‌بندی

نگارنده: میلاد محمدی پارچینی

استاد راهنما

دکتر هدی مشایخی

استاد مشاور

دکتر منصور فاتح

بهمن ۱۳۹۷

شماره: ۱۴۶۳/ف/ک
تاریخ: ۹۷/۱۲/۵

باسمه‌تعالی

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد **آقای میلاد محمدی پارچینی** با شماره دانشجویی۹۵۱۳۳۸۴ رشته مهندسی کامپیوتر **گرایش هوش مصنوعی و رباتیکز** تحت عنوان **تحلیل روابط گروهی در جریان‌داده با استفاده از خوشه‌بندی** که در تاریخ ۱۳۹۷/۱۱/۸ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

قبول (با درجه: ..عالی با تشویق هیئت داوران.....) ☑ مردود ☐

نوع تحقیق: نظری ☒ عملی ☐

عضو هیأت داوران	نام ونام خانوادگی	مرتبهٔ علمی	امضاء
۱ـ استادراهنمای اول	هدی مشایخی	استادیار	
۲- استادراهنمای دوم	-	استادیار	
۳- استاد مشاور	منصور فاتح	استادیار	
۴- نماینده تحصیلات تکمیلی	اسماعیل طحانیان	استادیار	
۵- استاد ممتحن اول	فاطمه جعفری نژاد	استادیار	
۶ـ استاد ممتحن دوم	علیرضا تجری	استادیار	

نام و نام خانوادگی رئیس دانشکده:

تاریخ و امضاء و مهر دانشکده: ۹۷/۱۲/۱۰

تبصره: در صورتی که کسی مردود شود حداکثر یکبار دیگر (در مدت مجاز تحصیل) می تواند از پایان نامه خود دفاع نماید (دفاع مجدد نباید زودتر از ۴ ماه برگزار شود).

تقدیم اول

به پدر و مادر عزیزم که مهر آسمانی شان آرامش بخش

و راه و روش شان چراغ راه زندگانی ام است

تشکر و قدردانی

سپاس خداوند یکتای عزتمندی که رحمت و دانش او در سراسر گیتی گسترده شده، آسمان ها و زمین همه از آن اوست و علم و دانش حقیقی را بر هر که بخواهد موهبت می فرماید. رحمت و لطف او را بی نهایت سپاس می گویم چرا که فهم و درک مطالب این پژوهش را بر من ارزانی داشت و مرا به این اصل رساند که علم و ایمان دو بال یک پرواز ند. توفیق تلاش به من داد، تا با امید، راه تازه ای را آغاز کنم و به خواست او به نتیجه ی مطلوب نائل آیم. به راستی که همه چیز از آن و به خواست اوست.

همچنین از استاد گرامی، سرکار خانم دکتر مهدی مشایخی بسیار سپاسگزارم که در تمامی دشواری های این مسیر، راهنمایی های بی دریغشان چاره ساز کارم بود و از جناب آقای دکتر منصور فاتح بابت کمک هایشان قدردانی به عمل می آورم.

تعهدنامه

اینجانب **میلاد محمدی پارچینی** دانشجوی دوره کارشناسی ارشد رشته **مهندسی کامپیوتر** دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **" تحلیل روابط گروهی در جریان‌داده با استفاده از خوشه‌بندی"** تحت راهنمائی سرکار خانم **دکتر هدی مشایخی** متعهد می‌شوم:

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

چکیده

امروزه با توجه به افزایش روند تولید داده‌ها و ظهور مفهوم جریان‌داده، می‌توان اطلاعات مفیدی را به کمک روش‌های داده‌کاوی از داده‌ها استخراج کرد. این اطلاعات می‌توانند در زمینه‌های متنوعی مانند تحلیل شبکه‌های اجتماعی، گروه‌بندی اطلاعات پزشکی بیماران و تشخیص نفوذ در شبکه استفاده شوند. از میان راهکارهای استخراج اطلاعات، می‌توان به خوشه‌بندی که روشی مناسب برای سازمان‌دهی داده‌ها است، اشاره داشت. در این پایان‌نامه، روشی جدید بر پایه‌ی راهکار خوشه‌بندی تطبیقی معرفی شده است. خوشه‌بندی تطبیقی با در نظر گرفتن یک دانش زمینه‌ای، سعی در ایجاد خوشه‌هایی با قدرت تفسیرپذیری بالاتر را دارد، تا بتواند به کشف روابط گروهی احتمالی موجود در میان داده‌ها و به دست آوردن اطلاعات مفید بیشتری از خوشه‌ها بپردازد. به طور کلی اگر داده‌ها از قبل گروه‌بندی مشخصی داشته باشند و بخواهیم از این پیش‌زمینه در فرایند خوشه‌بندی بهره ببریم، به نحوی که رابطه‌ای بین خوشه‌ها و گروه‌های پیشین وجود داشته باشد، می‌توان از خوشه‌بندی تطبیقی استفاده کرد. خوشه‌بندی تطبیقی، با وجود رویکرد تازه‌ای که به مقوله‌ی خوشه‌بندی داشته است، قابل اجرا بر روی جریان‌داده‌ها نیست و به صورت متمرکز عمل می‌کند.

در روش پیشنهادی این پایان‌نامه، فرایند خوشه‌بندی تطبیقی به صورت برخط و افزایشی ارائه شده است تا بتوان مفهوم موجود در خوشه‌بندی تطبیقی را بر روی جریان‌داده‌ها به کار گرفت و به کشف روابط گروهی بر روی آن‌ها پرداخت. روش پیشنهادی، با بهره‌گیری از دانش‌زمینه‌ای نگاه تازه‌ای در مفهوم خوشه‌بندی ایجاد کرده است. راهکار ارائه شده در این تحقیق تاکنون جزء نخستین روش‌های برخط خوشه‌بندی تطبیقی به حساب می‌آید. از مزایای مهم روش پیشنهادی می‌توان به امکان یافتن روابط گروهی موجود در یک جریان‌داده، کاهش حافظه‌ی مورد نیاز برای ذخیره‌سازی داده‌ها با بهره‌گیری از خلاصه‌سازی نمونه‌های جریان‌داده، کاهش پیچیدگی زمانی و محاسباتی و

همچنین عدم وابستگی به تعداد خوشه‌ها به منظور ایجاد خوشه‌های نهایی اشاره کرد. نتایج به دست آمده از انجام آزمایش‌ها روی مجموعه‌داده‌های مصنوعی و واقعی، عملکرد مناسب روش پیشنهادی را در مقایسه با خوشه‌بندی تطبیقی پایه و خوشه‌بندی K-Means تایید می‌کند.

کلمات کلیدی: جریان‌داده، خوشه‌بندی، خوشه‌بندی تطبیقی، یادگیری افزایشی، دانش‌زمینه‌ای

لیست مقالات مستخرج از پایان نامه

۱- محمدی،م. مشایخی،ه. فاتح،م. (۱۳۹۷)، " تحلیل روابط گروهی در جریان داده با استفاده از خوشه‌بندی تطبیقی" کنفرانس مشترک سیستم‌های فازی و هوشمند ایران، سال هفتم، بجنورد، ایران.

فهرست مطالب

فهرست جداول **ل**

فهرست اشکال **م**

فهرست واژگان **ن**

فصل۱ : مقدمه **۱**

۱-۱ مقدمه ۲

۱-۲ شرح مسئله ۳

۱-۳ اهمیت پژوهش ۵

۱-۴ هدف پژوهش ۶

۱-۵ روش پیشنهادی ۷

۱-۶ روش ارزیابی ۸

۱-۷ مروری بر فصل‌ها ۸

فصل۲ : ادبیات پژوهش **۱۱**

۲-۱ جریان‌داده ۱۳

۲-۲ خوشه‌بندی ۱۴

۲-۳ خوشه‌بندی مقید ۱۸

۲-۴ خوشه‌بندی جریان‌داده ۱۹

۲-۵ بررسی پژوهش‌های انجام شده ۲۳

۲-۶ خوشه‌بندی K-Means ۲۵

۲-۷ خوشه‌بندی تطبیقی ۲۶

۲-۸ جمع‌بندی ۳۰

فصل۳ : معرفی روش پیشنهادی **۳۱**

۳-۱ چالش‌های موجود ۳۳

۳-۱-۱ خلاصه‌سازی داده‌ها (فاز اول) ۳۵

۳-۱-۲ خوشه‌بندی تطبیقی وزن‌دار (فاز دوم) ۴۳

۳-۲ جمع‌بندی ۴۹

فصل۴ : ارزیابی روش پیشنهادی و نتایج **۵۱**

۴-۱ تنظیمات و راه‌اندازی سیستم ۵۲

۴-۲ مجموعه داده ۵۲

۴-۲-۱ مجموعه داده‌ی ساختگی ۵۲

۴-۲-۲ مجموعه داده‌های واقعی ۵۳

۴-۳ پیاده‌سازی روش ارائه شده ۵۴

۴-۴ آزمایش‌ها و ارزیابی نتایج ۵۵

۴-۵ جمع‌بندی ۶۸

فصل۵ : جمع بندی و پژوهشهای آینده **۶۹**

۵-۱ جمع بندی ۷۰

۵-۲ پژوهشهای آینده ۷۱

مراجع **۷۳**

فهرست جداول

جدول ۲-۱. خلاصه‌ای از پژوهش‌های انجام شده ۲۹
جدول ۳-۱. نمونه‌ای از بردار گروه-تعداد ۳۷
جدول ۳-۲. پارامترهای موجود در الگوریتم SAKmeans ۳۹
جدول ۳-۳. پارامترهای موجود در الگوریتم WAKmeans ۴۴
جدول ۳-۴. بردار گروه-تعداد برای مثال ۳-۲ ..... ۴۷
جدول ۳-۵. بردار گروه-تعداد مربوط به تعدادی از داده‌های هسته ۴۸
جدول ۴-۱. مقادیر پارامترها در پیاده‌سازی ۵۵
جدول ۴-۲. مقادیر پارامتر k در آزمایش نخست ۵۶
جدول ۴-۳. مقادیر پارامتر r در آزمایش دوم ۶۲
جدول ۴-۴. مقایسه‌ی زمان اجرای دو روش خوشه‌بندی تطبیقی جریان‌داده‌ای و خوشه‌بندی تطبیقی پایه ۶۷
جدول ۴-۵. مشخصات سیستم مورد استفاده ۶۷

فهرست اشکال

شکل ۱-۱. خوشه‌ی مربوط به کاندیدای x ۷
شکل ۲-۱. نمایش مدل‌های مختلف پنجره‌گذاری [۱۲] ۲۱
شکل ۲-۲. خوشه‌بندی **(1 , 0.75)-accordant** [۲] ۲۷
شکل ۳-۱. فرایند کلی خوشه‌بندی تطبیقی جریان‌داده ۳۵
شکل ۴-۱. نمایش مجموعه داده‌ی ساختگی ۵۳
شکل ۴-۲. نمودار **MSE** مربوط به داده‌ی ساختگی ۵۸
شکل ۴-۳. نمودار **MSE** مربوط به مجموعه‌داده‌ی **Pendigits** ۵۹
شکل ۴-۴. نمودار **MSE** مربوط به مجموعه‌داده‌ی **magic04** ۶۰
شکل ۴-۵. نمودار **MSE** مربوط به مجموعه‌داده‌ی **shuttle** ۶۱
شکل ۴-۶. نمودار **MSE** مربوط به مجموعه‌داده‌ی ساختگی با تغییر پارامتر **r** ۶۳
شکل ۴-۷. نمودار **MSE** مربوط به مجموعه‌داده‌ی **Pendigits** با تغییر پارامتر **r** ۶۴
شکل ۴-۸. نمودار **MSE** مربوط به مجموعه داده‌ی **magic04** با تغییر پارامتر **r** ۶۵
شکل ۴-۹. نمودار **MSE** مربوط به دو روش جریان‌داده‌ای بر روی مجموعه‌داده‌ی ساختگی ۶۶
شکل ۴-۱۰. نمودار **MSE** مربوط به دو روش جریان‌داده‌ای بر روی مجموعه‌داده‌ی **Pendigits** ۶۶

فهرست واژگان

افزایشی	Incremental
اکتشافی	Heuristic
با ناظر (تحت نظارت)	Supervised
برخط	Online
بُعد	Dimention
بهم ریختن، مخلوط کردن	Shuffle
پنجره‌گذاری	Windowing
تجزیه و تحلیل برخط داده‌های حجیم	Massive Online Analysis (MOA)
تدریجی	Gradual
تغییر مفهوم	Concept Drift
تکامل مفهوم	Concept Evolution
تکامل ویژگی	Feature Evolution
تکرار شونده	Recurring
تکه‌ها	Chunks
جزئی	Partitional
خوشه‌بندی تطبیقی	Accordant Clustering
خوشه‌بندی کلاسیک سخت	Classical Hard Clustering
خوشه‌بندی نرم	Soft Clustering
خوشه‌پذیر	Clusterable
دانش زمینه‌ای	Background knowledge
داده‌های هسته	Coresets
سلسله مراتبی	Hierarchical
طبقه‌بندی شده	Categorical
عناوین	Topics
فراموشی	Forgetness
متن باز	Open Source
مقید	Constrained
میانگین مربعات خطا	Mean Squared Error (MSE)
ناگهانی	Sudden
نمونه‌های فردی	Individual Elements

فصل۱ : مقدمه

۱-۱ مقدمه

جریان‌داده از حوزه‌هایی است که در سال‌های اخیر اهمیت کاربردی قابل توجهی پیدا کرده است. یک جریان‌داده، توالی نامحدودی از داده‌ها است. این داده‌ها از منابع مختلفی مانند شبکه‌های اجتماعی، تراکنش‌های بانکی، شبکه‌های حسگر بی‌سیم و غیره تولید می‌شوند. می‌توان با تحلیل این‌گونه داده‌ها به کمک روش‌های داده‌کاوی مانند خوشه‌بندی، رده‌بندی و غیره، اطلاعات مفیدی را از جریان‌داده‌ها به دست آورد. مواردی نظیر برخط بودن[۱]، اندازه‌ی نامحدود داده‌ها، پیچیدگی در سازمان‌دهی، ماهیت پویا و افزایشی[۲] و عدم دسترسی به کل داده‌ها، از چالش‌های موجود در این حوزه است [۱]. امروزه به منظور ایجاد یک ساز و کار مناسب برای تحلیل جریان‌داده، روش‌های مختلفی معرفی شده که در این پژوهش از خوشه‌بندی استفاده شده است.

خوشه‌بندی به عنوان یکی از اساسی‌ترین ابزارهای داده‌کاوی در حوزه‌های مختلفی کاربرد دارد. خوشه‌بندی به طور گسترده با هدف پیدا کردن بخش‌های مشابه موجود در داده‌ها استفاده می‌شود. به طور معمول فرایند یادگیری در خوشه‌بندی بدون ناظر صورت می‌گیرد. یعنی عمدتا داده‌هایی که در اختیار روش‌های خوشه‌بندی قرار می‌گیرند، بدون برچسب هستند و با استفاده از خوشه‌بندی می‌توانیم اطلاعات و جزئیات بیشتری در مورد این داده‌ها به دست بیاوریم [۲].

روش‌های کلاسیک خوشه‌بندی موجود، بر روی افتراق بین نمونه‌های فردی[۳] (بدون در نظر گرفتن گروه (کلاس) نمونه‌ها) تمرکز دارند. به بیان دیگر این روش‌ها بخش‌هایی (خوشه‌هایی) با شباهت درون بخشی زیاد و بین بخشی کم ایجاد می‌کنند [۲]. برای کشف روابط معنی‌دار در داخل و میان گروه‌های موجود در یک مجموعه‌داده، به مفهوم تازه‌ای از خوشه‌بندی نیاز است. این مفهوم با عنوان

---

[۱] Online

[۲] Incremental

[۳] Individual Elements

خوشه‌بندی تطبیقی[۱] در [۲] مطرح شده است. هدف این روش، ایجاد یک توازن میان کشف ساختار ذاتی و پیدا کردن روابط گروهی در داده‌هایی است که به یک گروه مشابه تعلق دارند. این روش به صورت متمرکز عمل کرده و مناسب محیط جریان‌داده‌ای نیست.

۱-۲ شرح مسئله

هدف روش‌های خوشه‌بندی سنتی، دسته‌بندی انواع داده‌ها به خوشه‌های معنی‌دار است. برای این منظور طیف گسترده‌ای از توابع هدف و الگوریتم‌ها معرفی و به کار گرفته شده‌اند. اساسی‌ترین نمونه-های خوشه‌بندی، یا مبتنی بر جزءبندی[۲] هستند (جایی که خروجی آن‌ها مجموعه‌ای از k خوشه است) و یا سلسله مراتبی[۳] (که به طور همزمان نشان‌دهنده‌ی دسته‌بندی چندگانه در یک ساختار درختی است). در روش‌های خوشه‌بندی سنتی به عناصری که دارای برچسب یا کلاس مشابهی هستند (به اصطلاح، عناصر گروهی) برای ایجاد خوشه‌های نهایی توجهی نمی‌شود. این در حالی است که ما به دنبال استفاده از این عناصر گروهی در شکل‌گیری خوشه‌های نهایی و کشف روابط گروهی در خوشه‌ها هستیم [۲].

نوع ورودی‌ای که تکنیک‌های مختلف خوشه‌بندی به منظور تاثیر دادن عناصر گروهی در نتیجه‌ی خوشه‌بندی می‌پذیرند بسیار گسترده است. یکی از این انواع مختلف به کاربر اجازه می‌دهد تا یک وزن را برای نمونه‌ها مشخص کند [۳]. این وزن، اهمیت عناصر فردی را مشخص می‌کند تا خوشه‌بندی بر اساس مواردی که اهمیت بیشتری دارند، هدایت شود. در حالی که این روش اجازه‌ی انعطاف پذیری بیشتری را فراهم می‌آورد، اما وزنی که به یک نمونه‌ی مشخص اختصاص داده می‌شود، هیچ شرطی را به گروه‌های نمونه‌هایی که در یک خوشه‌ی مشابه قرار گرفته‌اند تحمیل نمی‌کند. شاید متداول ترین

[۱] Accordant Clustering
[۲] Partitional
[۳] Hierarchical

تنظیم نیمه نظارتی[۱] موجود این اجازه را می‌دهد که برخی از جفت نمونه‌ها به عنوان Must Link (ML) و برخی دیگر به عنوان Cannot Link (CL) باشند [۴]. اگر چنین محدودیت‌هایی امکان پذیر باشد، خوشه‌بندی نهایی احتمالا ارزش معنایی بالاتری خواهد داشت که این امر مفید خواهد بود. از روش‌های دیگر موجود برای خوشه‌بندی، خوشه‌بندی نرم است[۲] که در آن ممکن است داده‌ها به خوشه‌های چندگانه متعلق باشند که در برابر مدل خوشه‌بندی کلاسیک سخت[۳] که در آن هر داده بخشی از یک خوشه‌ی منحصر به فرد است، قرار دارد.

با توجه به موارد بیان شده، اگر یک دانش زمینه‌ای[۴] (دانش اولیه) نیز در اختیار روش‌های خوشه‌بندی قرار بگیرد، می‌توان انتظار شکل‌گیری خوشه‌های معنادارتری را داشت. یعنی داده‌هایی که در یک خوشه قرار گرفته‌اند علاوه بر مشابه بودن، ویژگی‌های دیگری نیز دارا باشند. منظور از دانش زمینه‌ای مورد اشاره در این پایان‌نامه، همان برچسب (کلاس یا گروه) مربوط به داده‌ها است. به عبارت دیگر یک گروه، شامل داده‌هایی با برچسب یکسان است. این دانش را می‌توان برای کشف روابط معنادار در داخل و میان گروه‌های موجود در یک مجموعه‌داده به کار گرفت. برای نمونه اگر در یکی از خوشه‌های شکل گرفته، تعداد مناسبی از یک گروه حاضر در مجموعه‌داده قرار گرفته باشد، این خوشه را می‌توان با تمرکز و دقت بیشتری مورد بررسی قرار داد تا رابطه‌ی احتمالی میان داده‌های آن گروه با خوشه‌ی شکل گرفته و یا داده‌های گروه‌های دیگر مشخص شود.

باید به این نکته توجه کرد که برخلاف روش‌های خوشه‌بندی نظارتی که هدف اصلی آن‌ها ایجاد خوشه‌هایی با برچسب‌های همگن است، خوشه‌بندی مد نظر در این پایان‌نامه چنین هدفی را دنبال نمی‌کند. زیرا این امکان وجود دارد که در خوشه‌های شکل گرفته، برچسب‌های مختلفی وجود داشته باشد (بستگی به هدف این خوشه‌بندی دارد).

---

[۱] Semi Supervised

[۲] Soft Clustering

[۳] Classical Hard Clustering

[۴] Background knowledge

۱-۳ اهمیت پژوهش

روش‌های خوشه‌بندی با خود یک ابهام به همراه دارند. به بیان دیگر یک مجموعه‌داده‌ی مشابه، اغلب می‌تواند با چندین روش خوشه‌بندی گردد. علت این امر آن است که روش‌های خوشه‌بندی، عمدتا با معیارهای مختلفی، فرایند خوشه‌بندی یک مجموعه‌داده را انجام می‌دهند. این معیارها در داده‌های مختلف، نتایج متفاوتی ایجاد می‌کنند. به این ترتیب بهینگی و سودمندی خوشه‌بندی داده‌ها به نوع کاربرد وابستگی پیدا می‌کند. آنچه که به دنبال آن هستیم، کشف و تحلیل روابط گروهی در جریان داده (از طریق آنالیز و بررسی اعضای گروه) به کمک روشی برای خوشه‌بندی است که علاوه بر توانایی کشف ساختار معنی‌دار خوشه‌ها، بتواند قسمت‌ها و بخش‌های مفیدی را برای یک کاربرد خاص پیدا کند.

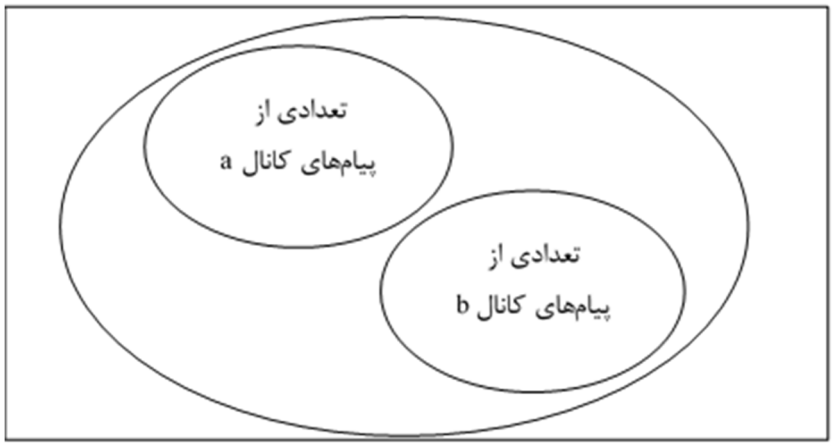
در این روش، داده‌های درون مجموعه‌داده بر اساس یک سری از ویژگی‌ها (بعد از انجام پیش‌پردازش‌های مورد نیاز بر روی داده‌ها) در یک دسته قرار می‌گیرند که هرکدام از این دسته‌ها، گروه نامیده می‌شوند. برای مثال، ممکن است هدف متخصصان پزشکی در یک زمینه‌ی خاص به دست آوردن عملکرد روش‌های مختلف برای درمان از طریق تجزیه و تحلیل چندین گروه درمانی باشد. در این مورد، هدف نه تنها کشف این موضوع است که کدام درمان موثرتر است، بلکه متخصصان به دنبال یافتن بهترین جمعیتی هستند که مناسب‌ترین شکل یا بیشترین تعداد از یک گروه را نیز داشته باشد. به این ترتیب، این امکان وجود دارد که ما به دنبال خوشه‌بندی بیماران در تمام گروه‌های درمانی بر اساس هر دو ویژگی جمعیت و نتایج حاصل از درمان (موثرترین روش برای درمان) باشیم. در حالی که راه‌های مختلفی برای تقسیم کردن معنی‌دار داده‌های مربوط به بیماران می‌تواند وجود داشته باشد، اما همه‌ی این روش‌ها به خوبی نمی‌توانند روش‌های درمانی را از یکدیگر متمایز کرده و ارتباط بین گروه‌های درمانی را آشکار سازند [۲].

۱-۴ هدف پژوهش

روش خوشه‌بندی تطبیقی [۲] که درسال ۲۰۱۷ مورد بررسی و پژوهش قرار گرفته، یک روش متمرکز بوده است. به عبارت دیگر، این روش خوشه‌بندی که توانایی یافتن روابط گروهی در یک مجموعه‌داده را داراست، قابلیت اجرا شدن بر روی جریانی از داده‌ها را ندارد. در این پایان‌نامه، مطابق با پژوهش‌های انجام شده، می‌خواهیم روشی بر پایه‌ی روش متمرکز قبلی و البته به صورت برخط، روی جریانی از داده‌ها پیاده‌سازی نماییم تا بتوانیم به تحلیل روابط گروهی مربوط به جریان‌داده‌ها تا حد مناسبی دست پیدا کنیم. همانگونه که گفته شد، هدف اصلی این پژوهش معرفی و استفاده از یک روش خوشه‌بندی تطبیقی جدید به منظور تحلیل روابط گروهی در جریان‌داده‌ها است. به این معنی که این روش با توجه به گروه‌های موجود در یک جریان‌داده و کمک گرفتن از یک دانش اولیه ، روابط بین گروه‌ها را کشف و تحلیل نماید. در ادامه یک نمونه از جریان‌داده‌هایی که در آن روابط گروهی وجود دارد برای روشن‌تر شدن مفهوم روابط گروهی در یک جریان‌داده قرار گرفته است.

شبکه‌ی اجتماعی تلگرام را در نظر بگیرید که در آن کانال‌هایی با تعداد اعضای زیادی وجود دارد. در یک زمان خاص، مثلا زمانی که به انتخابات ریاست جمهوری نزدیک شده‌ایم، در هر یک از کانال‌های خبری معتبر پیام‌هایی در این زمینه منتشر می‌شود. این مثال می‌تواند نمایانگر جریان‌داده‌ای باشد که در آن روابط گروهی وجود دارد. در واقع منظور ما از گروه در این حالت همان نام کانال خبری و یا شناسه‌ی منحصر به فرد آن کانال است. از این روابط گروهی می‌توان در خوشه‌بندی این جریان‌داده به عنوان یک ویژگی کمکی (همان دانش اولیه) استفاده کرد. مثلا همه‌ی پست‌های ایجاد شده توسط این کانال‌ها را در یک بازه‌ی زمانی مشخص روی هم ریخته و عمل خوشه‌بندی (یک روش خوشه‌بندی که در آن روابط گروهی مد نظر باشد (مثلا خوشه‌بندی تطبیقی)) را به فرض براساس نام کاندیداها به انجام برسانیم. در این حالت از خوشه‌های ایجاد شده می‌توان مفاهیم تازه و مفیدی را استخراج کرد. به عنوان یک نمونه‌ی شهودی فرض کنیم که یکی از

خوشه‌های ایجاد شده مانند شکل ۱-۱ بوده و به میزان ۰,۷۵ درصد از هر دو کانال a و b در این خوشه قرار گرفته باشد. در این حالت می‌توان به این نتیجه رسید که به طور فرض این دو کانال دارای گرایش سیاسی مشابهی هستند.



شکل ۱-۱. خوشه‌ی مربوط به کاندیدای x

در این پژوهش ما به بررسی چالش‌های مربوط به مدل خوشه‌بندی ارائه شده پرداخته و نتایج حاصل از غلبه‌ی روش خود بر این چالش‌ها را بیان خواهیم کرد. از جمله مزیت‌های مهمی که در معرفی و پیاده‌سازی روش خوشه‌بندی جدید به دنبال آن هستیم (علاوه بر برخط بودن و قابل اجرا بودن این روش بر روی جریان‌داده‌ها) کمتر بودن حافظه‌ی مصرفی در الگوریتم ارائه شده به دلیل استفاده از خلاصه کردن داده‌ها با هدف سازماندهی همه‌ی داده‌ها در تعداد کمی از نمونه‌ها، بالا بردن سرعت اجرای الگوریتم و همچنین کاهش دادن پیچیدگی زمانی و محاسباتی است.

۱-۵ روش پیشنهادی

در این پایان‌نامه، یک روش خوشه‌بندی تطبیقی برخط با هدف انجام فرایند خوشه‌بندی تطبیقی روی جریان‌داده‌ها ارائه شده است. به این منظور با معرفی چند ویژگی جدید برای خوشه‌بندی تطبیقی پایه و ایجاد تغییرات کلی در ساختار موجود در الگوریتم آن (متناسب با ویژگی‌های جدید تعریف شده در روش پیشنهادی) برای برخط عمل کردن روش خوشه‌بندی تطبیقی پایه، دو الگوریتم جدید، یکی

برای خلاصه کردن داده‌های موجود در جریان‌داده و دیگری برای انجام فرایند خوشه‌بندی تطبیقی وزن‌دار بر روی خلاصه‌های به دست آمده پیشنهاد شده است. الگوریتم‌های پیشنهادی، به زبان جاوا و در محیط برنامه‌نویسی IntelliJ IDEA پیاده‌سازی شده‌اند.

۱-۶ روش ارزیابی

برای ارزیابی نتایج حاصل از روش پیشنهادی در این پژوهش از معیار میانگین مجموع مربعات خطا[۱] استفاده شده است. از آنجایی که روش پیشنهادی به نوعی نسخه‌ی برخط شده‌ی خوشه‌بندی تطبیقی پایه است، برای بررسی چگونگی عملکرد این روش، معیار مذکور بر روی مجموعه‌داده‌های ساختگی و واقعی برای روش پیشنهادی و دو روش خوشه‌بندی تطبیقی پایه (AKmeans) و روش K-means به اجرا درآمده که در فصل چهارم به طور کامل نتایج حاصل از آن مورد بحث و بررسی قرار گرفته است.

۱-۷ مروری بر فصل‌ها

پس از شرح مقدماتی که در این فصل آورده شد، در فصل دوم ابتدا مفاهیم نظری و عملیاتی و به طور کلی مفاهیم مرتبط با موضوع پژوهش و سپس وارسی برخی از روش‌های مرتبطی که تاکنون در این زمینه صورت گرفته (خوشه‌بندی جریان‌داده و خوشه‌بندی با در نظر گرفتن دانش زمینه)، ارائه می‌شود. در فصل سوم، راهکار پیشنهادی این پژوهش به همراه جزئیاتی که (برای مثال، پارامترهای مربوط به روش پیشنهادی) لازم است تا تمامی زوایای این مبحث روشن گردد، بیان خواهد شد. در فصل چهارم به بررسی نتایج و تحلیل نمودارهای خروجی حاصل از این پژوهش پرداخته شده و ارزیابی‌های صورت گرفته گزارش خواهد شد. در نهایت در فصل پنجم یک جمع‌بندی و نتیجه‌گیری از

[۱] Mean squared error (MSE)

کل پژوهش به عمل آمده و پیشنهاداتی با هدف بسط و توسعه این پژوهش در اختیار خواننده قرار خواهد گرفت.

فصل ۲ : ادبیات پژوهش

در فرایند داده‌کاوی به استخراج الگوها و اطلاعات پنهان از داده‌های موجود پرداخته می‌شود. امروزه یکی از نکات مهمی که در استفاده از علم داده‌کاوی مورد توجه قرار می‌گیرد، ویژگی ایستا بودن و پویایی داده‌های مورد بررسی است. مشخصا کنترل داده‌های ایستا ساده‌تر و دارای پیچیدگی پردازشی کمتری نسبت به داده‌های از نوع پویا است. مجموعه‌داده‌های ایستا، قبل از پردازش در دسترس قرار دارند و به طور کلی بر خلاف داده‌های پویا با گذشت زمان تغییر نمی‌کنند. موضوعی که مطرح می‌شود نحوه‌ی استفاده از داده‌کاوی در تحلیل و استخراج الگوها است. به این منظور، استفاده از الگوریتم‌ها و روش‌های موجود در داده‌کاوی با توجه به نوع داده‌ها و کاربرد مد نظر کاربر، اهمیت پیدا می‌کند. روش‌هایی که به کار گرفته می‌شوند، باید متناسب با نوع داده‌ها بوده و اهداف مورد نظر را برآورده سازند. چالش اصلی، با توجه به موارد بیان شده، انتخاب یک روش مناسب برای برخورد با داده‌ها در جهت تحلیل آن‌ها است. یکی از مهم‌ترین دلایلی که این امر به عنوان یک چالش مطرح می‌شود، موضوع بهینگی روش انتخابی در جهت حل مسئله‌ی مد نظر است [۵].

الگوریتم‌های سنتی داده‌کاوی برای کنترل جریان‌های داده مناسب نیستند، چرا که این الگوریتم‌ها چندین گذر بر روی داده‌ها انجام می‌دهند. این مسئله یک چالش واقعی بر سر راه محققانی قرار می‌دهد که در حوزه‌ی جریان‌های‌داده فعال هستند. علاوه بر الگوریتم‌های موجود داده‌کاوی که برای خوشه‌بندی، رده‌بندی و یافتن یک الگو وجود دارند، عمدتا برای مجموعه‌داده‌های ثابت مناسب هستند و از لحاظ عملی برای کنترل و یا استخراج جریان‌های‌داده به کار گرفته نمی‌شوند. امروزه الگوریتم‌های داده‌کاوی با به کارگیری روش‌های نوینی مانند سازوکار فراموشی داده‌ها[۱]، خلاصه‌سازی داده‌ها و استفاده از پنجره‌گذاری[۲] این چالش را مدیریت کرده‌اند [۵]. در ادامه‌ی این فصل به بیان و بررسی چند مفهوم اساسی مورد نیاز برای انجام پژوهش حاضر، پرداخته‌ایم.

---

[۱] Forgetness

[۲] Windowing

۲-۱ جریان‌داده

در سال‌های اخیر، حوزه‌ی جریان‌داده اهمیت قابل توجهی پیدا کرده است. یک جریان‌داده، توالی نامحدودی از داده‌ها است. نمونه‌هایی از جریان‌داده شامل ترافیک شبکه، داده‌های سنسور، شبکه‌های اجتماعی و غیره است. از جمله مهم‌ترین چالش‌های موجود در جریان‌داده می‌توان به برخط بودن، اندازه‌ی نامحدود داده‌ها، پیچیدگی در سازمان‌دهی آن‌ها، ماهیت پویا و افزایشی و عدم دسترسی به کل داده‌ها اشاره داشت. این چالش‌ها فرایندهای داده‌کاوی را در استخراج و تحلیل اطلاعات با مشکل مواجه کرده است. در یک جریان‌داده، تعداد داده‌های برچسب‌دار نسبت به داده‌های بدون برچسب بسیار کمتر است. دلیل این امر آن است که برچسب‌گذاری داده‌ها، یا باید به صورت دستی انجام گیرد و یا آنکه توسط یک ناظر برچسب‌گذاری انجام شود. از جمله ویژگی‌های جریان‌داده که آن‌ها را متمایز می‌سازد، طول بی‌نهایت، تغییر مفهوم[۱]، تکامل مفهوم[۲]، تکامل ویژگی[۳] و تعداد محدود داده‌های برچسب‌دار است [۱]. در ادامه‌ی این بخش به شرح تغییر مفهوم پرداخته‌ایم.

تغییر مفهوم زمانی رخ می‌دهد که مفهوم اصلی موجود در داده‌ها در طول زمان تغییر یابد. اگر این تغییر مفهوم با پدیدار شدن کلاس‌های جدید همراه شود، تکامل مفهوم اتفاق افتاده است. تغییر رفتار داده‌ها در طول زمان موجب شناسایی و پدیدار شدن تغییر مفهوم می‌شود. با شناسایی این تغییرات، می‌توان تصمیمات بهتری در حوزه‌هایی مانند رشد جمعیت، تغییرات آب و هوایی، بازار سهام و غیره اتخاذ نمود. تغییر مفهوم از نظر چگونگی رخداد به چهار دسته‌ی: ناگهانی[۴]، تدریجی[۵]، افزایشی و تکرارشونده[۶] تقسیم می‌شود [۱] که شرح مربوط به آن‌ها در ادامه قرار گرفته است.

---

[۱] Concept Drift
[۲] Concept Evolution
[۳] Feature Evolution
[۴] Sudden
[۵] Gradual
[۶] Recurring

- **تغییر مفهوم ناگهانی:** در این نوع تغییر، مفهوم به صورت ناگهانی تغییر پیدا می‌کند. به عبارت دیگر، معنا قبل و بعد از یک زمان مشخص با یکدیگر متفاوت است. برای مثال، در یک بازه‌ی زمانی معین، میان گروه‌های شبکه‌های اجتماعی، بحث‌هایی با محوریت سیاست انجام می‌گیرد و در بازه‌ی زمانی دیگری بحث‌های فرهنگی صورت می‌گیرد. در چنین حالتی بیان می‌شود که تغییر ناگهانی مفهوم رخ داده است.
- **تغییر مفهوم تدریجی:** تغییر مفهوم در این نوع، به صورت تدریجی از حالتی به حالت دیگر اتفاق می‌افتد. برای مثال، در یک شبکه‌ی اجتماعی به مرور زمان و نه به طور ناگهانی بحث‌های صورت گرفته دچار تغییر در محتوا می‌شوند.
- **تغییر مفهوم افزایشی:** در این دسته، تغییر مفهوم مشابه تغییر مفهوم تدریجی است. با این تفاوت که نمونه‌های موجود، در یک بازه‌ی زمانی از حالتی به حالت دیگر تغییر پیدا می‌کنند. به این صورت که توزیع مفاهیم به تدریج از یک حالت فاصله گرفته و به حالت دیگری تغییر پیدا می‌کند.
- **مفاهیم تکرار شونده:** در این نوع، مفهومی که از قبل وجود داشته است، می‌تواند مجددا تکرار شود. برای مثال در یک شبکه‌ی اجتماعی به یک مبحث مشخص پرداخته می‌شود و بعد از مدتی دوباره آن مبحث میان کاربران مطرح شده و به گفت و گو گذاشته می‌شود.

۲-۲ خوشه‌بندی

خوشه‌بندی فرایندی است که در آن یک مجموعه از داده‌ها به مجموعه‌ای از خوشه‌ها طبقه‌بندی می‌شود، به نحوی که داده‌های درون یک خوشه به یکدیگر شبیه بوده و تا جای ممکن با داده‌های خوشه‌های دیگر متمایز باشند. از کاربردهای خوشه‌بندی در علوم مختلف می‌توان به مهندسی، پزشکی، علوم اجتماعی، بازاریابی و غیره اشاره کرد. به طور معمول فرایند یادگیری در خوشه‌بندی،

بدون نظارت صورت می‌گیرد که در طی آن، نمونه‌ها به دسته‌هایی که اعضای آن مشابه یکدیگر هستند تقسیم می‌شوند که از این دسته‌ها به عنوان خوشه یاد می‌شود. از جمله معیارهای شباهت مورد استفاده در خوشه‌بندی می‌توان به معیار فاصله اشاره داشت که در آن اعضایی که به یکدیگر نزدیک‌تر هستند به عنوان یک خوشه در نظر گرفته می‌شوند. به این نوع خوشه‌بندی، خوشه‌بندی مبتنی بر فاصله نیز اطلاق می‌شود [۶].

فرایند خوشه‌بندی با رده‌بندی تفاوت دارد. به این دلیل که در رده‌بندی برچسب نمونه‌های ورودی از قبل مشخص است در حالی که در خوشه‌بندی نمونه‌های ورودی برچسب اولیه ندارند. در واقع یکی از کاربردهای اصلی خوشه‌بندی، اختصاص برچسب به داده‌هایی است که در مورد آن‌ها اطلاعات و آگاهی کافی وجود ندارد.

از جمله مشخصات یک الگوریتم خوشه‌بندی مناسب، می‌توان به چند مورد زیر اشاره کرد [۷]:

۱. **قابلیت مقیاس‌پذیری:** یافتن خوشه در داده‌هایی با ابعاد بالا به عنوان چالشی در الگوریتم‌های خوشه‌بندی مطرح است. بنابراین باید تکنیک‌هایی که در داده‌کاوی به کار برده می‌شوند توانایی مدیریت حجم و ابعاد بالای داده‌ها را داشته باشند. از طرفی اگر تعداد نمونه-هایی که در اختیار روش خوشه‌بندی قرار می‌گیرد ناکافی باشد، خوشه‌های مناسبی شکل نگرفته و نتایجی که به دست می‌آیند جهت‌گیرانه خواهند بود.

۲. **توانایی مواجه با انواع داده‌ها:** الگوریتم‌های خوشه‌بندی باید توانایی کار با انواع داده‌ها، مانند داده‌های عددی و اسمی (طبقه‌بندی شده)[۱] را داشته باشند.

۳. **توانایی مقابله با داده‌های نویزی، نادرست و ناقص:** با وجود آن‌که پیش‌پردازش‌هایی قبل از اعمال داده‌ها به فرایند خوشه‌بندی انجام می‌گیرد، با این حال نمی‌توان به صورت قطعی در مورد بدون نقص بودن فرایند پیش‌پردازش اطمینان داشت. از این‌رو الگوریتم خوشه‌بندی

[۱] Categorical

نباید نسبت به نقصان موجود در داده‌ها حساسیت داشته باشد.

۴. **عدم حساسیت به ترتیب ورود داده‌ها:** نتایج الگوریتم‌های خوشه‌بندی نباید با تغییر ترتیب داده‌های ورودی به میزان قابل توجهی دچار تغییر شوند. از طرفی این الگوریتم‌ها باید قابلیت افزودن داده‌های جدید را نیز داشته باشند.

انواع روش‌های خوشه‌بندی

تاکنون روش‌های خوشه‌بندی متعددی معرفی شده است که در این بخش به بررسی چند مورد از روش‌های مطرح موجود در خوشه‌بندی پرداخته و به طور خلاصه اشاره‌ای به آن‌ها داشته‌ایم.

۱. **خوشه‌بندی مبتنی بر توزیع**: در این روش خوشه‌بندی نیازی به مشخص کردن تعداد خوشه‌های نهایی به عنوان پارامتر ورودی الگوریتم خوشه‌بندی نیست. این روش جزء روش‌های خوشه‌بندی سریع به شمار می‌آید. در این نوع خوشه‌بندی یک فرایند تکراری بر روی داده‌های ورودی انجام شده و هر داده‌ی ورودی به صورت متوالی خوانده می‌شود. سپس شباهت هر داده با خوشه‌های موجود بررسی شده و در صورتی که شبیه‌ترین خوشه از یک مقدار آستانه (که از قبل در نظر گرفته شده) بیشتر باشد، داده به آن خوشه تعلق می‌گیرد. در غیر این صورت، داده‌ی ورودی به عنوان یک خوشه‌ی جدید تلقی می‌شود. از جمله پارامترهای ورودی خوشه‌بندی مبتنی بر توزیع می‌توان به مقیاس شباهت، حداکثر تعداد خوشه‌ها و ماتریس شباهت اشاره داشت. از حداکثر تعداد خوشه‌ها در راستای جلوگیری از تولید خوشه‌های بسیار کوچک به تعداد زیاد و هم‌چنین صرفه‌جویی در زمان اجرای الگوریتم استفاده می‌شود. به عنوان یک نمونه از الگوریتم‌های موجود در این دسته می‌توان الگوریتم GMM را نام برد [۸].

۲. **خوشه‌بندی مبتنی بر مدل:** در این خوشه‌بندی به صورت پیش‌فرض یک مدل برای هر خوشه در نظر گرفته می‌شود. در ادامه‌ی فرایند خوشه‌بندی مبتنی بر مدل، نمونه‌های ورودی

به مدل مربوط به خود اضافه می‌شوند. در این نوع تکنیک، روش‌های آماری و شبکه‌ی عصبی مورد استفاده قرار می‌گیرند. از نمونه الگوریتم‌هایی که دراین نوع خوشه‌بندی وجود دارد، می‌توان به الگوریتم SOM اشاره کرد [۸].

۳. **خوشه‌بندی سلسله مراتبی:** از این روش خوشه‌بندی با رویکرد دسته‌بندی داده‌ها استفاده می‌شود. در این روش، داده‌های ورودی بر اساس معیار شباهت در دسته‌ها و زیردسته‌هایی قرار داده می‌شوند. خوشه‌بندی سلسله مراتبی خود به دو دسته‌ی بالا به پایین و پایین به بالا تقسیم می‌شود [۶].

الف) بالا به پایین یا تقسیم کننده : در این روش ابتدا تمامی داده‌ها به عنوان یک خوشه در نظر گرفته می‌شوند و سپس در طی یک فرایند تکراری در هر مرحله داده‌هایی که شباهت کمتری به یکدیگر دارند به خوشه‌های مجزایی شکسته می‌شوند و این روال تا رسیدن به خوشه‌هایی که دارای یک عضو هستند، ادامه پیدا می‌کند.

ب) پایین به بالا یا متراکم کننده: در این روش ابتدا هر داده به عنوان خوشه‌ی مجزا در نظر گرفته می‌شود و در طی فرایندی تکراری در هر مرحله خوشه‌هایی که شباهت بیشتری به یکدیگر دارند ترکیب می‌شوند تا در نهایت یک خوشه و یا تعداد مشخصی خوشه ایجاد شود.

۴. **خوشه‌بندی مبتنی بر جزءبندی:** در این روش خوشه‌بندی، ابتدا داده‌ها به مجموعه‌هایی واحد تقسیم‌بندی می‌شوند (در مجموعه‌های واحدی قرار می‌گیرند) به گونه‌ای که هر داده دقیقا در یک مجموعه قرار بگیرد. سپس با در نظر گرفته شدن یک تقسیم‌بندی اولیه، هر نمونه به صورت تکراری از یک خوشه به خوشه‌ی دیگر انتقال می‌یابد. این فرایند تا جایی ادامه پیدا می‌کند که به یک حالت بهینه در تقسیم‌بندی برسد. از معایب این روش می‌توان به مشخص بودن تعداد خوشه‌ها از ابتدا (به عنوان پارامتر ورودی) اشاره کرد. برای نمونه، روش خوشه‌بندی K-Means در این دسته قرار می‌گیرد که در همین فصل به بررسی و شرح

مربوط به آن پرداخته می‌شود [۸].

۵. **خوشه‌بندی مبتنی بر چگالی:** در این روش خوشه‌ها به عنوان بخش‌هایی از داده‌ها در نظر گرفته می‌شوند که چگالی بالایی دارند و به وسیله‌ی بخش‌های دارای چگالی پایین از یکدیگر متمایز شده‌اند. به عبارت دیگر، مجموعه‌داده‌ها به زیرمجموعه‌هایی تقسیم می‌شوند که چگالی و توزیع داده‌ها در آن‌ها مد نظر قرار می‌گیرد. از دلایل استفاده از این روش می‌توان به ناکارآمدی روش‌های خوشه‌بندی مبتنی بر افراز و سلسله‌مراتبی به منظور یافتن خوشه‌های بیضی شکل و S مانند، اشاره کرد. الگوریتم خوشه‌بندی Mean-Shift را می‌توان به عنوان یک نمونه از الگوریتم‌هایی که در این دسته از خوشه‌بندی قرار می‌گیرند، معرفی نمود [۸].

۲-۳ خوشه‌بندی مقید

در این نوع خوشه‌بندی، برخلاف روش‌های خوشه‌بندی مورد بحث تاکنون، به نوعی یک فرایند با ناظر به اجرا در می‌آید. ایده‌ی استفاده از اطلاعات جانبی (دانش زمینه‌ای) [۹] با هدف کمک کردن به الگوریتم‌های خوشه‌بندی در تشخیص و شکل‌گیری خوشه‌ها، دیدگاه و رویکردی است که روش‌های خوشه‌بندی مقید[۱] در نظر گرفته‌اند [۱۰]. در این دیدگاه می‌توان از یک دانش زمینه‌ای برای انجام فرایند خوشه‌بندی استفاده کرد. به‌کارگیری قیدها قابلیت به وجود آوردن خوشه‌هایی با شکل‌های دلخواه را به روش‌های خوشه‌بندی می‌دهد. در بسیاری از کاربردها می‌توان از دانش زمینه‌ای و یا دانش موجود در داده‌ها، قیدهایی تعریف کرد و از این قیدها در راستای بهبود دقت خوشه‌بندی استفاده نمود. از نمونه کاربردهای خوشه‌بندی مقید می‌توان مواردی مانند پردازش داده‌های وب، پردازش داده‌های شبکه، پردازش متون و غیره را نام برد [۱۱]. از روش‌های خوشه‌بندی مقید موجود می‌توان به مواردی مانند روش‌های مبتنی بر ارضای قید، روش‌های سلسله مراتبی، روش‌های مبتنی

[۱] Constrained

بر تغییر ماتریس فاصله و روش‌های مبتنی بر یادگیری معیار فاصله اشاره داشت.

تمامی موارد بیان شده از الگوریتم‌های خوشه‌بندی در این قسمت، قابلیت اجرا بر روی جریان‌داده را ندارند. به عبارت دیگر، این الگوریتم‌ها باید از ابتدا همه‌ی داده‌ها را در اختیار داشته باشند تا بتوانند فرایند خوشه‌بندی را انجام دهند. تکنیک‌های داده‌کاوی معمولی که بر روی مجموعه‌داده‌های ایستا به کار گرفته می‌شوند، برای استفاده بر روی جریان‌داده‌ها مناسب نیستند. چرا که در یک محیط جریان‌داده‌ای، توالی نامحدودی از داده‌ها وجود دارد و این در حالی است که در داده‌کاوی معمولی مجموعه‌ی داده‌ها از قبل شناخته شده است. در نتیجه نیاز است که پردازش سریع و مناسب‌تری روی نمونه‌های یک جریان‌داده انجام شود. در همین راستا خوشه‌بندی جریان‌داده را شرح داده‌ایم.

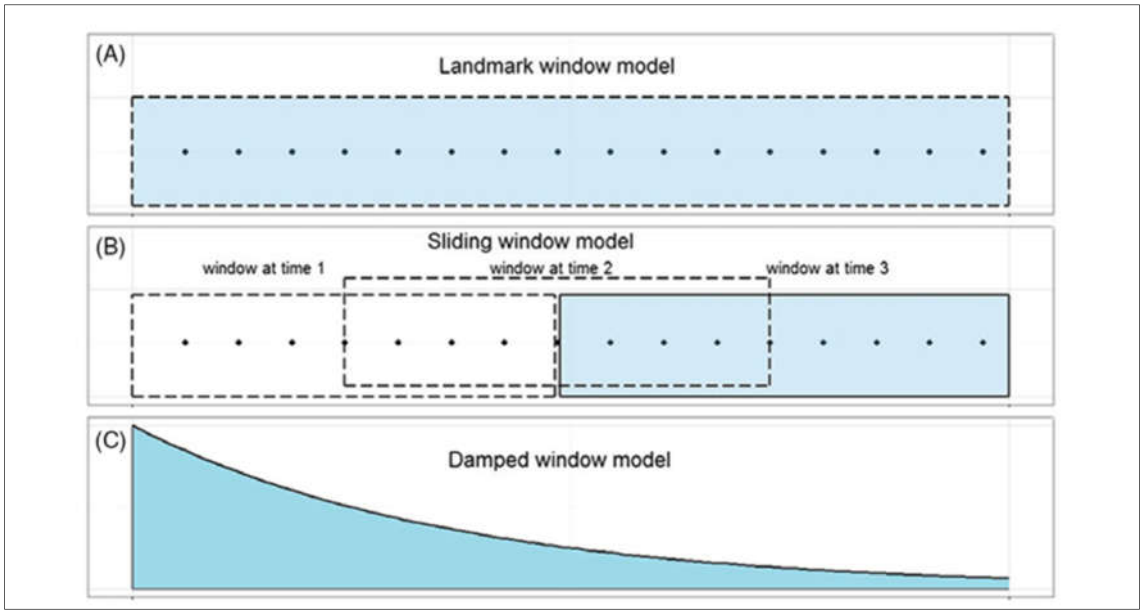
۲-۴ خوشه‌بندی جریان‌داده

با توجه به فراوانی جریان‌داده‌ها در سال‌های اخیر، خوشه‌بندی جریان‌داده از حوزه‌های مهم و پرکاربرد به حساب می‌آید. در همین راستا، نیاز به درک مفاهیم مربوط به چگونگی برخورد و کار با این‌گونه از داده‌ها تا حد زیادی احساس می‌شود. خوشه‌بندی بدون ناظر در بردارنده‌ی یکی از مرسوم‌ترین روش‌های داده‌کاوی برای به دست آوردن اطلاعات و داشتن یک برداشت و بینش نسبت به درون این‌گونه از داده‌ها به شمار می‌آید. فرایند خوشه‌بندی به خودی‌خود، یک عمل چالشی محسوب می‌شود. این در حالی است که خوشه‌بندی بر روی جریانی از داده‌ها، شامل چالش‌های اضافی دیگری از قبیل نداشتن همه‌ی داده‌ها از ابتدا است. علاوه بر این، سروکار داشتن با جریان‌داده‌هایی که به صورت دائم و با سرعت بسیار زیاد تغییر می‌کنند، بر این نکته تاکید دارد که مدل خوشه‌بندی استخراج شده بر روی این‌گونه از داده‌ها درحقیقت قابلیت تکامل به مرور زمان را نیاز خواهد داشت. به لحاظ تئوری، یک جریان‌داده بی‌نهایت است. برای کمک به این موضوع که کدام بخش از جریان‌داده به الگوهای داده‌کاوی کمک می‌کند، مدل‌های پنجره مورد استفاده قرار می‌گیرد. چندین مدل پنجره

تاکنون معرفی شده است که محبوب‌ترین آن‌ها عبارت‌اند از LandMark Window، Sliding Window و Damped Window [۱۲].

- LandMark Window: در این مدل، خوشه‌بندی از یک نقطه‌ی شروع که LandMark نام دارد آغاز شده و تا لحظه‌ی مورد نظر ادامه پیدا می‌کند. برای این پنجره می‌توان تعدادی نمونه‌ی مشخص (مثلا برای هر ۱۰۰۰ نمونه) و یا یک بازه‌ی زمانی (هفتگی، ماهانه و غیره) تعیین کرد. در مدل LandMark Window، هنگامی که یک پنجره‌ی جدید از داده‌ها شروع می‌شود، تمامی نمونه‌هایی که از قبل نگهداری شده‌اند حذف می‌شوند. همچنین در این مدل، داده‌های موجود در پنجره از اهمیت یکسانی برخورد هستند.
- **Sliding Window** : در این مدل، پنجره‌ای با اندازه‌ی ثابت w، از یک نقطه‌ی زمانی مانند t وجود دارد. با گذشت زمان از نقطه‌ی زمانی t، این پنجره مقدار ثابت w را حفظ و شروع به لغزش می‌کند. این مدل بیشتر برای کاربردهایی مناسب است که در آن‌ها جدیدترین نمونه‌های داده از اهمیت بیشتری برخوردار باشند.
- **Damped Window** : نمونه‌های جریان‌داده در این مدل دارای یک وزن هستند که به زمان ورود هر نمونه بستگی دارد. زمانی که یک نمونه‌ی جدید وارد می‌شود، بیشترین وزن ممکن به آن نمونه اختصاص پیدا می‌کند. این وزن با گذشت زمان کاهش می‌یابد. برخلاف دو مدل قبل که در مورد آن‌ها توضیح داده شد، مدل Damped Window نمونه‌ها را به طور کامل حذف نمی‌کند بلکه داده‌هایی که دارای وزن کمتری هستند (مدت زمان بیشتری از ورودشان می‌گذرد) کمتر مورد استفاده قرار می‌گیرند [۱۲, ۱۳].

به منظور روشن‌تر شدن تعاریف مربوط به مدل‌های پنجره‌ای مطرح شده برای کار با جریان‌داده‌ها شکل ۲-۱ در ادامه قرار گرفته است.



شکل ۲-۱. نمایش مدل‌های مختلف پنجره‌گذاری [۱۲]

یکی از اولین الگوریتم‌های خوشه‌بندی کاربردی در زمینه‌ی جریان‌داده‌ها، الگوریتم BIRCH[۱] است. این الگوریتم یک روش اکتشافی[۲] است که با استفاده از مشاهداتی که انجام می‌دهد، فضای مربوط به نقاطی که معمولا به صورت غیر یکنواخت اشغال شده است را به دست می‌آورد. در این روش مجموعه‌داده‌های ورودی اسکن شده و هر کدام از این مجموعه‌داده‌ها به صورت نواحی متراکم خلاصه‌سازی می‌شوند. سپس با استفاده از روش‌های خوشه‌بندی سلسله مراتبی سنتی، نواحی خلاصه‌شده مورد خوشه‌بندی قرار می‌گیرند. در نتیجه با این کار، مسئله‌ی خوشه‌بندی کل داده‌های ورودی به مسئله‌ی خوشه‌بندی مجموعه نقاط خلاصه شده که اندازه‌ای به مراتب کوچکتر از اندازه‌ی مجموعه‌ی اصلی داده‌ها دارد، کاهش پیدا می‌کند. الگوریتم BIRCH به صورت پیوسته نزدیکترین جفت خوشه‌ها را تا زمان رسیدن به خوشه‌های نهایی با یکدیگر ادغام می‌کند [۱۴]. یکی دیگر از الگوریتم‌های شناخته شده‌ی موجود بر روی جریان‌دادها، الگوریتم StreamLS است. این الگوریتم،

[۱] Balanced Iterative Reducing And Clustering Using Hierarchies (BIRCH)

[۲] Heuristic

جریان‌داده‌ی ورودی را به تکه‌های[1] کوچکی تقسیم کرده و الگوریتم خوشه‌بندی K-Means را با استفاده از یک روش جستجوی محلی بر روی این تکه‌ها به اجرا در می‌آورد. در انتهای فرایند مربوط به الگوریتم StreamLS، روش جستجوی محلی بر روی خوشه‌های ایجاد شده مجددا اجرا می‌شود تا خوشه‌بندی نهایی بر روی تمامی داده حاصل گردد [۱۵].

از روش‌های دیگری که در زمینه‌ی خوشه‌بندی جریان‌داده می‌توان به آن اشاره داشت، روش StreamKM++ است. این روش یکی از تازه‌ترین الگوریتم‌های توسعه یافته برای خوشه‌بندی جریان‌داده با استفاده از روش پنجره‌گذاری LandMark است. الگوریتم StreamKM++ در دسته‌ی روش‌هایی قرار می‌گیرد که از خلاصه‌کردن داده‌ها به منظور خوشه‌بندی جریانی از داده بهره می‌برند. در این الگوریتم، درختی از داده‌های هسته (خلاصه‌ی داده‌ها) با هدف خلاصه‌سازی داده‌ها بر پایه‌ی روش KMeans++ ایجاد می‌شود. هزینه‌ی ساخت نسبتا بالای مورد نیاز برای شکل‌گیری درخت مورد اشاره در این روش، از معایب موجود برای این الگوریتم خوشه‌بندی به شمار می‌آید [۱۶].

الگوریتم DGClust از جمله روش‌های دیگر موجود برای خوشه‌بندی جریان‌داده است. DGClust یک الگوریتم خوشه‌بندی توزیع شده برای داده‌های حسگر است که سلول‌های مشبک را برای خلاصه کردن جریان‌داده به کار می‌گیرد. همچنین این روش داده‌ها را از حسگرهای مختلف دریافت کرده و عملیات خوشه‌بندی را به انجام می‌رساند. هر کدام از حسگرها یک داده را تولید می‌کند که این داده‌های تولید شده به صورت محلی در هر حسگر مورد پردازش قرار می‌گیرد. زمانی که یک تغییر وضعیت در سلول مشبک اتفاق می‌افتد، برای اطلاع رسانی در مورد این تغییر با پایگاه مرکزی پردازش داده‌ها ارتباط برقرار می‌شود [۱۷, ۱۸] .

روش دیگری که در زمینه‌ی خوشه‌بندی جریان‌داده می‌توان به آن اشاره داشت، الگوریتم G2CS است. این روش که از یک پنجره‌ی لغزان استفاده می‌کند محدودیتی برای تعداد داده‌هایی که

[1] Chunks

خلاصه‌سازی می‌شوند، در نظر نمی‌گیرد. همچنین این روش پیچیدگی زمانی زیادی دارد. علاوه بر آن، روش مذکور سرباری را برای ایجاد و نگهداری شبکه‌ی توری به صورت غیر ضروری تولید می‌کند تا بتواند اندازه‌های مختلفی که در جریان‌داده تولید می‌شود را به خوبی مدیریت نماید [۱۹, ۲۰].

روش دیگری که در انتهای این بخش به صورت خلاصه به توضیح آن پرداخته شده است، روش ID-AP نام دارد. این روش یک الگوریتم خوشه‌بندی نیمه نظارتی است که فرایند خوشه‌بندی را با استفاده از داده‌های برچسب‌گذاری شده و داده‌هایی که برچسب ندارند، به انجام می‌رساند. روش مورد اشاره جهت مدیریت افزایش و کاهش داده‌ها طراحی شده است اما نمی‌تواند به صورت دقیق پنجره‌های چندتایی منقضی را از بین ببرد [۱۹, ۲۱].

۲-۵ بررسی پژوهش‌های انجام شده

از جمله مطالعات موجود در راستای استفاده از برچسب داده‌ها در خوشه‌بندی می‌توان به خوشه‌بندی با ناظر[۱] [۲۲]، خوشه‌بندی مقید [۴] و خوشه‌بندی تطبیقی [۲] اشاره کرد. خوشه‌بندی مقیدی که در [۴] معرفی شده است، بر پایه‌ی روش K-Means بوده و برای قرار گرفتن یا نگرفتن جفت داده‌ها در یک خوشه، قیدهایی در نظر گرفته است. همچنین این روش به کاربران اجازه می‌دهد که با در اختیار داشتن یک دانش قبلی در مورد داده‌ها، به تجزیه و تحلیل داده‌ها بپردازند.

خوشه‌بندی تطبیقی [۲] که بر پایه‌ی خوشه‌بندی متداول K-Means بنا شده است، با در نظر گرفتن برچسب داده‌ها رویکرد نوینی را برای خوشه‌بندی معرفی کرده است. در این رویکرد، تعدادی از خوشه‌ها باید حاوی حداقلی از اعضای یک گروه در مجموعه‌ی داده باشند، تا بتوان تحلیل کاربردی مناسبی از خوشه‌ها ارائه نمود. این روش به صورت متمرکز عمل کرده و به دلیل نیاز به تمام داده‌ها مناسب محیط جریان‌داده‌ای نیست.

---

[۱] Supervised

جین و همکارانش [۲۲]، یک روش جدید به منظور پردازش سریع‌تر داده‌ها و بالا بردن دقت فرایند خوشه‌بندی ارائه کرده‌اند. در این روش برای کاهش خطای خوشه‌بندی، از یک راهکار بهینه‌سازی در محاسبه‌ی تعداد تکرار مراحل الگوریتم استفاده شده است.

خوشه‌بندی جریان‌داده در مطالعات متعددی مورد بررسی قرار گرفته است. سینها و همکارانش [۲۳] یک الگوریتم خوشه‌بندی مبتنی بر K-Means برای خوشه‌بندی داده‌های بزرگ به صورت توزیع شده، ارائه داده‌اند که در آن نیازی به مشخص بودن تعداد خوشه‌ها نیست. در این روش، داده‌ها به قسمت‌های کوچک تقسیم شده و در واحدهای محاسباتی موجود توزیع می‌شوند. پردازش داده‌ها به صورت موازی در این واحدها صورت می‌گیرد.

فیتریانی و همکارانش [۲۴]، برای رفع مشکل کُند عمل کردن روش خوشه‌بندی K-Means بر روی مجموعه داده‌های بزرگ، یک روش برخط برای خوشه‌بندی این داده‌ها ارائه کرده‌اند. این روش از دسته‌های کوچک تصادفی که حاوی خلاصه‌ای از داده‌ها هستند، استفاده می‌کند.

لیبرتی و همکارانش [۲۵]، الگوریتمی به منظور برخط کردن خوشه‌بندی K-Means معرفی کرده‌اند. علاوه بر این، روش‌هایی به منظور سرعت بخشیدن به عمل خوشه‌بندی روی داده‌ها نیز در برخی مطالعات ارائه شده است. برای نمونه، ژانگ و همکارانش [۲۶] یک روش مبتنی بر ایده‌ی خلاصه‌سازی داده‌ها ارائه کرده‌اند که از این خلاصه‌ها با هدف سرعت بخشیدن به فرایند خوشه‌بندی نمونه‌ها استفاده کرده‌اند.

کمیتو و همکارانش [۲۷]، یک الگوریتم خوشه‌بندی جریان‌داده، با هدف پیدا کردن عناوین[۱] در داده‌های یک شبکه‌ی اجتماعی معرفی کرده‌اند. در این الگوریتم مراکز خوشه‌ها به شکل خلاصه‌سازی شده، نگهداری می‌شوند. هر نمونه‌ی ورودی با استفاده از یک معیار شباهت به یکی از این مراکز تعلق گرفته و مرکز انتخاب شده به‌روزرسانی می‌شود. همچنین الگوریتم خوشه‌بندی مذکور به مباحث

[۱] Topics

مربوط به روش‌های خوشه‌بندی جریان‌داده‌ها در شبکه‌ی اجتماعی توییتر پرداخته و آن‌ها را در سه دسته‌ی مبتنی بر توییت، مبتنی بر هشتگ و مبتنی بر کلمات کلیدی پر تکرار قرار داده است. هر کدام از این سه مورد می‌تواند به عنوان یک دانش زمینه‌ای یا همان برچسب داده‌ها تلقی شود.

بابلارگئون و همکارانش نیز، روشی برای خوشه‌بندی پست‌های شبکه‌ی اجتماعی با هدف گروه‌بندی توییت‌های مشابه یکدیگر طراحی کرده‌اند [۲۸]. این روش سعی دارد با پردازش توییت‌ها در دسته‌های کوچک تشکیل شده در واحد زمان به پردازش داده‌های بزرگ بپردازد. در واقع با انجام این کار، جریان‌داده‌ی موجود در زمان واقعی، به سرعت در مجموعه‌داده‌ها فهرست می‌گردد که در نتیجه کاربران می‌توانند آن را در زمان واقعی مورد استفاده قرار دهند.

در ادامه‌ی این فصل، ابتدا به دلیل آن‌که روش خوشه‌بندی K-Means به عنوان یک ابزار در فرایند خوشه‌بندی تطبیقی پایه و همچنین الگوریتم خوشه‌بندی جریان‌داده‌ای ارائه شده در این پژوهش به کار گرفته شده است (که در فصل سوم به صورت کامل در مورد آن توضیح داده شده)، توضیحات مختصری به صورت تیتر گونه آورده شده است. پس از این توضیحات، شرح مربوط به مراحل الگوریتم خوشه‌بندی تطبیقی ارائه شده در [۲] که ایده‌ی اصلی موجود در آن منجر به شکل گرفتن تفکر معرفی یک روش برخط با نگاه تحلیل روابط گروهی در جریان‌داده شد را قرار داده‌ایم.

۲-۶ خوشه‌بندی K-Means

الگوریتم K-Means یکی از روش‌های متداول خوشه‌بندی بوده که علی‌رغم ساده بودن، یک روش پایه به حساب می‌آید. فرایند اجرای این الگوریتم برای یک تعداد مشخص از خوشه‌ها به صورت زیر است:

- به تعداد خوشه‌ها نقاطی به صورت تصادفی انتخاب می‌شود (مراکز خوشه) که این نقاط در واقع همان میانگین نقاط هر خوشه هستند.
- هر نمونه داده‌ی ورودی به خوشه‌ای نسبت داده می‌شود که کمترین فاصله را تا آن مرکز

خوشه داشته باشد.

- در هر بار تکرار، مراکز جدید از روی میانگین نقاطی که در هر خوشه قرار گرفته‌اند محاسبه می‌شوند. این کار تا زمانی که تغییری در خوشه‌ها ایجاد نشود، ادامه پیدا می‌کند.

در ادامه باید به این نکته توجه داشت که معیار فاصله‌ی مورد استفاده برای این روش خوشه‌بندی عمدتا فاصله‌ی اقلیدسی است. در حالی که معیارهای فاصله‌ی دیگری نیز از جمله فاصله‌ی همینگ، فاصله‌ی جیسون و غیره وجود دارند که می‌توانند برای محاسبه‌ی فاصله مورد استفاده قرار بگیرند [۲۹]. در این روش خوشه‌بندی، خوشه‌هایی که در انتها شکل می‌گیرند همواره یکسان نبوده و یک جواب بهینه نخواهند داشت. به صورت کلی می‌توان برای این روش، مشکلات زیر را نام برد:

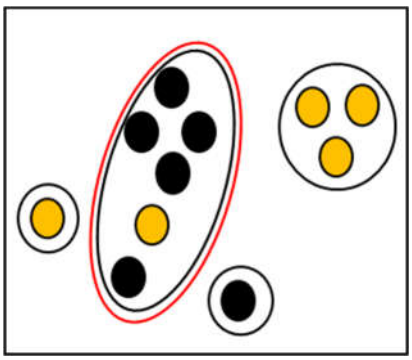
- خوشه‌های انتهایی به انتخاب مراکز اولیه وابسته است.
- یک راهکار مشخص برای به دست آوردن مراکز اولیه‌ی مناسب برای خوشه‌ها وجود ندارد.
- در این روش تعداد خوشه‌ها باید از ابتدا مشخص باشد. در حالی که در کاربردهای واقعی این تعداد از قبل مشخص نیست.

۲-۷ خوشه‌بندی تطبیقی

در این قسمت به معرفی روش خوشه‌بندی تطبیقی ارائه شده در [۲] به صورت خلاصه پرداخته‌ایم. هدف روش خوشه‌بندی تطبیقی که بر پایه‌ی خوشه‌بندی K-Means عمل می‌کند، کشف روابط گروهی از طریق تجزیه و تحلیل نمونه‌های موجود در یک مجموعه‌داده است. منظور از روابط گروهی، بررسی گروه یا گروه‌های موجود در خوشه‌های ایجاد شده است. چنین امکانی با تعریف چند مفهوم از جمله t-accordant و (r, t)-accordant به وجود می‌آید. فرض کنیم مجموعه‌داده‌ی $X \subset R^n$ شامل m گروه به صورت $X = \{X_1 \cup \dots \cup X_m\}$ باشد. یک خوشه‌بندی بر روی مجموعه‌داده‌ی X، k مرکز $C = \{C_1, C_2, \dots, C_k\}$ را تولید می‌کند. تناسب موجود در داده‌های یک گروه X_i که در یک خوشه

قرار گرفته‌اند، توسط بردار $f = \frac{1}{|X_i|}[|C_1 \cap X_i|, \dots, |C_k \cap X_i|]$ تعریف می‌شود. خوشه‌بندی C بر روی گروه X_i دارای خاصیت t-accordant است، اگر $\exists\, j \in \{1, \dots, k\}$ به طوری که j اُمین عنصر بردار f بزرگتر یا مساوی با t باشد ($f_j \geq t$).

یک خوشه‌بندی، (r, t)- accordant است، اگر داده‌های حداقل r گروه، شرط t-accordant را داشته باشند. به عنوان مثال، در شکل ۲-۲ مفهوم (1, 0.75)-accordant رعایت شده است، به این معنی که داده‌های حداقل یک خوشه (که با حاشیه‌ی قرمز رنگ مشخص شده است)، حداقل ۷۵ درصد از نمونه‌های یک گروه (دایره‌های سیاه رنگ) را دارا هستند.



شکل ۲-۲. خوشه‌بندی (1 , 0.75)-accordant [۲]

در ادامه‌ی این قسمت، مراحل الگوریتم خوشه‌بندی تطبیقی برای انجام فرایند خوشه‌بندی داده‌ها به k خوشه شرح داده شده است. در این روش، ابتدا از مجموعه‌داده‌ی موجود k مرکز به صورت تصادفی انتخاب می‌شود. سپس تا زمانی که الگوریتم به شرط پایانی برسد، مراحل زیر به اجرا در می‌آیند:

۱- برای هر نمونه‌ی موجود در مجموعه‌داده با کمک گرفتن از رابطه‌ی ۲-۱، میزان هزینه[۱] محاسبه می‌شود. در واقع با این کار ماتریسی از هزینه بر اساس رابطه‌ی اشاره شده به دست می‌آید.

[۱] Penalty

$$p_{ij} = dist(x_i, c_j) - min_\gamma(x_i, c_\gamma) \qquad (۱-۲)$$

در این فرمول، x_i نماینده‌ی داده i اُم، c_j برابر با مرکز خوشه j اُم، p_{ij} بیانگر میزان هزینه داده i اُم نسبت به خوشه j اُم، dist نشان‌دهنده‌ی فاصله‌ی اقلیدسی و min بیانگر کمترین فاصله‌ی آن نمونه تا نزدیک‌ترین مرکز است. به این ترتیب می‌توان یک ماتریس هزینه با ابعاد $N \times K$ تشکیل داد که N تعداد داده‌ها و K تعداد کل خوشه‌ها است [۲].

۲- سپس برای هر گروه و هر مرکز، داده‌های موجود در گروه بر اساس میزان هزینه‌ای که در مرحله‌ی ۱ محاسبه شده است به صورت صعودی مرتب می‌گردد. حال، هزینه‌ی مربوط به t درصد از داده‌هایی که به این صورت مرتب شده‌اند، با یکدیگر جمع شده و بدین ترتیب برای هر جفت گروه - مرکز[۱] موجود در کل داده‌ها، یک هزینه‌ی جدید به دست می‌آید.

۳- در ادامه، r عدد از جفت‌های گروه - مرکزی که دارای کم‌ترین میزان هزینه‌ی جدید هستند، انتخاب می‌شوند.

۴- t درصد از اولین جفت‌های انتخاب شده از مرحله‌ی ۳، به مرکز خوشه‌ی مرتبط با خود نسبت داده می‌شوند.

۵- داده‌های دیگری که باقی می‌مانند، به روش K-Means به نزدیکترین مراکز خوشه نسبت داده می‌شوند.

در ادامه‌ی توضیحات مربوط به روش خوشه‌بندی تطبیقی، می‌توان به این نکته اشاره داشت که پیچیدگی زمانی این روش از مرتبه‌ی $O(m \times n_{max} \times \log(n_{max}) \times k \times q)$ است. در این مرتبه‌ی زمانی، m برابر با تعداد گروه‌های موجود در مجموعه‌داده، n_{max} بیشترین تعدادی که از یک گروه در مجموعه‌داده وجود دارد، k تعداد خوشه‌ها و q بُعد[۲] مربوط به داده‌هاست.

---

[۱] Group_Center

[۲] Dimention

حال در انتهای این فصل به منظور آنکه بتوانیم مقایسه‌ی جامعی بین پژوهش‌های انجام شده داشته باشیم، در جدول ۲-۱ خلاصه‌ای از روش‌های مختلف به همراه شرح و ویژگی‌های هر یک از آن‌ها در فرایند خوشه‌بندی ارائه شده است.

جدول ۲-۱. خلاصه‌ای از پژوهش‌های انجام شده

روش	شرح مختصر	قابلیت اجرا بر روی جریان‌داده
۱- خوشه‌بندی با ناظر	یک رویکرد جدید با استفاده از برچسب داده‌ها برای بالا بردن سرعت و دقت فرایند خوشه‌بندی [۲۲].	ندارد
۲- خوشه‌بندی مقید	استفاده از یک دانش زمینه‌ای و تعریف یک سری از قیدها به منظور بهبود نتایج ناشی از خوشه‌بندی K-Means [۴].	ندارد
۳- خوشه‌بندی تطبیقی	استفاده از یک دانش زمینه‌ای و معرفی یک الگوریتم خوشه‌بندی جدید با نگاهی تازه برای شکل‌گیری خوشه‌های با معناتر [۲].	ندارد
۴- خوشه‌بندی مبتنی بر K-Means برای خوشه‌بندی داده‌های بزرگ به صورت توزیع شده	استفاده از رویکرد تقسیم داده‌ها به قسمت‌های کوچک در واحدهای محاسباتی توزیع شده بدون نیاز به مشخص بودن تعداد خوشه‌ها از ابتدای کار [۲۳].	دارد
۵- خوشه‌بندی K-Means روی مجموعه داده‌های بزرگ	ارائه‌ی یک روش برخط با رویکرد خلاصه‌سازی داده‌ها برای بالا بردن سرعت اجرای الگوریتم خوشه‌بندی K-Means [۲۴].	دارد
۶- خوشه‌بندی برخط K-Means	الگوریتمی با هدف برخط کردن خوشه‌بندی K-Means با رویکرد خلاصه‌سازی داده‌ها معرفی شده است [۲۵, ۲۶].	دارد

۷- الگوریتم خوشه‌بندی جریان‌داده، با هدف پیدا کردن عناوین در داده‌های یک شبکه‌ی اجتماعی	معرفی رویکرد نگهداری مراکز خوشه‌ها به شکل خلاصه‌سازی شده و تخصیص هر نمونه‌ی ورودی با استفاده از یک معیار شباهت به یکی از این مراکز [۲۷].	دارد
۸- خوشه‌بندی پست‌های شبکه‌ی اجتماعی با هدف گروه‌بندی توییت‌های مشابه یکدیگر	ارائه‌ی روشی با رویکرد پردازش توییت‌ها در دسته‌های کوچک تشکیل شده در واحد زمان و همچنین بهره‌گیری از یک دانش زمینه‌ای (هشتگ توییت‌ها) با هدف پردازش داده‌های بزرگ با سرعت بالا [۲۸].	دارد

۲-۸ جمع‌بندی

در این فصل مفاهیمی از قبیل داده‌کاوی، خوشه‌بندی، جریان‌داده و خوشه‌بندی جریان‌داده را معرفی کردیم. نمونه‌هایی از روش‌های خوشه‌بندی سنتی و خوشه‌بندی جریان‌داده‌ای موجود را نیز بیان کرده و در مورد معایب و فواید موجود در برخی از آن‌ها توضیحاتی آوردیم. پژوهش‌های پیشین انجام شده را بررسی کرده و دیدگاه هر کدام برای استفاده از یک دانش‌زمینه‌ای به منظور بهبود خوشه‌های شکل گرفته را شرح داده‌ایم. در انتهای فصل نیز به توضیح روش خوشه‌بندی تطبیقی پایه و مراحل الگوریتم آن پرداخته‌ایم و چند روش خوشه‌بندی مطرح شده در متن این پایان‌نامه را در قالب یک جدول با هدف مقایسه‌ی بهتر میان روش‌ها گردآوری نموده‌ایم.

فصل ۳ : معرفی روش پیشنهادی

بر اساس مطالبی که در فصل‌های پیشین در ارتباط با خوشه‌بندی تطبیقی مطرح شد، مشاهده کردیم که این روش نگاه تازه‌ای را به مسئله‌ی خوشه‌بندی به وجود آورده است. هدف اصلی این روش، شکل‌گیری خوشه‌های تفسیرپذیرتر و معنادارتر در انتهای فرایند خوشه‌بندی بوده است. بدان معنا که در پایان کار بتواند از خوشه‌های حاصل شده در جهت یک هدف خاص بهره بِبرد. نگاه جدیدی که این روش ایجاد کرده است، توجه به کلاس داده‌های موجود در یک مجموعه‌داده و همچنین تعداد نهایی موجود از هر کدام از این کلاس‌ها در برخی از خوشه‌های نهایی می‌باشد. به عنوان مثال، مجموعه داده‌ی مربوط به یک بیماری خاص که در آن جمعیت زنان و مردان قرار دارد را در نظر بگیرید (در اینجا، مجموعه داده شامل دو کلاس مردان و زنان خواهد بود). به طور فرض اگر قرار باشد سه خوشه بر روی این مجموعه داده شکل بگیرد و یکی از این خوشه‌ها شامل ۷۵ درصد از جمعیت زنان باشد، می‌توان خوشه‌ی حاصل شده را در اختیار متخصصان قرار داد تا داده‌های موجود در آن را با دقت بیشتری مورد بررسی قرار دهند. با این کار می‌توان این انتظار را داشت که عوامل موثر در ابتلای جمعیت زنان موجود در آن خوشه به یک بیماری خاص، بهتر و دقیق‌تر آشکار شود.

همانگونه که در فصل قبل نیز به آن اشاره شد، به طور کلی روش خوشه‌بندی تطبیقی به صورت متمرکز عمل می‌کند و قابل اجرا بر روی مجموعه‌داده‌هایی است که تمام داده‌های موجود در آن‌ها از قبل مشخص و در دسترس الگوریتم خوشه‌بندی باشد. به بیان روشن‌تر، این روش را نمی‌توان به صورت قالب فعلی برای خوشه‌بندی جریان‌های‌داده به کار گرفت. لذا در این پژوهش، ما یک روش خوشه‌بندی تطبیقی جریان‌داده‌ای جدیدی را ارائه کرده‌ایم. این روش مفهوم اصلی موجود در خوشه‌بندی تطبیقی که کشف روابط گروهی در یک مجموعه‌داده است را دارا می‌باشد. همچنین لازم به ذکر است که تاکنون روش خوشه‌بندی جریان‌داده‌ای دیگری، مطابق با دیدگاه مطرح شده در این پژوهش، وجود نداشته است. از ویژگی‌های مهم روش معرفی شده می‌توان افزایشی بودن، میزان استفاده‌ی مناسب از حافظه، سرعت بالای اجرا، پایین بودن پیچیدگی زمانی و همچنین عدم نیاز به

تعداد خوشه‌ها از ابتدای فرایند خوشه‌بندی را نام برد.

۳-۱ چالش‌های موجود

میزان حافظه‌ی مربوط به پردازشگر که وظیفه‌ی پردازش اطلاعات جریان‌داده‌ی ورودی را بر عهده دارد، از جمله مهم‌ترین مواردی است که در مباحث خوشه‌بندی نقش به‌سزایی را دارا می‌باشد. به همین دلیل پیشنهاد یک روش خوشه‌بندی مطلوب که بتواند مصالحه‌ی مناسبی را بین میزان حافظه‌ی مصرفی و زمان انجام پردازش ایجاد نماید، می‌تواند اولویت بالایی داشته باشد. حافظه‌ی مصرفی روش خوشه‌بندی از جمله چالش‌های مهمی است که راهکار خوشه‌بندی معرفی شده، توانسته است به خوبی بر آن غلبه کند. علت اصلی این موضوع، آن است که روش خوشه‌بندی پیشنهادی ما بر روی خلاصه‌ای از داده‌ها عملیات مربوط به خوشه‌بندی را به انجام می‌رساند. بنابراین واضح است که حافظه‌ی مصرفی برای نگهداری داده‌های خلاصه در این روش، بسیار کمتر از حالتی است که همه‌ی داده‌ها در حافظه نگهداری می‌شوند.

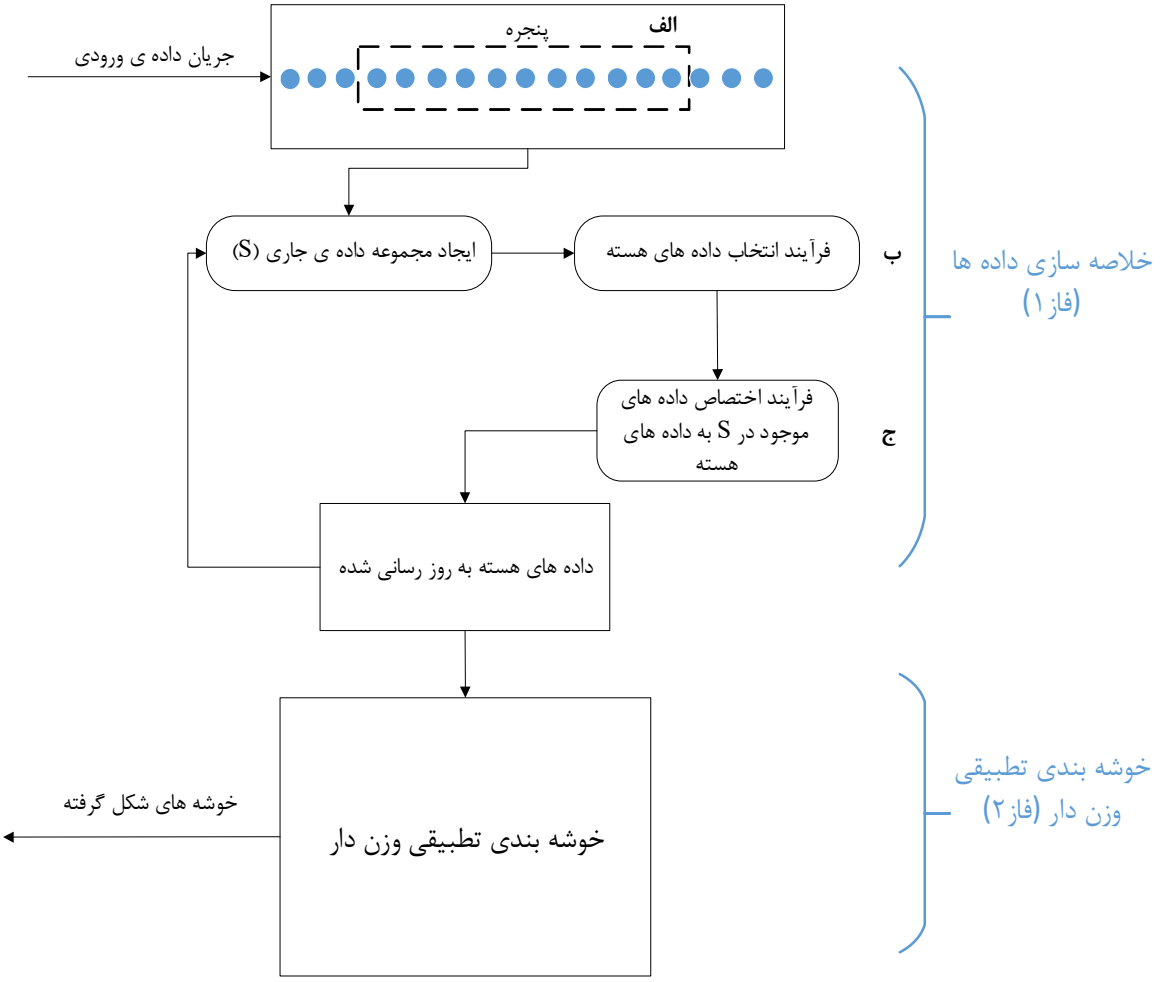
از جمله چالش‌های دیگری که روش پیشنهادی به آن پرداخته، افزایش سرعت اجرای فرایند خوشه‌بندی است. با توجه به این‌که خوشه‌بندی معرفی شده در این پژوهش بر روی خلاصه‌ای از داده‌ها عملیات خود را به انجام می‌رساند، طبیعی است که زمان اجرای این روش نسبت به روش‌های سنتی خوشه‌بندی که تمامی داده‌ها را مد نظر قرار می‌دهند، کمتر باشد. از آنجایی که روش‌های سنتی بر روی تمامی مجموعه‌ی داده‌ها عمل می‌کنند، لذا میزان مصرف حافظه و زمان اجرای الگوریتم در آن‌ها ممکن است طولانی‌تر شود. بنابراین روش ما از طریق اعمال فرایند خوشه‌بندی بر روی خلاصه‌ای از داده‌ها سعی در کنترل این مشکل نیز داشته است.

با توجه به چالش‌های مطرح شده از جمله میزان حافظه‌ی مصرفی و سرعت اجرای الگوریتم خوشه‌بندی، می‌توان این انتظار را داشت که پیچیدگی محاسباتی الگوریتم ارائه شده نیز کنترل شده و بهبود یافته است. ما در ادامه‌ی این فصل، به بیان هدف کلی خود از انجام پایان‌نامه، شرح کامل

مراحل و قسمت‌های موجود در فرایند خوشه‌بندی جریان‌داده‌ای جدید پیشنهادی می‌پردازیم.

هدف اصلی که در این پژوهش دنبال می‌شود، ایجاد یک روش خوشه‌بندی تطبیقی جریان‌داده‌ای بر پایه الگوریتم خوشه‌بندی تطبیقی است که در نهایت به کشف روابط گروهی میان داده‌ها در جریان داده می‌انجامد. اجرای این عمل مستلزم طی نمودن مراحلی است که به صورت گام به گام در ادامه آورده شده است. همچنین برای آن‌که ما قابلیت انجام خوشه‌بندی تطبیقی به صورت جریان‌داده‌ای را داشته باشیم، به طور کامل نسبت به ایجاد تغییرات در نسخه‌ی اولیه‌ی الگوریتم خوشه‌بندی تطبیقی پرداختیم. به منظور درک بهتر روش پیشنهادی، فرایند کلی آن در شکل ۳-۱ آورده شده است. در این شکل، شاهد دو فاز کلی هستیم. در فاز اول، خلاصه‌سازی داده‌ها و در فاز دوم فرایند نهایی خوشه‌بندی که در این پژوهش برای آن عنوان خوشه‌بندی تطبیقی وزن‌دار در نظر گرفته شده است، به انجام می‌رسد. این دو فاز می‌توانند به صورت مستقل اجرا شوند. همچنین لازم به ذکر است که فرایند خوشه‌بندی در هر زمانی می‌تواند برای هر تعداد خوشه‌ی دلخواه اجرا شده و نتیجه‌ی خوشه‌بندی تا لحظه‌ی مورد نظر محاسبه گردد.

فرایند کلی موجود در شکل ۳-۱ به این صورت است که ابتدا یک پنجره از داده‌های جریان‌داده وارد فاز یک یا همان خلاصه‌سازی داده‌ها شده و پس از انجام مراحلی که در بخش بعدی معرفی می‌شوند، یک خروجی تولید خواهد شد. خروجی حاصل از این فاز به عنوان ورودی به فاز دوم یا همان خوشه‌بندی نهایی داده خواهد شد. در ادامه به صورت کامل و دقیق به توضیح هر فاز و عناصر آن پرداخته می‌شود و جزئیات هر بخش به همراه عملیات مربوط به آن، بازگو می‌گردد.



شکل ۳-۱. فرایند کلی خوشه‌بندی تطبیقی جریان‌داده

۳-۱-۱ خلاصه‌سازی داده‌ها (فاز اول)

اولین گام مورد نیاز درجهت برخط کردن نسخه‌ی اولیه‌ی خوشه‌بندی تطبیقی، فرایند خلاصه‌سازی داده‌ها است. برای این منظور از مفاهیم موجود در روش StreamKM++ الهام گرفته شد [۱۶]. از جمله‌ی این مفاهیم می‌توان به مفهوم مدل پنجره‌ی LandMark و داده‌های هسته[۱] اشاره کرد که شرح مربوط به این مدل پنجره‌ای در فصل ۲ مورد بررسی قرار گرفت. اما به طور خلاصه می‌توان این‌گونه بیان کرد که در این مدل، جریان‌داده‌ی ورودی در قالب تعدادی نمونه‌های مشخص وارد پنجره‌ای (محل ذخیره‌ی داده‌ها) با اندازه‌ی ثابت شده و مورد پردازش قرار می‌گیرد. پس از آن‌که

[۱] Coresets

پردازش مورد نظر به انجام رسید، داده‌هایی که در پنجره قرار گرفته‌اند حذف شده و داده‌های جدید از جریان‌داده جایگزین آن‌ها می‌شوند.

در ادامه‌ی این بخش به بررسی مفهوم داده‌های هسته پرداخته می‌شود و چگونگی ایجاد آن‌ها را به طور کامل شرح می‌دهیم. داده‌های هسته به منظور خلاصه‌سازی تک‌داده‌ها به کار می‌روند. به این صورت که زیرمجموعه‌ای از داده‌ها در یک داده‌ی‌هسته خلاصه می‌شود. یک تعریف ریاضی از داده‌های هسته در قالب فرمول ۳-۱ آورده شده است.

$$D_{cs} = \{\, \mathrm{d} \mid \mathrm{d} \in D \ \wedge\ \forall\, cs' \neq cs \ .\ \| \mathrm{d} - \mathrm{cs} \| \leq \| \mathrm{d} - \mathrm{cs}' \| \,\}$$

$$cs = \frac{\sum_{d \in D_{cs}} d}{|D_{cs}|} \quad , \quad D_{cs} \subset D \qquad (۱-۳)$$

در این فرمول cs یک داده هسته‌ی نمونه است که مجموعه‌ی D_{cs} را خلاصه می‌کند، D نشان‌دهنده‌ی کل داده‌ها ،D_{cs} زیرمجموعه‌ای از کل داده‌ها و d بیانگر یک نمونه از داده است. داده‌های هسته به عنوان یک مفهوم انتزاعی درنظر گرفته شده‌اند که این مفهوم خود شامل پارامترهای از جمله : ۱- وزن (w) ۲- بردار گروه-تعداد[۱] ($\widehat{\mathrm{gc}}$) و ۳- بُعد مربوط به داده‌ها است. این سه پارامتر در طی فرایند خلاصه‌سازی شکل گرفته و به‌روزرسانی می‌شوند.

به طور کلی منظور از وزن هر داده‌ی هسته در این پژوهش، تعداد نمونه‌های ورودی تعلق گرفته به هر داده‌ی هسته است. به عنوان مثال اگر وزن یک داده‌ی هسته برابر با ۸ باشد، به این معناست که تعداد ۸ نمونه از کل نمونه‌های یک جریان‌داده، کمترین فاصله (فاصله اقلیدسی) را با آن داده‌ی هسته داشته‌اند. در ادامه باید گفت بردار گروه-تعداد، پارامتری است که به کمک آن می‌توان مشخص کرد که چه تعدادی از هر کلاس موجود در مجموعه‌داده به وسیله‌ی داده‌های هسته خلاصه‌سازی شده‌اند. به عبارت دیگر این بردار نشان‌دهنده‌ی توزیع مربوط به کلاس داده‌های خلاصه شده توسط داده‌های

[۱] Group_Count

هسته است و در هنگام فرایند خلاصه‌سازی به وجود می‌آید. برای درک بهتر این پارامتر به مثال ۳-۱ توجه کنید.

مثال ۳-۱. فرض کنید که تعدادی از نمونه‌های یک جریان‌داده در اختیار روش خلاصه‌سازی قرار گرفته است و در انتها برداری همانند جدول ۳-۱ برای یکی از خلاصه‌داده‌ها (داده‌های هسته) شکل گرفته باشد.

جدول ۳-۱. نمونه‌ای از بردار گروه-تعداد

گروه<۳>	گروه<۲>	گروه<۱>
۱	۴۰	۱۰

مفهوم موجود در این بردار آن است که از میان داده‌هایی (در اینجا ۵۱ نمونه) که در داده‌ی هسته خلاصه شده‌اند، ۱۰ داده از گروه یا کلاس اول، ۴۰ داده از گروه دوم و یک داده از گروه سوم بوده است. از این بردار در بخش بعدی که مربوط به فاز دوم است، استفاده شده است. همچنین منظور از بُعد هر نمونه از داده‌ها، در حقیقت همان ویژگی‌های مربوط به داده است که همه‌ی این ویژگی‌ها به غیر از ویژگی کلاس (یا همان گروه) به صورت عددی در نظر گرفته شده‌اند.

در فاز اول به طور خلاصه سه بخش اصلی وجود دارد که در شکل ۳-۱ با حروف الفبا مشخص شده‌اند. در ادامه به توضیح این سه بخش پرداخته‌ایم.

الف) پنجره: در این قسمت، عملیات پنجره‌گذاری به اجرا در می‌آید. به عبارت دیگر بر روی جریان‌داده‌ای که قرار است وارد فاز پردازشی شود، یک قاب یا پنجره با اندازه از پیش تعریف شده قرار می‌گیرد.

ب) انتخاب داده‌های هسته: در این بخش انتخاب داده‌های هسته به این صورت انجام می‌پذیرد که در آن به تعداد مورد نیاز، داده‌ی هسته با استفاده از روش انتخاب مراکز اولیه‌ی KMeans++

انتخاب می‌شود. روش انتخاب موجود در الگوریتم KMeans++ به این واقعیت اشاره دارد که کیفیت خوشه‌بندی ناشی از روش K-Means به شدت به مجموعه‌ی اولیه‌ی مراکز بستگی دارد. در همین راستا روش انتخاب مراکز اولیه در الگوریتم KMeans++ موجب بهتر شدن سرعت و دقت خوشه‌بندی K-Means شده است [۳۰].

به طور خلاصه فرایند انتخاب نقاط اولیه در روش KMeans++ به این صورت است که ابتدا یک نقطه به صورت کاملا تصادفی از میان داده‌ها انتخاب می‌شود. سپس برای انتخاب نقاط دیگر، به هر کدام از نقاط مجموعه‌داده یک احتمال نسبت داده می‌شود. احتمال انتخاب نقطه‌ی a متناسب است با $D_i(a)^2$ که $D_i(a)$ فاصله‌ی نقطه‌ی a تا نزدیک‌ترین نقطه‌ی انتخاب شده‌ی قبلی است. در واقع بعد از انتخاب یک نقطه، نقطه‌ی بعد هرچه از نقاط قبل‌تر خود دورتر باشد، احتمال بالاتری برای انتخاب دارد. روند انتخاب نقاط تا زمانی ادامه پیدا می‌کند که به تعداد نهایی نقاط اولیه‌ی مورد نظر برسیم.

دلیل اصلی استفاده از دیدگاه الگوریتم KMeans++ در روش خوشه‌بندی جریان‌داده‌ای پیشنهادی که برای انتخاب داده‌های هسته استفاده می‌شود، آن است که قصد داشتیم تا شرایطی را به وجود آوریم که در آن نقاط انتخابی تا جای ممکن از پراکندگی مناسبی برخوردار باشند و جنبه‌های مختلف داده را پوشش دهند.

ج) تخصیص داده‌ها: در این مرحله داده‌هایی که درون یک پنجره قرار گرفته‌اند، توسط یک معیار فاصله به داده‌های خلاصه شده اختصاص پیدا می‌کنند. این معیار در روش پیشنهادی، فاصله اقلیدسی در نظر گرفته شده است. وزن داده‌ی هسته به صورت تعداد داده‌هایی که به آن نزدیک بوده به‌روزرسانی شده و متناسب با آن مطابق با فرمول ۳-۱، بُعد مربوط به آن داده‌ی هسته نیز تغییر می‌یابد. لازم به ذکر است که در همین مرحله، بردار گروه-تعداد به وجود آمده و مقداردهی می‌شود.

با توجه به مطالب بیان شده، با انجام این عملیات، خلاصه‌ای از داده‌ها تولید می‌شود که این خلاصه‌ها، در واقع خروجی فرایندهای اعمال شده بر روی داده‌ها در فاز یک هستند. ما به منظور شکل‌گیری این خلاصه‌ها الگوریتمی را طراحی کردیم که در ادامه به صورت گام به گام به توضیح مراحل موجود در آن خواهیم پرداخت. داده‌های خلاصه شده (داده‌های هسته‌ی به‌روزرسانی شده) به عنوان یک ورودی در اختیار عناصر پردازشی موجود در فاز دو مطابق با فرایند شکل ۳-۱ قرار می‌گیرد.

در ابتدا برای مطالعه ساده‌تر و همچنین روشن بودن تمامی زوایای الگوریتم خلاصه‌سازی داده‌ها، کلیه پارامترهای به کار رفته در جدول ۳-۲ گردآوری شده است.

جدول ۳-۲. پارامترهای موجود در الگوریتم SAKmeans

پارامترهای الگوریتم	مفهوم
Window	پنجره
M	تعداد خلاصه‌داده‌ها (داده‌های هسته)
r	پارامتر r در خوشه‌بندی تطبیقی
t	پارامتر t در خوشه‌بندی تطبیقی
k	تعداد خوشه‌ها

پس از مشخص شدن پارامترهای به کار گرفته شده در الگوریتم مذکور، در ادامه به بررسی شبه کد مربوط به شکل‌گیری این خلاصه‌داده‌ها توسط الگوریتم SAKmeans مطابق با الگوریتم ۱-۳ پرداخته می‌شود و سپس جزئیات این شبه کد و روال کلی موجود در آن را توضیح می‌دهیم.

Algorithm: Streaming Accordant Kmeans (SAKmeans)

Input: $data\ stream, M$

$CS = \emptyset$

repeat $for\ each\ received\ data\ batch\ (window)\ W$:

$S = CS \cup W$

$CS = choose\ M\ coresets\ from\ S\ using\ K-means++ initialization.$

For $each\ cs \in CS$:

$D_{cs} = \{ \mathrm{d} \mid \mathrm{d} \in D \wedge \forall cs' \neq cs \ . \ \| \mathrm{d} - \mathrm{cs} \| \leq \| \mathrm{d} - \mathrm{cs}' \| \}$

$cs = \frac{\sum_{d \in D_{cs}} d}{|D_{cs}|}$

$w_{cs} = |D_{cs}|$

$\widehat{gc}_{cs} = < \widehat{gc}_{cs}^{1}, \widehat{gc}_{cs}^{2}, \dots, \widehat{gc}_{cs}^{G} >, \ gc_{cs}^{j} = |\{ d \mid d \in D_{cs} \wedge label(d) = j \}|$

Upon $clustering\ request\ with\ parameters\ t, r, k$:

Clusters = $WAKmeans(t, r, k, CS)$

Output: A set of clusters $\{\mathrm{c}_1, \mathrm{c}_2, \dots, \mathrm{c}_k\}$ ***//k is arbitrary***

الگوریتم۳-۱. الگوریتم خلاصه‌سازی داده‌ها

به طور کلی این الگوریتم در ابتدا با استفاده از پنجره‌گذاری، جریان‌داده‌ی ورودی را دریافت می‌کند. سپس با دریافت هر پنجره، به جای ذخیره‌سازی کل داده‌ها، تنها خلاصه‌ای از داده‌ها یا همان داده‌های هسته که تعداد آن به مراتب کمتر از کل داده‌ها می‌باشد را حفظ می‌نماید. به بیان دیگر، داده‌های هسته خلاصه‌ای از داده‌های اصلی هستند که در طول پردازش جریان‌داده شکل گرفته و در مفاهیم موجود در جریان‌داده مورد استفاده قرار می‌گیرند. در هر زمان دلخواه می‌توان داده‌های هسته را به عنوان ورودی به الگوریتم WAKmeans اعمال کرد که این الگوریتم فرایند خوشه‌بندی نهایی را بر روی داده‌های هسته انجام می‌دهد. در واقع باید به این نکته اشاره شود که الگوریتم WAKmeans، نسخه برخط شده الگوریتم خوشه‌بندی تطبیقی است که در بخش‌های بعدی به صورت کامل‌تر به

شرح آن خواهیم پرداخت. ذکر این نکته ضروری است که پارامتر S به کارگرفته شده در این الگوریتم (پارامتر S همان پارامتر موجود در فرایند شکل ۳-۱ است)، ناشی از اجتماع خلاصه‌داده‌ها با داده‌هایی است که درون یک پنجره قرار گرفته‌اند. داده‌هایی که در مجموعه S وجود دارند، به نقاطی که به عنوان خلاصه انتخاب شده‌اند، اختصاص پیدا می‌کند. این عمل به معنای سنجیدن میزان نزدیکی داده‌های قرار گرفته در S با خلاصه‌داده‌ها است که با بهره‌گیری از فاصله اقلیدسی به انجام می‌رسد. پس از انجام عملیات مربوط به سنجش نزدیکی داده‌های قرار گرفته در S، باید به محاسبه وزن هرکدام از داده‌های هسته بپردازیم.

به منظور محاسبه وزن باید به این نکته اشاره شود که فرایند تخصیص نمونه‌های وارد شده به این صورت انجام می‌گیرد که از میان داده‌هایی که درون پنجره قرار دارند، با استفاده از روش انتخاب مراکز اولیه‌ی KMeans++ به تعداد مورد نیاز، نقاطی به عنوان خلاصه‌ی اولیه با وزن یک (واحد) انتخاب می‌شوند. به تدریج با ورود داده‌های بیشتر به پنجره‌ای که در نظر گرفته‌ایم، این خلاصه‌ی اولیه به مجموعه نقاطی که درون پنجره جای گرفته‌اند افزوده شده و از میان آن‌ها مجددا فرایند انتخاب توسط روش KMeans++ صورت می‌گیرد. این فرایند متناسب با تعداد دفعاتی که برای اعمال پنجره‌گذاری درنظر گرفته شده ادامه می‌یابد. به عبارت دیگر هر زمان که کاربر بخواهد نمونه‌های جریان‌داده تا لحظه‌ی جاری را مورد تحلیل قرار دهد، این فرایند می‌تواند به اتمام برسد. ترتیب کلی انجام عملیات خلاصه‌سازی داده‌ها به ازای هر بار دریافت پنجره‌ای از جریان‌داده، مطابق مراحل زیر دنبال می‌شود:

۱- مجموعه داده‌های هسته جاری به داده‌های دریافت شده از جریان‌داده اضافه می‌شوند. باید درنظر داشت که در بار اول اجرای این الگوریتم، مجموعه‌ی نقاط هسته تهی است.

۲- داده‌های هسته جدید با روش انتخاب نقاط اولیه در الگوریتم KMeans++، از مجموعه‌داده گام ۱ انتخاب می‌شوند.

۳- تمامی نقاط مجموعه‌داده‌ای که در گام ۱ ایجاد شده است، بر اساس نزدیک‌ترین فاصله به داده‌های هسته اختصاص پیدا کرده و وزن مربوط به داده‌های هسته به‌روز می‌شود. بردار گروه-تعداد نیز متناسب با گروه داده‌های اختصاص یافته به خلاصه‌داده‌ها مقداردهی می‌شود. همچنین باید به این نکته اشاره کرد که مطابق با فرمول ۳-۱، بُعد مربوط به داده‌های هسته نیز متناسب با وزن آن داده‌ها مقدار جدیدی می‌گیرد. در این مرحله، فاصله‌ی میان داده‌هایی که داخل پنجره قرار گرفته‌اند با نقاطی که به عنوان خلاصه انتخاب شده‌اند، برحسب فاصله اقلیدسی سنجیده می‌شود.

لازم به یادآوری است که ما در الگوریتم پیشنهادی از مدل پنجره‌ی LandMark برای پردازش داده‌ها استفاده کرده‌ایم. اندازه‌ای که برای این پنجره در نظر گرفته شد به صورت تجربی و آزمون و خطا به دست آمده است. به بیان دیگر برای بهتر شکل گرفتن خلاصه‌داده‌ها در این پژوهش نسبت‌های مختلفی از اندازه‌ی پنجره مورد بررسی و آزمون قرار گرفت. همچنین باید درنظر داشته باشیم تعداد داده‌های هسته نیز به صورت تجربی تعیین شده است. مقداری که برای اندازه‌ی پنجره در نظر گرفته شده متناسب با اندازه‌ی جریان‌داده است و از سوی دیگر تعدادی که برای خلاصه‌داده‌ها نیز لحاظ شده متناسب با اندازه‌ی پنجره می‌باشد.

از میان آزمون و خطاهای انجام شده می‌توان به نکاتی از جمله کم یا زیاد بودن اندازه‌ی پنجره و تعداد خلاصه‌داده‌ها و تاثیر آن بر فرایند خلاصه‌سازی و خوشه‌بندی نهایی اشاره کرد. برای مثال اگر تعداد خلاصه‌داده‌های انتخابی خیلی کم باشد، این امر موجب شکل‌گیری خلاصه‌های نامناسب و در نتیجه بروز مشکل در خوشه‌بندی نهایی می‌شود و یا اگر اندازه‌ی پنجره بسیار بزرگ انتخاب شود، می‌تواند باعث بالا رفتن پیچیدگی زمانی و حافظه‌ی مصرفی گردد. بیش از حد بودن تعداد خلاصه‌داده‌ها نیز به نوبه‌ی خود می‌تواند مشکلاتی نظیر موارد فوق را ایجاد کند. بنابراین لازم است که به این پارامترها مقادیر مناسبی اختصاص یابد تا بتوانیم انتظار شکل‌گیری خلاصه‌های مناسب و یک خوشه‌بندی دقیق را داشته باشیم.

همچنین با توجه به این‌که نمی‌توان انتهایی برای جریان‌داده در نظر گرفت، لذا در ابتدا چنین برداشت می‌شود که به علت نداشتن کل اندازه‌ی جریان‌داده‌ی مورد بررسی، نمی‌توانیم اندازه‌ی مناسبی برای پنجره تعیین کنیم. بنابراین به منظور مقابله با این چالش، یک مقدار ثابت برای اندازه‌ی جریان‌داده درنظر گرفته‌ایم. در ادامه‌ی این بخش به بررسی فاز دوم یا همان خوشه‌بندی تطبیقی وزن‌دار پرداخته‌ایم.

۳-۱-۲ خوشه‌بندی تطبیقی وزن‌دار (فاز دوم)

داده‌های خلاصه شده مطابق با شکل ۳-۱ وارد فاز دوم این فرایند شده و بر روی آن عملیات خوشه‌بندی تطبیقی وزن‌دار به انجام می‌رسد. پس از انجام این فرایند، خوشه‌های نهایی شکل می‌گیرند که تعداد این خوشه‌ها را می‌توان بر اساس مقتضیات کاربردی که مدنظر داریم، تعیین کنیم. روال انجام عملیات خوشه‌بندی در فاز دوم مطابق با الگوریتم شکل ۳-۲ است. خوشه‌های تولید شده ناشی از این الگوریتم همان اهداف خوشه‌بندی تطبیقی نسخه‌ی اولیه را دنبال می‌کنند که در واقع کشف روابط گروهی میان داده‌ها است. به بیان دیگر مفهوم (r, t)- accordant در این خوشه‌بندی رعایت می‌شود. دلیل این امر آن است که در ایجاد خلاصه‌ی داده‌ها به تعداد داده‌های موجود از هر گروه توجه شده است تا درصدی از داده‌ها (t) که قرار است در انتهای خوشه‌بندی در تعدادی از خوشه‌ها قرار بگیرد، به درستی برقرار شود. همانطور که در فصل دوم در ارتباط با این مفهوم به طور مفصل شرح داده شد، می‌توان در نظر داشت که یک خوشه‌بندی، (r, t)- accordant است، اگر داده‌های حداقل r گروه، شرط t-accordant را داشته باشند.

پارامترهای مورد استفاده به منظور درک بهتر الگوریتم ارائه شده (WAKmeans) در جدول ۳-۳ آورده شده است. در این الگوریتم روش خوشه‌بندی تطبیقی پایه تغییر داده شده است و الگوریتم مذکور، روی مجموعه‌ای از داده‌های هسته و یا همان خلاصه‌ی داده‌ها به اجرا در می‌آید.

جدول ۳-۳. پارامترهای موجود در الگوریتم WAKmeans

پارامترهای الگوریتم	مفهوم
#iterations	حداکثر تعداد دورهای خوشه‌بندی تطبیقی
R	پارامتر r در خوشه‌بندی تطبیقی
T	پارامتر t در خوشه‌بندی تطبیقی
K	تعداد خوشه‌ها

به منظور توضیح دقیق‌تر و همچنین بیان جزئیات مربوط به الگوریتم ۳-۲، قسمت‌هایی در شبه‌کد با استفاده از اعداد انگلیسی مشخص شده است که در ادامه توضیح اجمالی در مورد نحوه‌ی عملکرد کلی این الگوریتم داده می‌شود و سپس تمامی قسمت‌هایی که شماره‌گذاری شده‌اند، به طور کامل مورد بررسی قرار خواهند گرفت.

الگوریتم ۳-۲:Weighted Accordant Kmeans (WAKmeans)

Input: $t, r, k,$ coresets

1:cluster_centers $=$ initial clusters centers

2: $groups\_count = initial\ groups\ count\ (using\ \ group\_count)$

repeat until convergence:

  For each $coreset_i \in$ Coresets :

    For each $cluster_j \in$ cluster_centers :

      3: $p_{ij} = dist(coreset_i, cluster_j) - \min_\gamma(coreset_i, cluster_\gamma)$

    End For

  End For

  4: $Calculate\ minimum\ weight\ \ in\ each\ group\ for\ coresets$.

  5: $Extract\ group\_portion\ \ vectors\ for\ all\ coresets.$

6:Calculate a measure for coresets sorting in group_center.

7:Sort group_center *containing* coresets using alpha measure calculated in step 6 in descending priority.

8: *Calculate the number of cores that satisfy* $t-$ *accordant condition by using group fraction calculated in step* 4.

9: *Calculate the sum of alpha measure for selected cores in step* 8 *and consider this sum as group_center score. Then sort group* $-$ *centers by them scores in descending priority.*

10: *Choose the r group_center from sorted list. In this group_centers assign cores that selected in step* 8 *to corresponding centers.*

11: *Assign remaining cores to the closest cluster center.*

12: Calculate new cluster centers.

Output: A set of clusters $\{c_1, c_2, \ldots, c_k\}$

الگوریتم۳-۲. الگوریتم خوشه‌بندی تطبیقی وزن‌دار

روال کلی این الگوریتم به این صورت است که ابتدا داده‌های خلاصه شده (coresets) به عنوان ورودی الگوریتم به آن داده می‌شود. باید در نظر داشته باشیم که خلاصه‌داده‌ها به صورت وزن‌دار هستند. با توجه به این‌که روش معرفی شده همانند روش خوشه‌بندی تطبیقی بر پایه‌ی K-Means است، لذا برای انجام فرایند خوشه‌بندی نیاز به تعدادی مرکز اولیه است. انتخاب این مراکز اولیه در نسخه ابتدایی و پایه این روش به صورت کاملا تصادفی صورت می‌گیرد. در حالی که در روش پیشنهادی، انتخاب مراکز اولیه به کمک KMeans++ انجام می‌پذیرد. در واقع می‌توان این عمل یعنی تغییر در انتخاب مراکز اولیه را به عنوان یک بهبود بر نسخه اولیه درنظر گرفت. اجرای این الگوریتم بر روی داده‌های هسته، نتیجه‌ای نزدیک به اجرا بر روی داده‌های خام را به همراه دارد. برای درک بهتر مفاهیم گفته شده در مطالب فوق، مراحل موجود در الگوریتم WAKmeans، به صورت گام به گام در ادامه بیان می‌شود.

۱- در گام اول مراکز اولیه‌ی خوشه‌ها با استفاده از روش انتخاب KMeans++ از میان

خلاصه‌داده‌ها مقداردهی اولیه می‌شوند.

۲- در این گام با استفاده از بردار گروه-تعداد مربوط به خلاصه‌داده‌های شکل گرفته، تعداد عناصر هر گروه محاسبه می‌شود (groups_count). این تعداد برای آن‌که بتواند شرط اصلی مربوط به خوشه‌بندی تطبیقی که همان شرط (r, t)- accordant است را برآورده نماید، به کار برده می‌شود. ارضا شدن این شرط بدان معنا است که می‌توان تحلیل روابط گروهی بر روی خوشه‌های شکل گرفته را انتظار داشت که این امر از جمله مهم‌ترین اهداف اصلی انجام این پایان‌نامه نیز بوده است.

۳- در این گام ماتریس هزینه‌ی مربوط به خلاصه‌داده‌ها (داده‌های هسته) و مراکز خوشه‌ها محاسبه می‌شود.

۴- در این قسمت، حداقل تعدادی که هر گروه باید داشته باشد (Grpfractionpoints) محاسبه می‌شود. این کار نیز برای ارضای شرط خوشه‌بندی تطبیقی به انجام می‌رسد. لازم به ذکر است که در محاسبه‌ی این حداقل وزن از groups_count که در گام ۲ به دست آمده استفاده می‌شود. چگونگی این محاسبه در فرمول ۳-۲ مشاهده می‌شود.

$$\text{Grpfractionpoints}_i = \lceil t \times \text{groups\_count}_i \rceil \qquad (۲-۳)$$

در این فرمول، $\text{Grpfractionpoints}_i$ نشان‌دهنده‌ی حداقل تعداد مورد نیاز برای گروه i اُم، $groups\_count_i$ تعداد داده‌های موجود در گروه i اُم و پارامتر t همان پارامتر موجود در خوشه‌بندی تطبیقی است.

۵- در این بخش برداری با عنوان group_portion برای هر خلاصه‌داده به دست می‌آید. بردار مورد اشاره از تقسیم تعداد داده‌های موجود از هر گروه در یک خلاصه‌داده، بر مجموع تعداد داده‌هایی که از کل گروه‌ها در همان خلاصه‌داده قرار دارند به دست می‌آید. به منظور روشن شدن موضوع، مثال ۳-۲ در ادامه آورده شده است.

مثال ۳-۲. فرض کنید که بردار گروه-تعداد مربوط به یک داده‌ی هسته به شکل زیر باشد (جدول ۳-۴). بنابراین بردار group_portion آن مطابق آنچه که توضیح داده شد، برای داده‌های موجود از گروه ۱ در این خلاصه داده برابر خواهد بود با $\frac{10}{10+60+0} = \frac{1}{7}$.

جدول ۳-۴. بردار گروه-تعداد برای مثال ۳-۲

گروه<۳>	گروه<۲>	گروه<۱>
۰	۶۰	۱۰

۶- در این بخش یک معیار برای مرتب‌سازی داده‌های هسته در زوج‌های گروه-مرکز که در خوشه‌بندی تطبیقی پایه به آن اشاره شد (بخش ۲-۷) تعریف شده است. این معیار با عنوان alpha از فرمول ۳-۳ محاسبه می‌شود.

$$alpha_{core,center,group} = \frac{group\_portion_{core,group} \times group\_count_{core,group}}{1+penalty_{core,center}} \qquad (۳-۳)$$

در این فرمول، پارامترهای group_count و group_portion در موارد قبلی توضیح داده شده‌اند. حال می‌توان در مورد پارامتر $Penalty$ نیز به این نکته اشاره کرد که همان متغیر P در شبه‌کد مربوط به الگوریتم خوشه‌بندی تطبیقی وزن‌دار است (الگوریتم ۳-۲). همچنین در مورد اندیس‌های core، center و group باید گفت که به ترتیب نشان‌دهنده‌ی خلاصه‌داده (ساده شده‌ی coreset)، مرکز و گروه هستند. این فرمول با هدف در نظر گرفتن ارزش واقعی هر خلاصه‌داده با توجه به تعداد داده‌هایی که از هر گروه در آن قرار گرفته‌اند، طراحی شده است. فرمول ۳-۳ ارزش مربوط به خلاصه‌داده‌ها را چه در حالتی که تنها از یک گروه به آن‌ها داده تعلق گرفته باشد (به اصطلاح خالص باشد) و چه در حالتی که از چند گروه به آن داده اختصاص پیدا کرده باشد (به اصطلاح ناخالص باشد) به‌خوبی می‌تواند محاسبه کند. الگوریتم خوشه‌بندی تطبیقی وزن‌دار پیشنهادی در حالتی که داده‌های هسته خالص باشند نسبت به حالتی که داده‌های هسته خلوص کمتری دارند، بهتر می‌تواند

فرایند خوشه‌بندی را به انجام برساند.

۷- در این مرحله، خلاصه‌داده‌های موجود در هر گروه-مرکز، بر اساس معیار آلفا (فرمول ۳-۳) به صورت نزولی مرتب‌سازی می‌شود.

۸- در این مرحله تعدادی از خلاصه‌داده‌ها که منجر به ارضای شرط t-accordant در یک گروه-مرکز می‌شود با استفاده از مقدار Grpfractionpoints که در گام ۴ به دست آمد انتخاب می‌شوند. برای ارضای این شرط، مجموع تعداد عناصر مربوط به گروه یک گروه-مرکز خاص، بایستی بزرگتر یا مساوی مقدار Grpfractionpoints متناسب با آن گروه خاص باشد. برای درک بهتر موضوع به مثال ۳-۳ توجه کنید.

مثال ۳-۳. فرض کنیم بردار گروه-تعداد مربوط به تعدادی از داده‌های هسته (جدول ۳-۵) به شکل زیر باشد (این داده‌های هسته مرتب شده هستند). آنگاه برای حالتی که به طور فرض Grpfractionpoints برای گروه ۱، مقدار ۲۴ محاسبه شده باشد، داده‌های هسته‌ی ۱، ۲ و ۳ این Grpfractionpoints را ارضا می‌کنند. مجموع تعداد داده‌های موجود از گروه ۱ در این داده‌های هسته بزرگتر یا مساوی با مقدار ۲۴ است ($24 \ll 10 + 5 + 10$).

جدول ۳-۵. بردار گروه-تعداد مربوط به تعدادی از داده‌های هسته

داده‌های هسته	گروه<۱>	گروه<۲>	گروه<۳>
Coreset 1	۱۰	۰	۲
Coreset 2	۵	۳	۱
Coreset 3	۱۰	۰	۰
Coreset 4	۳	۰	۰

۹- سپس مقدار معیار آلفا برای خلاصه‌داده‌های انتخاب شده در گام ۸ با یکدیگر جمع می‌شود و این عدد به عنوان امتیاز گروه-مرکز در نظر گرفته می‌شود. گروه-مرکزهای موجود، بر اساس مقادیر امتیاز آن‌ها به صورت نزولی مرتب می‌شوند.

۱۰- در این مرحله، r عدد از جفت‌های گروه-مرکزی که مرتب شده هستند، به ترتیب انتخاب می‌شوند. سپس همان خلاصه‌داده‌هایی که در گام ۸ برای هر گروه-مرکز انتخاب شده‌اند به مرکز مربوطه اختصاص داده می‌شوند. منظور از مرکز مربوطه همان مرکزی است که زوجِ گروه-مرکز دارند.

۱۱- در این مرحله، خلاصه‌داده‌هایی که باقی‌مانده‌اند به کمک روش K-Means به نزدیکترین مراکز خوشه‌ی موجود اختصاص داده می‌شوند.

۱۲- در مرحله‌ی انتهایی، مراکز خوشه‌ی ایجاد شده به صورت وزن‌دار به‌روزرسانی می‌شوند. چرا که همانطور که قبلا به آن اشاره شده است، خلاصه‌داده‌ها وزن‌دار هستند.

۳-۲ جمع‌بندی

در این فصل به معرفی و توضیح روش پیشنهادی خود پرداختیم. اشاره کردیم که روش ما در این پایان‌نامه بر پایه‌ی روش خوشه‌بندی تطبیقی بوده است که تغییراتی در آن با هدف برخط نمودن این روش ایجاد شد. مفاهیمی از قبیل خلاصه‌سازی داده‌ها بیان و مفهوم پارامترهایی از قبیل وزن، بُعد و بردار گروه-تعداد برای هر داده تبیین گردید. پارامترهای استفاده شده در الگوریتم‌های این بخش در قالب جداولی برای روشن‌تر شدن مفهوم الگوریتم پایه و روش پیشنهادی ارئه گردید. در این بخش همچنین فرایندی برای تشریح بهتر فازهای پیاده‌سازی این روش رسم شد که به صورت دقیق و مرحله به مرحله خلاصه‌سازی داده‌ها و البته انجام عملیات خوشه‌بندی تطبیقی وزن‌دار نهایی را بیان نموده است.

فصل ۴ : ارزیابی روش پیشنهادی و نتایج

۴-۱ تنظیمات و راه‌اندازی سیستم

در این فصل به ارزیابی روش پیشنهادی و بیان نتایج حاصل از آن می‌پردازیم. در ابتدا به معرفی مجموعه‌داده‌های به کار گرفته شده و سپس به بیان جزئیات مربوط به فرایند پیاده‌سازی اقدام می‌کنیم. در ادامه مراحل راه‌اندازی سیستم و تنظیمات مورد نیاز برای اجرای الگوریتم خوشه‌بندی تطبیقی جریان‌داده را شرح می‌دهیم. در انتهای فصل نیز به انجام آزمایش‌ها و ارزیابی نتایج حاصل از آن‌ها و مقایسه‌ی راهکار پیشنهادی با تعدادی از روش‌های معمول در این زمینه می‌پردازیم. لازم به ذکر است که به منظور پیاده‌سازی راهکار پیشنهادی خود، از زبان برنامه‌نویسی جاوا و محیط IntelliJ IDEA استفاده کردیم.

۴-۲ مجموعه داده

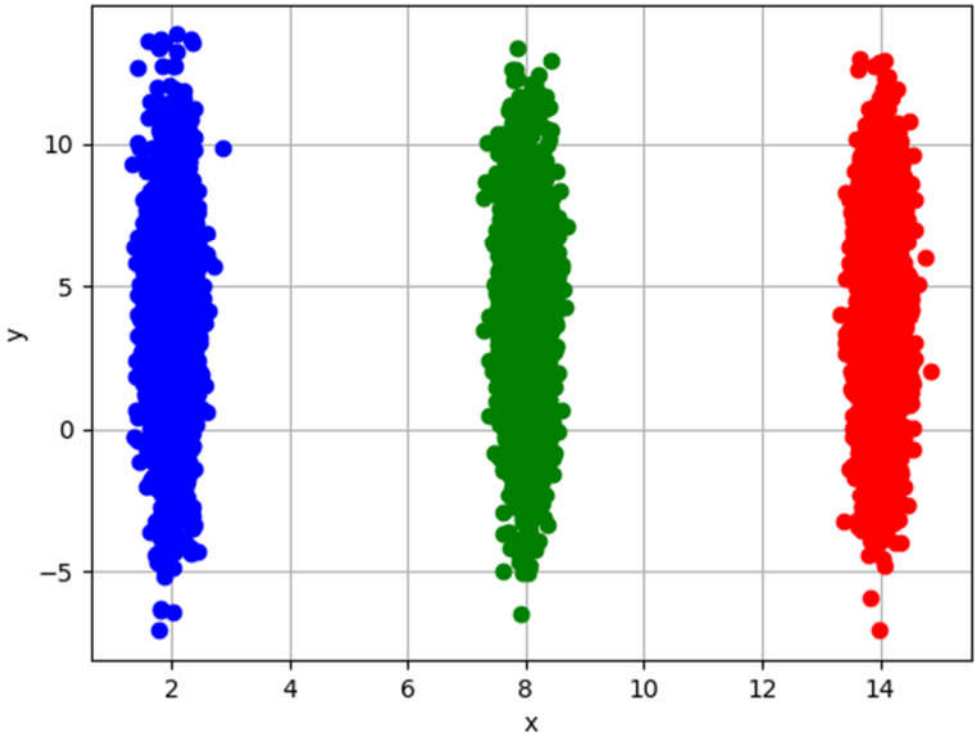
به منظور ارزیابی روش پیشنهادی از دو نوع مجموعه‌داده استفاده شده است. نوع اول، مجموعه‌داده‌ی ساختگی است که هدف اصلی از به کارگیری آن، بررسی مفاهیم پایه‌ای موجود در خوشه‌بندی می‌باشد. به عبارت دیگر به منظور کسب اطمینان از فرایند خوشه‌بندی بر روی داده‌هایی که به طور ذاتی خوشه‌پذیر[۱] هستند، نسبت به استفاده از این مجموعه‌داده اقدام کردیم. نوع دوم، مجموعه‌داده‌های واقعی است که با هدف تعمیم‌پذیری روش پیشنهادی و ارزیابی سیستم در شرایطی نزدیک به دنیای حقیقی به کار گرفته شده است.

۴-۲-۱ مجموعه داده‌ی ساختگی

ما به منظور ساخت این مجموعه‌داده از یک توزیع نرمال در ۳ کلاس مختلف با میانگین متفاوت و واریانس یکسان اقدام کردیم. در این مجموعه‌داده ۱۰۰۰۰ نمونه وجود دارد. هر نمونه دارای دو بُعد

[۱] Clusterable

است. لازم به ذکر است که ما برای انتخاب مراکز مربوط به خوشه‌ها، بازه‌ی مناسبی را در نظر گرفتیم به گونه‌ای که خوشه‌ها در فاصله‌ی معقولی نسبت به یکدیگر قرار گرفته باشند. در شکل ۴-۱ مجموعه‌داده‌ی ساختگی شرح داده شده، مشاهده می‌شود.



شکل ۴-۱. نمایش مجموعه داده‌ی ساختگی

۴-۲-۲ مجموعه داده‌های واقعی

همان‌طور که در ابتدای این بخش نیز اشاره شد، ما به منظور سنجش راهکار پیشنهادی خود در کاربردهای دنیای حقیقی و بررسی تعمیم‌پذیری سیستم، از تعدادی مجموعه‌داده‌ی واقعی استاندارد موجود در UCI که در ادامه توضیح داده می‌شوند، استفاده کردیم [۳۱].

مجموعه‌داده‌ی نخست با نام Pendigits مورد استفاده قرار گرفت. این مجموعه‌داده شامل ۱۰۹۹۲ نمونه برچسب‌دار است. در این مجموعه‌داده، هر نمونه ۱۶ ویژگی با ۱۰ کلاس دارد. بنابراین از دیدگاه مدل‌سازی، مجموعه‌داده‌ی مذکور شامل ۱۰ گروه است. Pendigits در واقع یک مجموعه‌داده‌ی رقمی

است که از ۴۴ نویسنده و با ۲۵۰ نمونه گردآوری شده است.

مجموعه‌داده‌ی دوم magic04 نام دارد. این مجموعه‌داده دارای ۱۰ ویژگی با دو کلاس است که تعداد ۱۹۰۲۰ نمونه را شامل می‌شود. این مجموعه‌داده با هدف شبیه‌سازی ذرات گاما با سطح انرژی بالا به کمک روش تصویربرداری یک تلسکوپ در اطراف جو زمین ایجاد شده است.

مجموعه‌داده‌ی سوم با نام Shuttle به کار گرفته شد. این مجموعه‌داده از دو بخش آموزش و آزمون تشکیل شده است که در مجموع ۵۷۷۱۰ نمونه دارد. هر نمونه شامل ۹ ویژگی با ۷ کلاس است. بنابراین، در این مجموعه‌داده از لحاظ مدل‌سازی ۷ گروه وجود دارد. لازم به ذکر است که مجموعه‌داده‌ی مذکور از جمله مجموعه‌داده‌های شناخته شده‌ی نجومی به حساب می‌آید.

۴-۳ پیاده‌سازی روش ارائه شده

ما در این بخش بعد از معرفی مجموعه‌داده‌های به کار گرفته شده، به بیان تنظیمات سیستم و همچنین نتایج به دست آمده از الگوریتم خوشه‌بندی تطبیقی جریان‌داده‌ای که در فصل سوم به طور کامل توضیح داده شد، می‌پردازیم.

به طور کلی، سیستم پس از دریافت مجموعه‌داده مورد نظر، فرایند خوشه‌بندی تطبیقی جریان‌داده را آغاز می‌کند. در مرحله‌ی نخست، بعد از مشخص کردن پارامترهای الگوریتم معرفی شده، فرایند خلاصه‌سازی داده‌ها انجام می‌شود و سپس در مرحله‌ی بعد خوشه‌بندی تطبیقی وزن‌دار بر روی خلاصه‌های به دست آمده از مرحله‌ی اول، به انجام می‌رسد. با این کار در نهایت خوشه‌های نهایی ناشی از جریان‌داده تا لحظه‌ی مورد نظر شکل می‌گیرد.

در جدول ۴-۱، مقادیر پارامترها برای اجرای الگوریتم خوشه‌بندی تطبیقی جریان‌داده‌ای آورده شده است. لازم به ذکر است که ما در انتخاب مقادیر پارامترهای اندازه‌ی پنجره و تعداد داده‌های هسته همانطور که در فصل سوم اشاره شد، متناسب با اندازه‌ی مجموعه‌داده‌ی مورد استفاده عمل می‌کنیم. به این صورت که اندازه‌ی پنجره برابر با یک دهم اندازه‌ی جریان‌داده و تعداد داده‌های هسته، یک دهم

اندازه‌ی پنجره انتخاب می‌شود. به عنوان مثال اگر اندازه‌ی مجموعه‌داده برابر ۱۰۰۰۰ نمونه باشد، اندازه‌ی پنجره و تعداد داده‌های هسته به ترتیب برابر ۱۰۰۰ و ۱۰۰ در نظر گرفته می‌شود.

جدول ۴-۱. مقادیر پارامترها در پیاده‌سازی

پارامتر	توضیح	مقدار
K	تعداد خوشه‌ها	۲ تا ۱۰
T	پارامتر t در خوشه‌بندی تطبیقی	۰,۷۵
R	پارامتر r در خوشه‌بندی تطبیقی	۱ تا ۱۰
#iterations	حداکثر تعداد دورهای خوشه‌بندی تطبیقی	۵۰
M	تعداد خلاصه داده‌ها (داده‌های هسته)	یک دهم اندازه‌ی پنجره
window size	اندازه پنجره دریافتی	یک دهم اندازه‌ی جریان‌داده

۴-۴ آزمایش‌ها و ارزیابی نتایج

با توجه به ماهیت داده‌ها در دنیای امروزی که به طور پیوسته در حال تولید هستند، این الزام به وجود آمده است که ساز و کارهای مناسبی برای مواجه با آن‌ها تدارک دیده شود. از این رو طراحی سیستم‌هایی که بتوانند به صورت برخط در مدیریت جریان‌داده‌ها به طور موثر واقع شوند، بیش از پیش احساس می‌شود. ما به منظور ارزیابی سیستم خود در چنین فضایی آزمایش‌های مختلفی را به کار بردیم. روند کلی کار به این صورت است که ما با تغییر پارامترها در آزمایش‌های مختلف اقدام می‌کنیم و سپس هر یک از آن‌ها را بر روی مجموعه‌داده‌های معرفی شده در بخش قبل، اجرا می‌نماییم.

معیار ارزیابی در این پژوهش، میانگین مجموع مربعات خطا است که در رابطه‌ی ۴-۱ مشاهده

می‌شود. همانطور که مشخص است این معیار از طریق جمع مربعات خطا محاسبه می‌شود که دلیل استفاده از آن نمایش کیفیت خوشه‌بندی انجام شده توسط الگوریتم پیشنهادی در مقایسه با روش‌های دیگر بوده است.

$$\frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{z}(C_i - Q_j)^2 \qquad (۱-۴)$$

در رابطه‌ی ۴-۱، n برابر با تعداد کل داده‌ها، k برابر تعداد نهایی خوشه‌های به دست آمده، C نماینده‌ی مرکز هر خوشه، z برابر تعداد نقاط قرار گرفته در خوشه‌ی i اُم و در نهایت Q داده‌ای است که به خوشه‌ی i اُم تعلق گرفته است.

در آزمایش نخست، با ثابت در نظر گرفتن پارامتر های r و t و تغییر در پارامتر k، نسبت به اجرای الگوریتم اقدام کرده که نتایج و توضیحات مربوط به آن برای هر مجموعه‌داده به صورت جداگانه آورده شده است. لازم به ذکر است که مقدار پارامتر k و همچنین اندازه‌ی پنجره و تعداد داده‌های هسته، متناسب با هر مجموعه‌داده و تعداد کلاس‌هایی که در آن مجموعه‌داده وجود دارد تعیین شده است. در جدول ۴-۲ مقادیر پارامترهای مذکور به همراه هر مجموعه‌داده آورده شده است.

جدول ۴-۲. مقادیر پارامتر k در آزمایش نخست

مجموعه‌داده	مقدار پارامتر k	اندازه‌ی پنجره	تعداد داده‌های هسته
مجموعه‌داده‌ی ساختگی	۲ تا ۶	۱۰۰۰	۱۰۰
Pendigits	۲ تا ۱۰	۱۰۰۰	۱۰۰
magic04	۲ تا ۶	۱۹۰۰	۱۹۰
Shuttle	۲ تا ۷	۴۰۰۰	۴۰۰

از دیگر مواردی که در این آزمایش می‌توان به آن اشاره کرد، مقایسه‌ی نتایج حاصل از روش خوشه‌بندی ارائه شده در این پایان‌نامه (SAKmeans) با دو روش K-Means و خوشه‌بندی تطبیقی

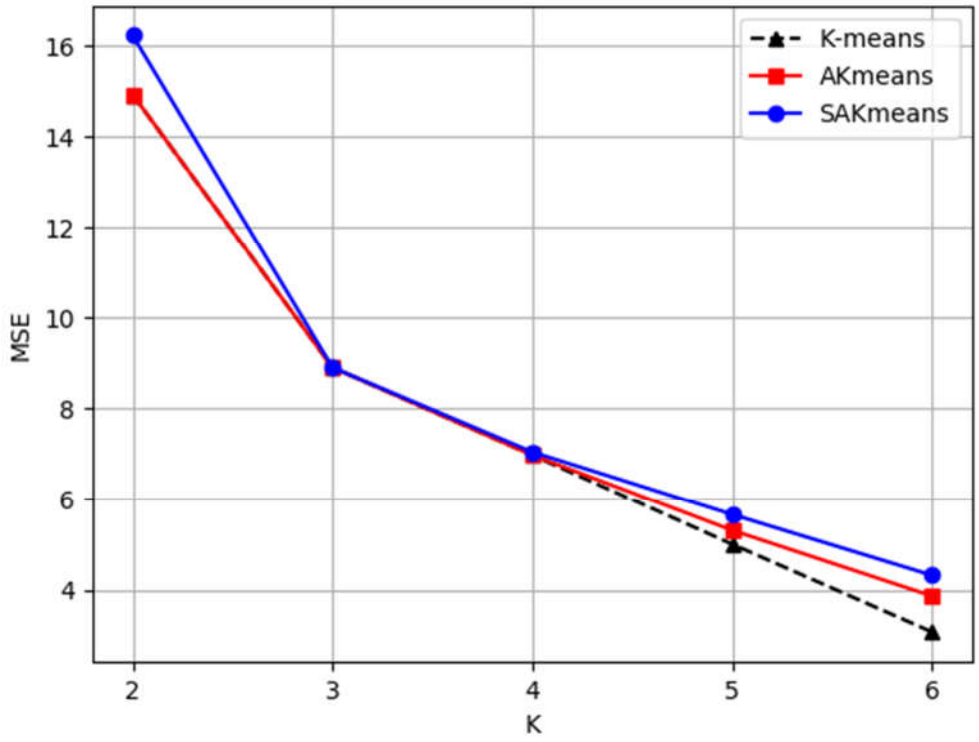
پایه (AKmeans) می‌باشد. به طور کلی روش K-Means و خوشه‌بندی تطبیقی نیاز به مراکز اولیه‌ی خوشه دارند. به همین دلیل برای این دو روش، مراکز اولیه یکسانی لحاظ شده است که به وسیله‌ی روش انتخاب مراکز در الگوریتم KMeans++ از میان مجموعه‌داده‌ی مورد استفاده، به دست آمده است. همچنین برای آن‌که بتوان روش SAKmeans ارائه شده را (که روی خلاصه‌ی داده‌ها فرایند خوشه‌بندی نهایی را انجام می‌دهد) با دو روش مورد اشاره که روی کل داده‌ها خوشه‌بندی می‌کنند مقایسه کرد، یک بار فرایند خوشه‌بندی تطبیقی پایه با استفاده از مراکزی که در روش SAKmeans برای تعداد خوشه‌ی مشخص تولید می‌شوند، بر روی تمامی داده‌ها صورت می‌گیرد. با این کار، معیار میانگین مجموع مربعات خطا به راحتی برای سه روش مورد اشاره محاسبه خواهد شد.

به منظور پیاده‌سازی ساختار جریان‌داده‌ای و یا به بیان دیگر برای آنکه مجموعه‌داده‌های موجود را به صورت جریان‌داده‌ای وارد الگوریتم SAKmeans نماییم، ابزار MOA[۱] را به کار گرفته‌ایم [۳۲]. MOA یک ابزار متن‌باز[۲] است که در آن مجموعه‌ای از الگوریتم‌های یادگیری ماشین (مانند روش‌های مختلف خوشه‌بندی، رده‌بندی و غیره) وجود دارد. از این ابزار با هدف کاوش و انجام ارزیابی‌های مختلف بر روی جریان‌داده‌ها استفاده می‌شود. در ادامه‌ی این فصل به بررسی خروجی‌های حاصل از آزمایش نخست می‌پردازیم.

در شکل ۴-۲ خروجی‌های به دست آمده بر روی مجموعه‌داده‌ی ساختگی قرار گرفته است. برای این مجموعه‌داده که قبلا در مورد آن توضیح داده شده، مقادیر پارامترهای t و r به ترتیب برابر ۰,۷۵ و ۲ در نظر گرفته شده است.

[۱] Massive Online Analysis

[۲] Open Source

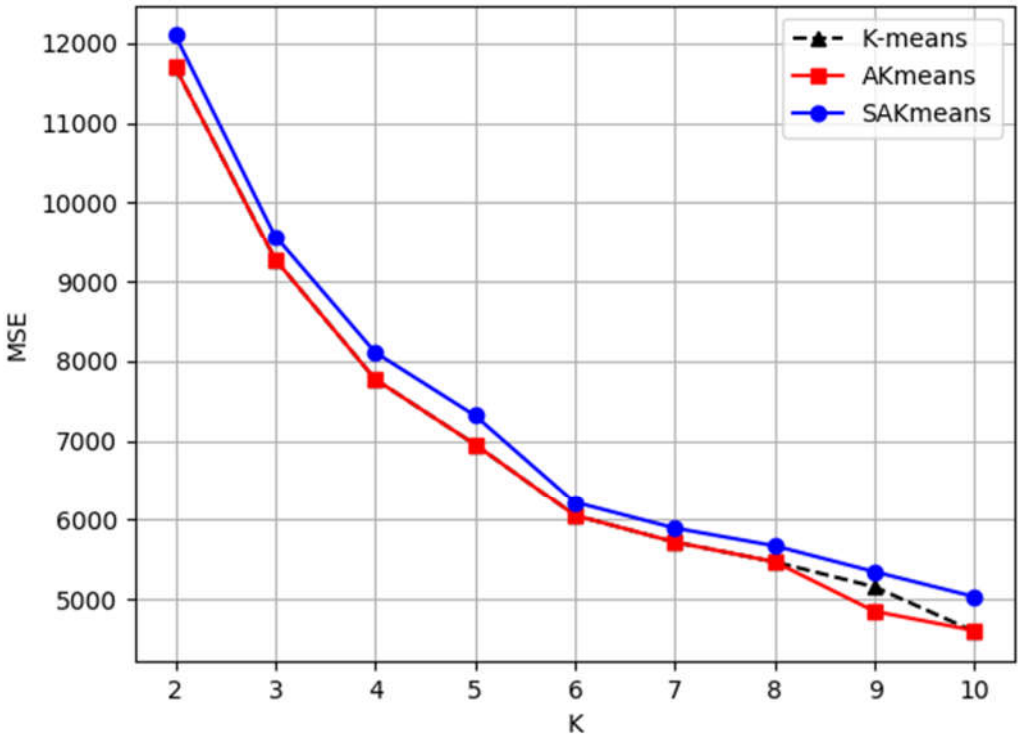


شکل ۴-۲. نمودار MSE مربوط به داده‌ی ساختگی

همانگونه که در شکل ۴-۲ مشاهده می‌شود، الگوریتم SAKmeans عملکرد قابل قبولی نسبت به روش K-Means و روش خوشه‌بندی تطبیقی داشته است. الگوریتم پیشنهادی، از تقریب خلاصه‌سازی داده‌ها برای کاهش پیچیدگی زمانی و ذخیره‌سازی استفاده کرده و با وجود این میزان کاهش داده، نتیجه‌ای بسیار نزدیک به دو الگوریتم دیگر به دست آورده است. اما از سوی دیگر میانگین مجموع مربعات خطای الگوریتم پیشنهادی مقداری بیشتر از دو روش دیگر است. دلیل اصلی این میزان بیشتر بودن، اعمال تقریبی از داده‌ها در اجرای فرایند کلی سیستم است. در واقع از آنجایی که روش پیشنهادی به صورت برخط اجرا می‌شود، بنابراین، کاهش پیچیدگی‌های زمانی و افزایش سرعت اجرا اهمیت ویژه‌ای را به همراه خواهد داشت. در نتیجه وجود چنین ویژگی‌ای در روش پیشنهادی، در مقایسه با اندکی بیشتر بودن مقدار میانگین مجموع مربعات خطا، چندان محسوس نبوده و وضعیت قابل قبولی را ایجاد می‌نماید. علت قرار گرفتن خروجی مربوط به روش خوشه‌بندی K-Means فراهم آوردن امکان مقایسه‌ی بهتر مقادیر میانگین مجموع مربعات خطای به دست آمده برای دو الگوریتم

دیگر (یعنی SAKmeans و AKmeans) با یک روش خوشه‌بندی استاندارد است.

در این قسمت به بررسی خروجی‌های به دست آمده از مجموعه‌داده‌ی Pendigits می‌پردازیم. در این مجموعه‌داده همانطور که قبل‌تر اشاره شد، ۱۰۹۹۲ نمونه وجود دارد که برای انجام این آزمایش از ۱۰۰۰۰ نمونه‌ی آن استفاده شده است. در شکل ۴-۳ نمودار مربوط به این خروجی به تصویر کشیده شده است. در این خروجی پارامتر r برابر با ۶ و پارامتر t برابر با ۰٫۷۵ لحاظ شده است. در این شکل نیز سه روش مذکور با یکدیگر مورد مقایسه قرار گرفته‌اند.

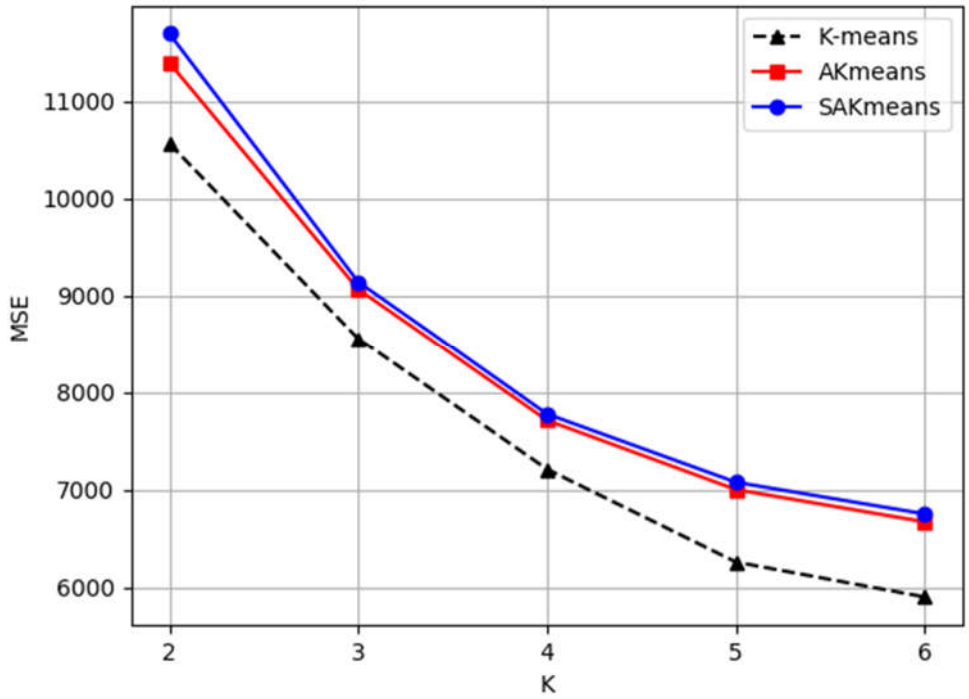


شکل ۴-۳. نمودار MSE مربوط به مجموعه‌داده‌ی Pendigits

مطابق با نمودار شکل ۴-۳، مقدار MSE سه الگوریتم نزدیک به یکدیگر هستند. بنابراین روش خوشه‌بندی جریان‌داده‌ای پیشنهاد شده با وجود اعمال تقریب و کاهش پیچیدگی‌های زمانی و ذخیره‌سازی، همچنان کارایی خود را حفظ کرده است.

در ادامه، خروجی‌های شکل گرفته بر روی مجموعه‌داده‌ی magic04 ارائه می‌شود. این مجموعه‌داده ۱۹۰۲۰ نمونه دارد که در آزمایش صورت گرفته از ۱۹۰۰۰ نمونه‌ی آن استفاده کرده‌ایم. ذکر این

نکته ضروری است که داده‌های این مجموعه‌داده به صورت مرتب شده گردآوری شده‌اند. یعنی ابتدا همه‌ی نمونه‌هایی که عضو کلاس اول هستند و در ادامه‌ی آن‌ها همه‌ی نمونه‌هایی که به کلاس دوم تعلق دارند، قرار گرفته‌اند. به همین دلیل، ما به منظور استفاده از این مجموعه‌داده در ابتدا ترتیب قرارگیری کل داده‌های موجود در آن را به هم ریختیم[1]. برای انجام آزمایش بر روی این مجموعه‌داده، پارامترهای t و r به ترتیب ۰,۷۵ و ۱ در نظر گرفته شد. در شکل ۴-۴ شاهد نتایج به دست آمده هستیم که سه روش خوشه‌بندی ذکر شده در این بخش را با یکدیگر مقایسه کرده است.



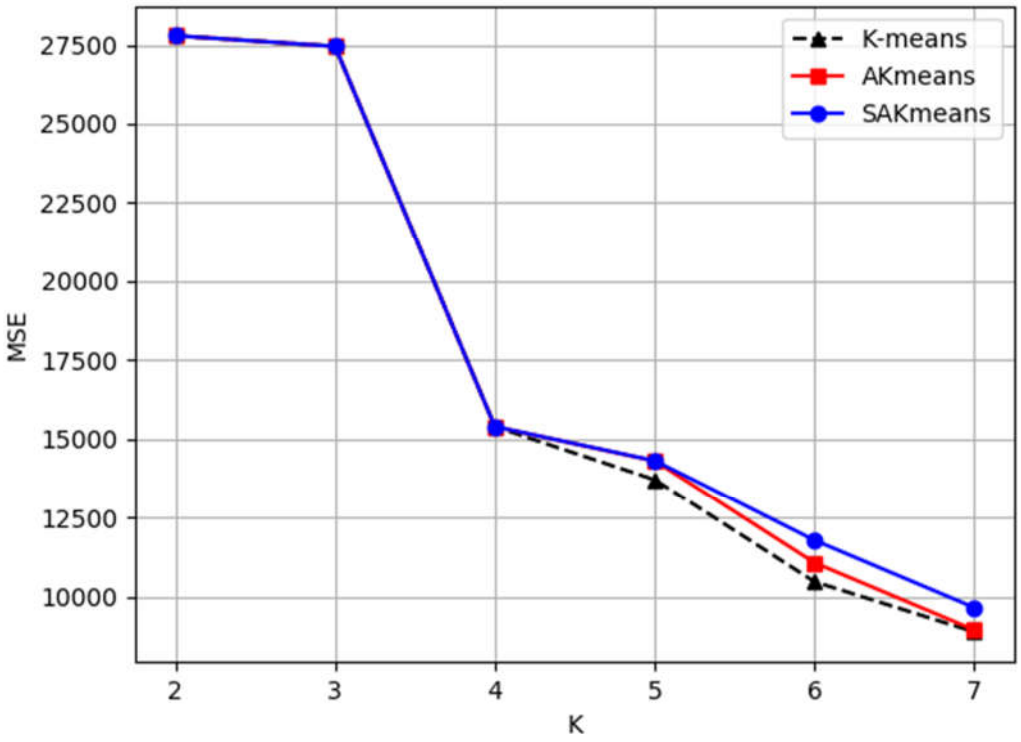
شکل ۴-۴. نمودار MSE مربوط به مجموعه‌داده‌ی magic04

همانطور که در شکل ۴-۴ دیده می‌شود، مقدار MSE حاصل شده از الگوریتم SAKmeans بسیار به روش خوشه‌بندی تطبیقی نسخه‌ی اولیه (AKmeans) نزدیک شده است. این امر نشان‌دهنده‌ی عملکرد مناسب روش ارائه شده است. چراکه این روش برخط است و مانند دو روش دیگر خوشه‌بندی مورد مقایسه، تمامی داده‌ها را از ابتدا در اختیار ندارد و تنها با استفاده از خلاصه‌ای از داده‌ها توانسته

[1] Shuffle

چنین خروجی‌هایی را ایجاد کند.

حال به بررسی خروجی‌های به دست آمده از مجموعه‌داده‌ی shuttle در شرایط انجام آزمایش نخست، می‌پردازیم. این مجموعه‌داده شامل داده‌هایی برای آزمون و داده‌هایی برای آموزش است. در این آزمایش از ۴۳۵۰۰ داده‌ی آموزش موجود برای این مجموعه‌داده، ۴۳۰۰۰ نمونه‌ی آن مورد استفاده قرار گرفته است. به منظور انجام آزمایش بر روی مجموعه‌داده‌ی shuttle، پارامترهای r و t به ترتیب برابر با ۳ و ۰٫۷۵ قرار داده شد. در شکل ۴-۵ نتایج به دست آمده از مقایسه‌ی سه روش خوشه‌بندی مذکور را شاهد هستیم.



شکل ۴-۵. نمودار MSE مربوط به مجموعه‌داده‌ی shuttle

شکل ۴-۵ نیز بیانگر عملکرد مناسب روش معرفی شده نسبت روش خوشه‌بندی تطبیقی پایه است. مطابق آنچه که برای خروجی‌های دیگر نیز توضیح داده شد، باید بر این نکته تاکید شود که با وجود اعمال تقریب و کاهش پیچیدگی‌های زمانی و ذخیره‌سازی، روش SAKmeans همچنان کارایی خود را حفظ کرده است.

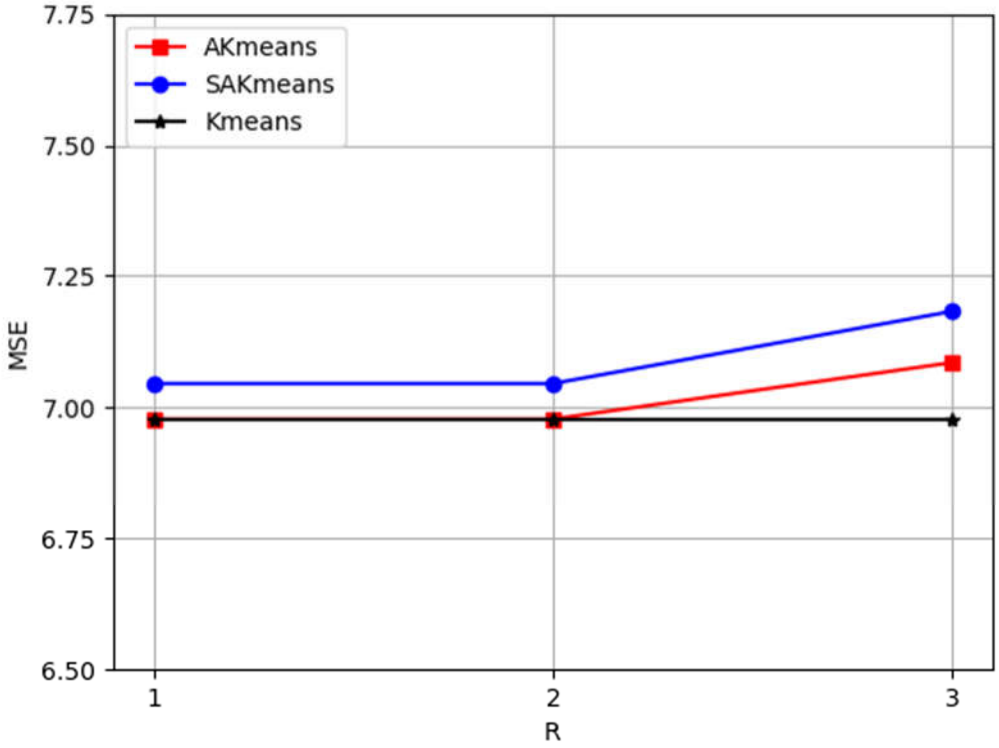
حال به معرفی شرایط موجود در آزمایش دوم می‌پردازیم. به طور کلی در این آزمایش به منظور مشاهده‌ی تاثیر پارامتر r بر روی خروجی‌های مربوط به الگوریتم معرفی شده در این پژوهش، پارامترهای k و t به صورت ثابت و پارامتر r، به طور متغییر در نظر گرفته شد. در ادامه نتایج و توضیحات مربوط به آن برای هر مجموعه‌داده به صورت جداگانه آورده شده است. لازم به ذکر است که مقدار پارامتر r و همچنین اندازه‌ی پنجره و تعداد داده‌های هسته، متناسب با هر مجموعه‌داده و تعداد کلاس‌هایی که در آن مجموعه‌داده وجود دارد لحاظ شده است. در جدول ۴-۳ مقادیر پارامترهای مذکور به همراه هر مجموعه‌داده به چشم می‌خورد.

جدول ۴-۳. مقادیر پارامتر r در آزمایش دوم

مجموعه‌داده	مقدار پارامتر r	اندازه‌ی پنجره	تعداد داده‌های هسته
مجموعه‌داده‌ی ساختگی	۱ تا ۳	۱۰۰۰	۱۰۰
Pendigits	۱ تا ۷	۱۰۰۰	۱۰۰
magic04	۱ تا ۲	۱۹۰۰	۱۹۰

در اجرای این آزمایش نیز شاهد نتایج حاصل از روش خوشه‌بندی ارائه شده در این پایان‌نامه (SAKmeans) و دو روش K-Means و خوشه‌بندی تطبیقی پایه (AKmeans) هستیم. البته با این تفاوت که در این آزمایش، به جای پارامتر k، پارامتر r دچار تغیر می‌شود. کیفیت خوشه‌بندی نیز در تمامی روش‌ها توسط معیار میانگین مجموع مربعات خطا اندازه‌گیری شده است. همچنین مراکز اولیه همانند آزمایش نخست برای دو روش K-Means و خوشه‌بندی تطبیقی پایه تعیین شده است.

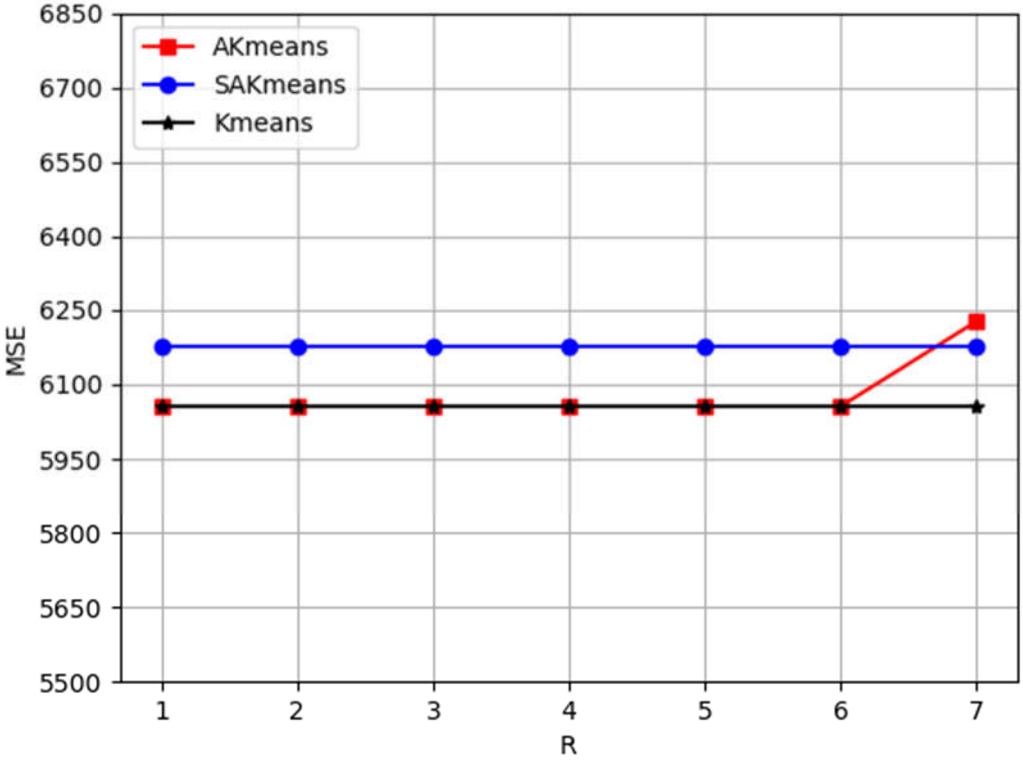
در شکل ۴-۶ خروجی‌های به دست آمده از مجموعه‌داده‌ی ساختگی به نمایش گذاشته شده است. برای این مجموعه‌داده نیز مقادیر پارامترهای t و k به ترتیب برابر ۰,۷۵ و ۴ در نظر گرفته شد.



شکل ۴-۶. نمودار MSE مربوط به مجموعه‌داده‌ی ساختگی با تغییر پارامتر r

بر اساس شکل ۴-۶، مشاهده می‌شود که روش SAKmeans بسیار نزدیک به خوشه‌بندی تطبیقی عمل کرده است. روش K-Means نیز منطقا باید دارای مقدار ثابتی باشد، چرا که مستقل از پارامتر r عمل می‌کند. دلیل قرارگیری خروجی K-Means فراهم آوردن امکان مقایسه‌ی بصری بهتر میان مقادیر MSE مربوط به دو الگوریتم خوشه‌بندی SAKmeans و AKmeans بوده است. لازم به ذکر است که در تمامی حالت‌های خوشه‌بندی جریان‌داده، شرط خوشه‌بندی تطبیقی نیز رعایت می‌شود که این شرط به صورت کامل در فصل‌های گذشته، بیان شده و مورد بررسی قرار گرفته است.

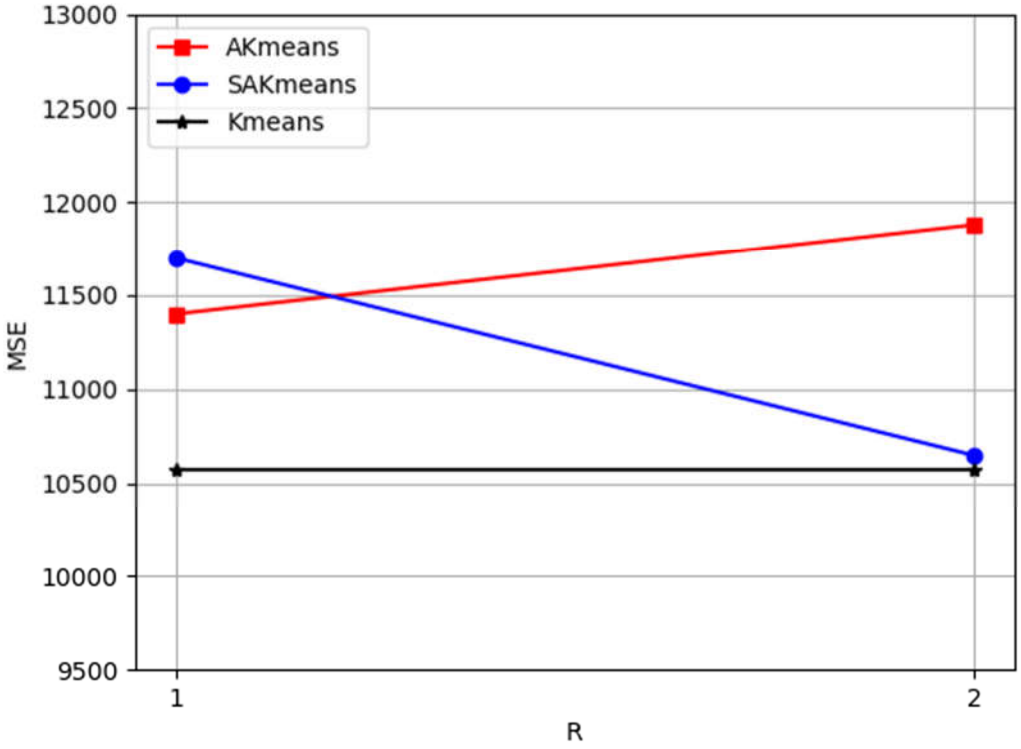
در شکل ۴-۷ نتایج به دست آمده بر روی مجموعه‌داده‌ی Pendigits را مشاهده می‌کنیم. به منظور انجام آزمایش دوم بر روی این مجموعه‌داده نیز از ۱۰۰۰۰ نمونه‌ی موجود در آن استفاده شده است. پارامترهای k و t نیز به ترتیب برابر با ۶ و ۰٫۷۵ در نظر گرفته شد.



شکل ۴-۷. نمودار MSE مربوط به مجموعه‌داده‌ی Pendigits با تغییر پارامتر r

همانگونه که در شکل ۴-۷ مشاهده می‌شود، روش SAKmeans در مقایسه با روش خوشه‌بندی تطبیقی پایه (AKmeans) برای بسیاری از مقادیر r (۱ تا ۶) نتایج بسیار نزدیکی داشته و حتی در جایی که پارامتر r برابر با ۷ شده، نتیجه‌ی بهتری نسبت به خوشه‌بندی تطبیقی از خود در خروجی نشان داده است. مجددا باید به این نکته اشاره شود که در تمامی حالت‌های خوشه‌بندی جریان‌داده، شرط خوشه‌بندی تطبیقی نیز رعایت می‌شود.

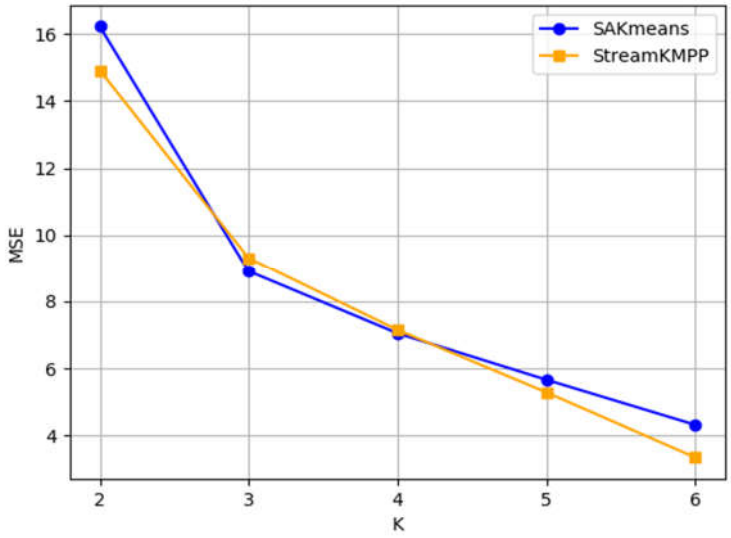
در ادامه، شاهد نتایج به دست آمده بر روی مجموعه‌داده‌ی magic04 در شکل ۴-۸ هستیم. به منظور انجام این آزمایش از ۱۹۰۰۰ نمونه‌ی این مجموعه‌داده استفاده شده است. همانطور که در قسمت‌های قبل توضیح داده شد، ترتیب نمونه‌های موجود در این مجموعه‌داده را قبل از استفاده به هم ریخته‌ایم. برای پارامترهای k و t نیز به ترتیب مقادیر ۲ و ۰,۷۵ لحاظ شده است.



شکل ۴-۸. نمودار MSE مربوط به مجموعه داده‌ی magic04 با تغییر پارامتر r

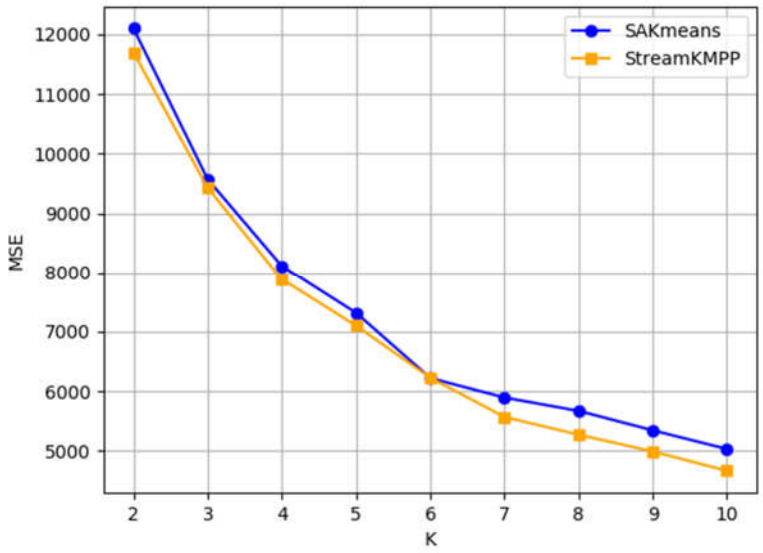
در شکل ۴-۸ مشاهده می‌کنیم که روش SAKmeans برای r=۱ دارای مقدار میانگین مجموع مربعات خطای بیشتری نسبت به روش خوشه‌بندی تطبیقی پایه و در حالت r=۲ مقدار میانگین مجموع مربعات خطای کمتری نسبت به روش خوشه‌بندی تطبیقی پایه داشته است. این‌گونه می‌توان گفت که در حالت r=۲، روش ارائه شده عملکرد بهتری نسبت روش خوشه‌بندی تطبیقی از خود نشان داده است.

در انتهای این فصل، روش معرفی شده‌ی خود را با خوشه‌بندی StreamKMPP [۱۶] که یک روش جریان‌داده‌ای است نیز مقایسه کرده‌ایم. نتایج مربوط به آن در دو شکل ۴-۹ و ۴-۱۰ به نمایش در آمده است. در شکل ۴-۹، خروجی‌های به دست آمده ناشی از مجموعه‌داده‌ی ساختگی هستند. اندازه‌ی پنجره و تعداد داده‌های هسته برای هر دو روش به ترتیب مقادیر ۱۰۰۰ و ۱۰۰ در نظر گرفته شده است.



شکل ۴-۹. نمودار MSE مربوط به دو روش جریان‌داده‌ای بر روی مجموعه‌داده‌ی ساختگی

در شکل ۴-۹ مشاهده می‌کنیم که روش خوشه‌بندی تطبیقی جریان‌داده‌ای معرفی شده، نتایج بسیار نزدیکی در مقایسه با روش خوشه‌بندی جریان‌داده‌ای StreamKMPP داشته و حتی برای مقادیر k=۳ و k=۴ میزان MSE کمتری از خود به نمایش گذاشته است.



شکل ۴-۱۰. نمودار MSE مربوط به دو روش جریان‌داده‌ای بر روی مجموعه‌داده‌ی Pendigits

در شکل ۴-۱۰ نیز واضح است که روش خوشه‌بندی تطبیقی جریان‌داده‌ای معرفی شده، مقادیر MSE بسیار نزدیکی نسبت به روش خوشه‌بندی جریان‌داده‌ای StreamKMPP دارد. اندازه‌ی پنجره و تعداد داده‌های هسته برای شکل‌گیری نتایج در این قسمت، به ترتیب دارای مقادیر ۱۹۰۰ و ۱۹۰ می‌باشد.

در این پایان‌نامه به این موضوع اشاره شده است که روش خوشه‌بندی تطبیقی جریان‌داده‌ای معرفی شده زمان اجرای کمتری (سرعت اجرای بیشتری) نسبت به روش خوشه‌بندی تطبیقی پایه دارد. این بیان به صورت کیفی است، به منظور نشان‌دادن این برتری در زمان اجرا، مقایسه‌ی کمّی زمان اجرای این دو روش در جدول ۴-۴ نشان داده شده است.

جدول ۴-۴. مقایسه‌ی زمان اجرای دو روش خوشه‌بندی تطبیقی جریان‌داده‌ای و خوشه‌بندی تطبیقی پایه

مجموعه‌داده	k	t	r	زمان اجرای خوشه‌بندی تطبیقی پایه (ثانیه)	زمان اجرای خوشه‌بندی تطبیقی جریان‌داده‌ای (ثانیه)	درصد کاهش زمان اجرا
مجموعه‌داده‌ی ساختگی	۳	۰,۷۵	۲	۵,۳۱۶	۳,۸۲۴	۲۸٪
	۵	۰,۵	۳	۴۲,۱۷۵	۵,۹۶۴	۸۵٪
Pendigits	۴	۰,۷۵	۲	۹,۰۱	۳,۴۸	۶۱٪
	۶	۰,۵	۴	۲۱,۵۹۱	۴,۰۵۸	۸۱٪
Magic04	۳	۰,۵	۲	۲۲۸,۶۰۷	۱۷,۵۰۱	۹۲٪
	۵	۰,۷۵	۲	۴۹۰,۳۵۹	۲۵,۳۹۶	۹۴٪

همانگونه که در جدول ۴-۴ مشاهده می‌شود، نسبت به مجموعه‌داده‌ی مورد استفاده و مقادیر مختلفی که برای پارامترها در نظر گرفته شده است، به این نتیجه خواهیم رسید که روش معرفی شده، زمان اجرای کمتری نسبت به روش خوشه‌بندی تطبیقی پایه داشته است. لازم به ذکر است که پردازش روش‌های مذکور، در سیستمی با مشخصات سخت افزاری مطابق جدول ۴-۵ صورت گرفته است.

جدول ۴-۵. مشخصات سیستم مورد استفاده

حافظه‌ی RAM	پردازنده	فرکانس پردازنده
4 GB	Intel core i3-2310M	2.10 GH

۵-۴ جمع‌بندی

ما در این فصل به معرفی و بررسی آزمایش‌های متفاوت سیستم پیشنهادی و سپس به مقایسه نتایج حاصل از آنها بر روی مجموعه‌داده‌های مختلف اقدام کردیم. همان‌طور که مشاهده شد، راهکار پیشنهادی توانست در تمامی آزمایش‌ها، عملکرد مناسبی را نسبت به رویکردهای قابل قیاس خود، به همراه داشته باشد. این روش جزء نخستین روش‌های خوشه‌بندی تطبیقی جریان‌داده به شمار می‌آید که بر اساس آزمایش‌های صورت گرفته وضعیت قابل قبولی را دارا بوده و می‌تواند به عنوان آغازگر مناسبی در پژوهش‌های مربوط به الگوریتم‌های خوشه‌بندی تطبیقی جریان‌داده‌ای در نظر گرفته شود.

فصل ۵ : جمع‌بندی و پژوهش‌های آینده

۵-۱ جمع بندی

آنچه که در این پایان‌نامه اشاره شد، معرفی و پیاده‌سازی یک روش خوشه‌بندی برخط با دیدگاهی تازه است. در این دیدگاه، خوشه‌هایی که در انتهای فرایند خوشه‌بندی به وجود می‌آیند به گونه‌ای هستند که می‌توان از آن‌ها برای تحلیل روابط گروهی موجود میان داده‌ها استفاده کرد. برای مثال اگر در یکی از خوشه‌های ایجاد شده تعداد قابل ملاحظه‌ای از یک گروه (برچسب مربوط به هر داده) قرار گرفته باشد، یک فرد متخصص می‌تواند جداگانه این خوشه را مورد بررسی قرار داده و روابطی را میان نمونه‌های اختصاص یافته به همان خوشه استخراج کند.

در این پژوهش به معرفی یک روش خوشه‌بندی تطبیقی جریان‌داده‌ای پرداخته شد. این روش، خلاصه‌ای از جریان‌داده‌ی ورودی را گرفته و فرایند خوشه‌بندی تطبیقی نهایی را بر روی آن به انجام می‌رساند. الگوریتم پیشنهادی ما با گرفتن پارامترهایی از قبیل اندازه‌ی پنجره، تعداد داده‌های خلاصه‌ی مورد نیاز و همچنین تعداد خوشه‌های مورد نظر، فرایند برخط کردن خوشه‌بندی تطبیقی را انجام می‌دهد. از مهم‌ترین مزایای روش ارائه شده می‌توان به کشف روابط گروهی درون یک جریان‌داده به کمک خلاصه‌سازی داده‌ها، افزایشی بودن، عدم وابستگی به تعداد خوشه‌ها و کاهش پیچیدگی زمانی و ذخیره‌سازی اشاره کرد که تا پیش از این کمتر در سایر پژوهش‌ها به آن پرداخته شده بود. راهکار پیشنهادی، در مقایسه با دو روش خوشه‌بندی تطبیقی پایه و K-Means، نتایج قابل قبولی داشته است.

در این پژوهش مشاهده شد که معیار میانگین مجموع مربعات خطا برای روش پیشنهادی کمی بیشتر از دو روش دیگر مورد مقایسه بوده است. این بیشتر بودن، به معنای مناسب نبودن روش پیشنهادی نیست بلکه روش پیشنهادی ما این قابلیت و نوآوری را دارد که علی‌رغم داشتن میانگین مجموع مربعات خطای بیشتر، بر روی خلاصه‌ای از داده‌ها فرایند خوشه‌بندی تطبیقی را انجام دهد.

همچنین باید در نظر داشته باشیم، دو روشی که با روش پیشنهادی ما مقایسه شده‌اند به صورت سنتی عمل کرده و همه‌ی داده‌ها را از ابتدا در اختیار دارند.

۵-۲ پژوهش‌های آینده

روش پیشنهادی، روی جریان‌داده‌ها عملیات خوشه‌بندی را انجام می‌دهد. از این رو، تغییر در شیوه‌ی پنجره‌گذاری و اضافه کردن ویژگی‌هایی از قبیل فراموشی[۱]داده‌ها، می‌تواند به بهبود عملکرد الگوریتم کمک کند. در واقع این ویژگی (فراموشی) به داده‌های جدید اهمیت بیشتری بخشیده و اثر داده‌های قدیمی را کم‌رنگ‌تر می‌کند. البته باید در نظر داشته باشیم که اضافه کردن این ویژگی باعث می‌شود که حافظه‌ی بیشتری برای ذخیره‌سازی خلاصه‌ی داده‌ها نیاز باشد. در صورتی که ما از نظر حافظه‌ی مورد نیاز برای ذخیره‌سازی دچار محدودیت باشیم، این امکان تا حد زیادی وجود دارد که افزودن ویژگی مذکور باعث بروز مشکلاتی در اجرای الگوریتم خوشه‌بندی شود. بهره‌گیری از این امر که با هدف ایجاد خلاصه‌های بهتر از داده‌ها برای توسعه‌ی الگوریتم خوشه‌بندی تطبیقی جریان‌داده انجام می‌شود، از اقدامات ما برای انجام کارهای آینده است.

از موارد دیگری که می‌توان به عنوان پژوهش‌های آینده در نظر گرفت پویا نمودن اندازه‌ی پنجره و همچنین تعداد داده‌های هسته است. به بیان دیگر، امکانی ایجاد شود که با اثبات روابط ریاضی مرتبط با این زمینه بتوانیم به اندازه‌ای از پنجره برسیم که به ازای ورود جریان داده‌های مختلف، متناسب با مقتضیات آن جریان و یا متناسب با مقتضیات نتیجه‌ی حاصل از پردازش جریان‌های‌داده، یک اندازه‌ی استاندارد و مناسب را برای پنجره‌ی مورد نظر تعیین نمود. همچنین با توجه به امکان ایجاد این استاندارد، می‌توان تعداد داده‌های هسته را نیز کم یا زیاد کرد که مطابق با نیاز یا صلاح‌دید کاربر، قابل تغییر باشد. با توجه به این‌که اندازه‌ی جریان‌داده در واقعیت، ثابت و مشخص نیست؛ لذا ایجاد این

[۱] Forgetness

استاندارد برای سهولت در خوشه‌بندی نهایی داده‌ها امری ضروری به نظر می‌رسد. بنابراین ایجاد این استاندارد در زمینه‌ی پژوهشی مورد اشاره، انتظار ایجاد نتایج مناسب‌تری را می‌تواند به دنبال داشته باشد.

از دیگر مواردی که می‌توان به آن در پژوهش‌های آینده اشاره کرد، ضرب کردن وزن خلاصه‌ی داده‌ها در معیار هزینه است. به عبارت دیگر، اگر ما بخواهیم تاثیر مربوط به هزینه‌ی ناشی از اختصاص داده‌های خلاصه شده به مراکز موجود را در نظر بگیریم، باید از یک ضریب ثابت استفاده کنیم که این ضریب نیز برای جریان‌داده‌های مختلف می‌تواند متفاوت باشد.

مراجع

[۱] Almalki, E. H., & Abdullah, M. (**2018**). A survey on big data stream mining. *Journal of Fundamental and Applied Sciences*, *10*(4S), 278-284.

[۲] Dhurandhar, A., Ackerman, M., & Wang, X. (**2017**). Uncovering Group Level Insights with Accordant Clustering. In *Proceedings of the 2017 SIAM International Conference on Data Mining* (pp. 228-236). Society for Industrial and Applied Mathematics.

[۳] Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (**2005**). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 657-668.

[۴] Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. (**2001**). Constrained k-means clustering with background knowledge. In *ICML* (Vol. 1, pp. 577-584).

[۵] Mining, W. I. D. (**2006**). Data Mining: Concepts and Techniques. *Morgan Kaufinann*.

[۶] Abbasi, S., & Vaziri, B. (**2015**). Clustering Algorithms in Big data. *International Academic Journal of Science and Engineering*, *2*, 26-36.

[۷] Xu, R., & Wunsch, D. (**2005**). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645-678.

[۸] Xu, D., & Tian, Y. (**2015**). A comprehensive survey of clustering algorithms. *Annals of Data Science*, *2*(2), 165-193..

[۹] Baghshah, M. S., & Shouraki, S. B. (**2010**). Non-linear metric learning using pairwise similarity and dissimilarity constraints and the geometrical structure of data. *Pattern Recognition*, *43*(8), 2982-2992.

[۱۰] Vu, V. V., Labroche, N., & Bouchon-Meunier, B. (**2012**). Improving constrained clustering with active query selection. *Pattern Recognition*, *45*(4), 1749-1758.

[۱۱] Wang, Y., Xiang, Y., Zhang, J., Zhou, W., Wei, G., & Yang, L. T. (**2014**). Internet traffic classification using constrained clustering. *IEEE transactions on parallel and distributed systems*, *25*(11), 2932-2943.

[۱۲] Mansalis, S., Ntoutsi, E., Pelekis, N., & Theodoridis, Y. (**2018**). An evaluation of data stream clustering algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *11*(4), 167-187.

[۱۳] Nguyen, H. L., Woon, Y. K., & Ng, W. K. (**2015**). A survey on data stream clustering and classification. *Knowledge and information systems*, *45*(3), 535-569.

[۱۴] Zhang, T., Ramakrishnan, R., & Livny, M. (**1996**). BIRCH: an efficient data clustering method for very large databases. In *ACM Sigmod Record* (Vol. 25, No. 2, pp. 103-114). ACM.

[۱۵] Guha, S., Meyerson, A., Mishra, N., Motwani, R., & O'Callaghan, L. (**2003**). Clustering data streams: Theory and practice. *IEEE transactions on knowledge and data engineering*, *15*(3), 515-528.

[۱۶] Ackermann, M. R., Märtens, M., Raupach, C., Swierkot, K., Lammersen, C., & Sohler, C. (**2012**). StreamKM++: A clustering algorithm for data streams. *Journal of Experimental Algorithmics (JEA)*, *17*, 2-4.

[۱۷] Silva, J. A., Faria, E. R., Barros, R. C., Hruschka, E. R., De Carvalho, A. C., & Gama, J. (**2013**). Data stream clustering: A survey. *ACM Computing Surveys (CSUR)*, *46*(1), 13.

[۱۸] Gama, J., Rodrigues, P. P., & Lopes, L. (**2011**). Clustering distributed sensor data streams using local processing and reduced communication. *Intelligent Data Analysis*, *15*(1), 3-28.

[۱۹] Youn, J., Shim, J., & Lee, S. G. (**2018**). Efficient Data Stream Clustering With Sliding Windows Based on Locality-Sensitive Hashing. *IEEE Access*, *6*, 63757-63776.

[۲۰] Badiozamany, S., Orsborn, K., & Risch, T. (**2016**). Framework for real-time clustering over sliding windows. In *Proceedings of the 28th International Conference on Scientific and Statistical Database Management* (p. 19). ACM.

[۲۱] Yang, C., Bruzzone, L., Guan, R., Lu, L., & Liang, Y. (**2013**). Incremental and decremental affinity propagation for semisupervised clustering in multispectral images. *IEEE Transactions on Geoscience and Remote Sensing*, *51*(3), 1666-1679.

[۲۲] Al-Harbi, S. H., & Rayward-Smith, V. J. (**2006**). Adapting k-means for supervised clustering. *Applied Intelligence*, *24*(3), 219-226.

[۲۳] Sinha, A., & Jana, P. K. (**2016**). A novel K-means based clustering algorithm for big data. In *Advances in Computing, Communications and Informatics (ICACCI), 2016 International Conference on* (pp. 1875-1879). IEEE.

[۲۴] Fitriyani, S. R., & Murfi, H. (**2016**). The k-means with mini batch algorithm for topics detection on online news. In *Information and Communication Technology (ICoICT), 2016 4th International Conference on* (pp. 1-5). IEEE.

[۲۵] Liberty, E., Sriharsha, R., & Sviridenko, M. (**2016**). An algorithm for online k-means clustering. In *2016 Proceedings of the Eighteenth Workshop on Algorithm Engineering and Experiments (ALENEX)* (pp. 81-89). Society for Industrial and Applied Mathematics.

[۲۶] Zhang, Y., Tangwongsan, K., & Tirthapura, S. (**2017**). Streaming algorithms for k-means clustering with fast queries. *arXiv preprint arXiv:1701.03826*.

[۲۷] Comito, C., Pizzuti, C., & Procopio, N. (**2016**). Online clustering for topic detection in social data streams. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)* (pp. 362-369). IEEE.

[۲۸] Baillargeon, S., Hallé, S., & Gagné, C. (**2016**). Stream clustering of tweets. In *Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1256-1261). IEEE Press.

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

[۳۰] Arthur, D., & Vassilvitskii, S. (**2007**). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035). Society for Industrial and Applied Mathematics.

[۳۱] Dua, D., & Karra Taniskidou, E. (**2017**). UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California. *School of Information and Computer Science*.

[۳۲] Bifet, A., Read, J., Holmes, G., & Pfahringer, B. (**2018**). Streaming Data Mining with Massive Online Analytics (MOA). *SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE*, *83*(1), 1-25.

Abstract

Nowadays, with the increasing rate of data generation and the emergence of the data stream concept, we can extract useful information from data by data mining techniques. This information can be used in a variety of areas such as social network analysis, patient medical information grouping, and network intrusion detection. Among the information extraction strategies, we can point to clustering, which is a convenient way of organizing data. In this thesis, a new approach is introduced based on accordant clustering. Accordant clustering attempts to create more interpretable clusters by considering the background knowledge in order to discover possible group relationships among data and obtain more useful information from clusters. If the data already has a predefined grouping and we want to use this information in clustering, such that a relationship exists between the clusters and the prior groups, accordant clustering can be used. Accordant clustering, despite adopting a new approach to clustering, is not applicable on data streams and executes centrally.

In the proposed method of this thesis, the accordant clustering process is presented in an online and incremental manner in order to apply the concept of accordant clustering on the data streams and discover the group relationships in such environments. Conventional clustering methods attempt to put similar data in a cluster; while the proposed method is using the background knowledge with a new approach in the concept of clustering. One of the important advantages of the proposed method is the possibility of finding the existing group relationships in a data stream, reducing the memory required for storing data, reducing the time and computational complexity, as well as the lack of dependence on the number of clusters in order to create the final clustering. The results of the experiments on artificial and real data sets confirm the proper performance of the proposed method in comparison with the basic accordant clustering and K-Means clustering.

Keywords: data stream, clustering, accordant clustering, incremental learning, background knowledge

Shahrood University of Technology

Faculty of Computer Engineering
M.Sc Thesis in Artificial Intelligence Engineering

Analysis of group relationships in data streams using clustering

By : Milad Mohammadi

Supervisor:
Dr. Hoda Mashayekhi

Advisor:
Dr. Mansor Fateh

January 2019