دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

سال پانزدهم، شماره‌های ۵۵ و ۵۶، بهار و تابستان
۱۴۰۲
صفحات: ۸۵تا ۱۰۳

Fuzzy Multicore Clustering of Big Data in the Hadoop Map Reduce Framework

Seyyed Omid Azarkasb[*], Seyyed Hosein Khaasteh[**], Mostafa Amiri[***]

[*]Lecturer and PhD student in computer engineering, artificial intelligence and robotics, Khajeh Nasiruddin Tousi University of Technology
[**]Specialized doctorate in computer engineering with artificial intelligence, assistant professor and faculty member of Khajeh Nasiruddin Tousi University of Technology
[***] Master's Degree in Computer Engineering, Software Orientation, Khajeh Nasiruddin Tosi University of Technology

Abstract

A logical solution to consider the overlap of clusters is assigning a set of membership degrees to each data point. Fuzzy clustering, due to its reduced partitions and decreased search space, generally incurs lower computational overhead and easily handles ambiguous, noisy, and outlier data. Thus, fuzzy clustering is considered an advanced clustering method. However, fuzzy clustering methods often struggle with non-linear data relationships. This paper proposes a method based on feasible ideas that utilizes multicore learning within the Hadoop map reduce framework to identify inseparable linear clusters in complex big data structures. The multicore learning model is capable of capturing complex relationships among data, while Hadoop enables us to interact with a logical cluster of processing and data storage nodes instead of interacting with individual operating systems and processors. In summary, the paper presents the modeling of non-linear data relationships using multicore learning, determination of appropriate values for fuzzy parameterization and feasibility, and the provision of an algorithm within the Hadoop map reduce model. The experiments were conducted on one of the commonly used datasets from the UCI Machine Learning Repository, as well as on the implemented CloudSim dataset simulator, and satisfactory results were obtained. According to published studies, the UCI Machine Learning Repository is suitable for regression and clustering purposes in analyzing large-scale datasets, while the CloudSim dataset is specifically designed for simulating cloud computing scenarios, calculating time delays, and task scheduling.

Keywords: Big Data Clustering, Fuzzy Multicore Learning, Hadoop Map Reduce, Task Scheduling, Cloud Computing, Pattern Recognition.

خوشه‌بندی فازی‌چندهسته‌ای کلان‌داده‌ها در چارچوب نگاشت‌کاهش هدوپ

سیدامید آذرکسب *[۱]، سیدحسین خواسته**، مصطفی امیری***

* مدرس و دانشجوی دکترای تخصصی مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز، دانشگاه صنعتی خواجه نصیرالدین طوسی

** دکترای تخصصی مهندسی کامپیوتر گرایش هوش مصنوعی، استادیار و عضو هیئت علمی دانشگاه صنعتی خواجه نصیرالدین طوسی

*** کارشناسی ارشد مهندسی کامپیوتر گرایش نرم افزار، دانشگاه صنعتی خواجه نصیرالدین طوسی

تاریخ دریافت:۱۴۰۰/۱۱/۲۳ تاریخ پذیرش:۱۴۰۱/۰۷/۰۷

نوع مقاله:پژوهشی

چکیده

یک راه‌حل منطقی برای لحاظ‌کردن همپوشانی خوشه‌ها، انتساب مجموعه‌ای از درجه عضویت به هر داده است. به‌دلیل کم‌شدن افرازها و کوچک‌شدن فضای‌جستجو، خوشه‌بندی فازی عموما دارای سربار محاسباتی کمتری بوده، تشخیص و مدیریت داده‌های مبهم، نویزدار و داده‌های‌پرت نیز در آن به‌سهولت انجام می‌گیرد. ازاینرو خوشه‌بندی فازی از نوع پیشرفته روش‌های خوشه‌بندی به شمار می‌رود. اما روش‌های خوشه‌بندی فازی در مواجه با روابط غیرخطی داده‌ها ناتوانند. روش پیشنهادی این مقاله می‌کوشد تا مبتنی بر ایده‌های امکان پذیری، از یادگیری چندهسته‌ای در چارچوب نگاشت‌کاهش هدوپ برای تشخیص خوشه‌های خطی‌جدایی‌ناپذیر با ساختار کلان‌داده‌های پیچیده، استفاده کند. مدل یادگیری چندهسته‌ای قادر به کشف روابط پیچیده بین داده‌ای بوده و در عین حال هدوپ ما را قادر خواهد ساخت تا به جای تعامل با سیستم عامل و پردازنده، با یک کلاستر منطقی از پردازش‌ها و گره‌های انباره داده تعامل داشته باشیم و عمده کار را بر عهده فریم‌ورک بیندازیم. به طور خلاصه مدل‌سازی روابط غیرخطی داده‌ها با استفاده از مدل یادگیری چندهسته‌ای، تعیین مقادیر مناسب برای پارامترهای فازی‌سازی و امکان‌پذیری، و ارائه الگوریتم در مدل نگاشت‌کاهش هدوپ از دستاوردهای کلیدی مقاله حاضر می‌باشد. آزمایشها برروی یکی از مجموعه داده‌های پر استفاده مخزن یادگیری UCI و همچنین برروی دیتاست شبیه‌ساز CloudSim پیاده سازی شده است و نتایج قابل قبولی به دست آمده است. طبق مطالعات منتشر شده، مخزن یادگیری UCI برای مقاصد رگرسیون و خوشه‌بندی کلان داده، و مجموعه داده CloudSim برای شبیه‌سازی موارد مربوط به رایانش ابری، محاسبه تأخیرهای زمانی و زمانبندی انجام وظایف معرفی شده‌اند.

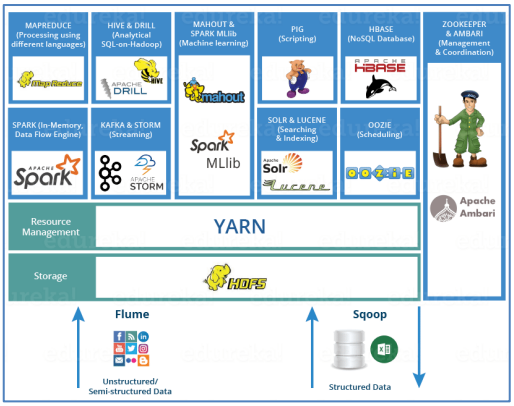
واژگان کلیدی: داده‌های کلان، خوشه‌بندی، منطق فازی، یادگیری چندهسته‌ای، هدوپ، نگاشت‌کاهش

[۱] نویسنده مسئول:سید امید آذرکسبazarkasb@ymail.com

۱. مقدمه

مساله خوشه‌بندی به عنوان یک مساله چالش برانگیز با دامنه کاربرد فراوان در حوزه یادگیری ماشین و داده‌کاوی مطرح است. در این بین کلان‌داده یک واقعیت در جهان کنونی به خصوص در رایانش‌های نوین نظیر رایانش ابری و رایانش مه است و مسأله‌ای است که سیستم‌های واقعی باید با آن دست‌وپنجه نرم کنند.

جوامع اطلاعاتی نوین به‌واسطه مخازن گسترده‌ای از داده تعریف می‌شوند. کلان‌داده اصطلاحی است که برای توصیف حجم زیادی جوامع اطلاعاتی نوین به‌واسطه مخازن گسترده‌ای از داده تعریف می‌شوند. جوامع اطلاعاتی نوین به‌واسطه مخازن گسترده‌ای از داده تعریف می‌شوند. کلان‌داده اصطلاحی است که برای توصیف حجم زیادی از داده‌های پیچیده و متغیر که با سرعت بالایی تولید می‌شوند، به‌کار می‌رود. برای ضبط، ذخیره، توزیع، تجزیه و تحلیل و همچنین مدیریت چنین داده‌هایی، روش‌ها و فناوری‌های پیشرفته‌ای مورد نیاز هستند. چالش اصلی در زمینه کلان‌داده‌ها نحوه پردازش داده‌های جمع‌آوری شده برای درک و به‌دست آوردن بینش جدید در یک زمان مناسب و با هزینه معقول است. یکی از ابزارهایی که در این خصوص تولید شده، هدوپ است. هدوپ یک چارچوب مدیریت و پردازش کلان‌داده به صورت توزیع شده می‌باشد. در نسخه اولیه هدوپ موتور پردازشی که از آن استفاده می‌کرد به نگاشت‌کاهش محدود می‌شد ولی پس از انتشار نسخه بعدی، موتورهای پردازشی دیگری اعم از اسپارک هم می‌توانند از هدوپ استفاده کنند. از قابلیت‌هایی که هدوپ دارد این است که می‌توانیم آنرا تبدیل به یک دیتابیس کنیم البته نه به تنهایی بلکه با کمک Hive یا Hbase [۱]. اکوسیستم جدید هدوپ در شکل ۱ نشان داده شده است.



شکل۱. اکوسیستم جدید هدوپ [۲]

این اکوسیستم با مجموعه ابزارهایی که در اختیار دارد می‌تواند به حل مشکلات کلان داده‌ها کمک کند. هدوپ را می‌توان به عنوان مجموعه‌ای در نظر بگیریم که دربر دارنده‌ی تعدادی از سرویس‌های کار با داده در درون خود می‌باشد. این سرویس‌ها به صور خلاصه عبارتند از [۲]:

- HDFS: سیستم فایل توزیع شده هدوپ،
- YARN: چارچوبی برای مدیریت منابع و زمان‌بندی،
- MapReduce: موتور پردازش داده‌ها با استفاده از APIهای زبان برنامه‌نویسی،
- Spark: موتور پردازش داده‌ها مبتنی بر رویکرد درون حافظه،
- HIVE&PIG: سرویس‌های پردازش داده با استفاده از زبان‌های پرس‌وجو نظیر SQL،
- HBase: پایگاه داده‌ی غیر رابطه‌ای،
- Spark MLlib&Mahout: چارچوب‌های یادگیری ماشین،
- Apache Drill: زبان پرس‌وجو SQL بروی هدوپ،
- Zookeeper: مدیریت خوشه (کلاستر)،
- Oozie: زمان‌بند کار،
- Sqoop&Flume: سرویس‌های جمع‌آوری داده،
- Solre & Lucene: جستجو و شاخص گذاری،
- Ambari: آماده‌سازی، نظارت و نگه‌داری از خوشه.

هدوپ با افزایش تحمل خطا در محیط توزیع شده مشکلات سرویس‌های بزرگی مانند گوگل، یاهو و فیسبوک را حل کرده‌است. در گوگل، نمایه کردن صفحات برای موتور جستجو و تحلیل ترجمه گوگل، در یاهو، جستجوی نقشه یاهو و شناسایی هرزنامه‌ها، و در فیسبوک، داده‌کاوی، بهینه‌سازی تبلیغات و شناسایی هرزنامه‌ها از طریق هدوپ انجام می‌شود [۳]. پردازش داده در هدوپ به‌صورت دسته‌ای است. پردازش دسته‌ای یک روش کارآمد پردازش مجموعه داده‌های استاتیک است. در این بین الگوریتم های خوشه‌بندی محبوب‌ترین، توانمندترین و متداول‌ترین روش ها، برای مواجه با کلان‌داده‌ها هستند. هدف اصلی مورد توجه این مقاله خوشه‌بندی کلان داده‌ها به صورت کارا می‌باشد. به دلیل پیچیدگی و زمان اجرای بالا، تکنیک‌های خوشه‌بندی سنتی را نمی‌توان برای چنین حجمی از داده‌ها استفاده کرد. از طرفی برخی از اشیا در این داده‌ها دارای صفات مشترکی از چند خوشه هستند. لذا توجه نویسندگان این مقاله به خوشه‌بندی فازی منعطف گردید که در آن یک راه حل منطقی برای لحاظ کردن همپوشانی خوشه

ها انتساب مجموعه‌ای از درجه عضویت به هر داده است. انتساب چندگانه درجه عضویت ناشی از ماهیت فازی خوشه‌ها است. خوشه بندی فازی عموما دارای سربار محاسباتی کمتری است. الگوریتم خوشه‌بندی قدرتمند بدون نیاز به داده‌های از دست‌رفته باید بتوانند کار خود را انجام دهند. اما روش‌های خوشه‌بندی فازی در مواجه با روابط غیرخطی داده‌ها و داده‌های از دست رفته ناتوانند. برای مقابله با چنین مواردی، تکنیک خوشه‌بندی مبتنی بر هسته پیشنهاد می‌گردد. در عین حال، در ابزار نگاشت کاهش هدوپ ما قادر خواهیم بود علاوه بر بکارگیری امکانات چندهسته‌ای، محاسبات خود را به فریم ورک واگذار کنیم و به قدرت محاسباتی بالا و بهینه دست یابیم. لذا در این مقاله ما با بهره‌گیری از ایده نگاشت‌کاهش هدوپ و ترکیب آن با خوشه‌بندی فازی چندهسته‌ای، شیوه جدیدی را برای خوشه‌بندی با کیفیت و کاراتر کلان‌داده‌ها ارائه کرده‌ایم.

ساختار مقاله در ادامه به اینصورت است که در بخش دوم به بیان پیشینه پژوهشی و معرفی خوشه‌بندی فازی و شیوه نگاشت کاهش می‌پردازیم و در بخش سوم شیوه پیشنهادی مد نظر را معرفی می‌کنیم، در بخش چهارم ملاحظات فنی و پیاده سازی روش پیشنهادی و در بخش پنجم به بیان نتایج می‌پردازیم. در پیوست نیز اثبات همگرایی تابع هدف بر اساس یکی از مراجع ارائه می گردد.

۲. ادبیات موضوعی و پیشینه پژوهشی

۲-۱- خوشه‌بندی فازی‌چندهسته‌ای

اخیرا شاهد استفاده متنوع الگوریتم‌های خوشه‌بندی براساس نوع کاربرد در محیط واقعی هستیم. خوشه‌بندی مبتنی بر یادگیری چندهسته‌ای [۴]. از این جمله است. برخی از روش‌های مبتنی بر یادگیری هسته‌ای سعی در یادگیری ماتریس هسته [۵] و برخی سعی در یادگیری پارامترهای هسته‌ای با فرم مشخص دارند. و با تعریف یک تابع هدف مناسب تلاشی در جهت یافتن ماتریس هسته انجام داده‌اند [۶]. روش PCK-Means تابع هدف را با انتساب وزن به متغیرها و در نظر گرفتن عبارت جریمه تعریف می کند. روش MPCK-Means، علاوه براینکه اجازه انتساب وزن به متغیرها را به همراه یک عبارت جریمه می‌دهد بلکه سعی دارد یادگیری تابع فاصله مناسب را در حین خوشه‌بندی انجام دهد [۷]. از جمله روش‌های خوشه‌بندی فازی می‌توان به [۸] اشاره کرد. روش خوشه‌بندی FCM یکی از روش‌های محبوب خوشه‌بندی فازی است که در بسیاری از موارد از معادل قطعی خود انعطاف پذیرتر است [۹]. با داشتن مجموعه داده $X=\{x_1,\dots,x_N\}$ که $x_i\in R^l$ و l بعد داده است، روش FCM با کمینه‌سازی تابع هدف معادله ۱، زیرمجموعه داده X را به C خوشه فازی افراز می‌کند [۱۰].

$$J_{FCM}(U,V) = \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}^{m} d_{ci}^{2} \quad \text{(معادله ۱)}$$

که $V=(v_1,\dots,v_C)$بردار C تایی نماینده خوشه‌ها ، d_{ci}^2 فاصله داده x_i تا نماینده خوشه c ام (برای مثال $||x_i - v_c||^2$)، N تعداد داده ها، C تعداد خوشه ها ، u_{ci} درجه عضویت فازی داده x_i به خوشه c با در نظر گرفتن شرط $\sum_{c=1}^{C} u_{ci} = 1$ ، m کنترل کننده درجه فازی خوشه‌بندی و$U \equiv [u_{ci}]$ماتریس $C\times N$ معروف به ماتریس افراز فازی است که سه شرط : ۱)$u_{ci}\in[0,1]$به ازای هر i و c، ۲) $\sum_{i=1}^{N} u_{ci} > 0$ به ازای هر c و ۳) $\sum_{c=1}^{C} u_{ci} = 1$ به ازای هر i را برآورده می‌سازد. روش FCMشرط مجموع برابر با ۱ را برای مجموع درجه عضویت یک داده به تمامی خوشه‌ها در نظر گرفته است. اگر چه این شرط در ایجاد خوشه‌های فازی مفید است، ولی روش FCM را نسبت به نویز و داده های پرت حساس می‌کند. با حذف این شرط روش PCM با تابع هدف معادله ۲ برای خوشه‌بندی معرفی می‌شود [۱۱].

$$J_{PCM}(T,V) = \sum_{c=1}^{C}\sum_{i=1}^{N} t_{ci}^{p} d_{ci}^{2} + \sum_{c=1}^{C} \mu_c \sum_{i=1}^{N} (1 - t_{ci})^{p} \quad (۲)$$

که t_{ci} درجه عضویت امکان‌پذیری داده x_i به خوشه c، $T \equiv [t_{ci}]$ ماتریس $C\times N$ معروف به ماتریس افراز امکان‌پذیری است که دو شرط: ۱)$t_{ci}\in[0,1]$ به ازای هر i و c و ۲) $\sum_{i=1}^{N} t_{ci} > 0$ به ازای هر c را برآورده می‌سازد، p فاکتور وزن درجه عضویت امکان‌پذیری وμ_c یک ثابت مثبت مناسب است. عبارت نخست در $J_{PCM}(T,V)$ تلاش می‌کند تا فاصله داده‌ها تا نماینده خوشه‌ها تا حد ممکن کمینه شود. در حالیکه عبارت دوم تلاش می‌کند t_{ci} تا حد ممکن بیشینه شود. روش PCM افرازی از امکان‌پذیری داده‌ها ارائه می‌دهد که درجه امکان‌پذیری هر داده بیانگر میزان خصوصیتی است که آن داده از خوشه‌ها دارد. این روش نسبت به نویز و داده‌های پرت مقاوم است زیرا داده‌های نویزی و پرت با درجه امکان‌پذیری کمتری به خوشه‌ها منتسب می‌شوند و همین امر سبب می‌شود تا این داده‌ها نتوانند نتیجه خوشه‌بندی را به‌طور موثری تحت تاثیر خود قرار دهند [۱۲]. با این وجود کارایی این روش بسیار حساس به مقدار دهی مناسب نمایندگان اولیه خوشه‌ها است. از معایب این روش می‌توان به تمایل روش به تولید خوشه‌های منطبق بر هم اشاره نمود. [۱۳] با ادغام دو روش FCM و PCM سعی در خوشه‌بندی مناسب داده‌ها به کمک راهکار فازی و مقاومت روش در مقابل داده‌های

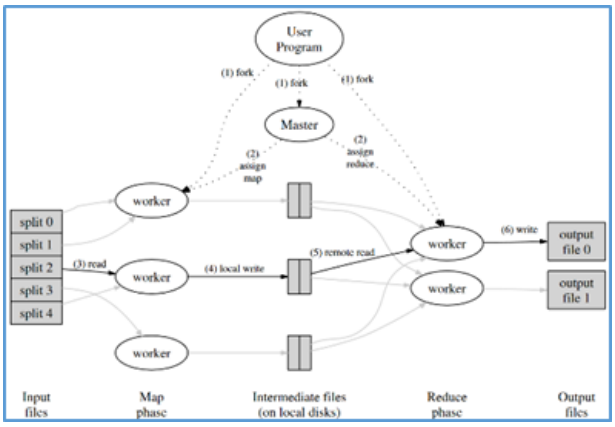
نویزی و پرت به کمک راهکار امکان‌پذیری دارد و با کمینه‌سازی تابع هدف معادله ۳، مجموعه داده X را به C خوشه فازی افراز می‌کند.

$$J_{PFCM}(T,U,V) = \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}^{m} t_{ci}^{p} d_{ci}^{2} + \sum_{c=1}^{C} \mu_c \sum_{i=1}^{N} u_{ci}^{m}(1-t_{ci})^{p} \quad (۳)$$

در این تابع هدف برای متغیرهای درجه عضویت فازی و درجه عضویت امکان‌پذیری شروط بیان شده در دو روش FCM و PCM صادق است. از کاستی‌های روش فوق می‌توان به ناتوانی این روش در تشخیص خوشه‌های غیرکروی و خوشه‌بندی داده‌های با ساختار پیچیده اشاره نمود. تحقیقات اخیر در این زمینه نشان می‌دهد که الگوریتم‌های خوشه‌بندی نیاز به طراحی مجدد برای معماری مدرن محاسبات دارند. بکارگیری ابزار نگاشت‌کاهش توزیع شده هدوپ یکی از این رویکردها است [۱۴].

۲-۲- نگاشت‌کاهش هدوپ

در حالت کلی، الگوریتم FCM، پیچیدگی زمانی بالایی دارد. از اینرو نویسندگان در مرجع [۱۵] روشی را ارائه کرده اند که از نگاشت‌کاهش برای بهبود سرعت اجرای الگوریتم و کیفیت خوشه بندی استفاده شده است، در عین‌حال، چارچوب محاسباتی موازی نگاشت‌کاهش را برای طراحی و پیاده‌سازی الگوریتم خود بکار گرفته اند. نگاشت‌کاهش، لایه پردازشی در معماری هدوپ است که بر پایه تقسیم و حل پیاده‌سازی شده است و یک چارچوب منبع باز و مدل برنامه‌نویسی ساده است که برای حل مسائل محاسباتی در مقیاس وسیع و همچنین به‌صورت توزیعی، مورداستفاده قرار می‌گیرد. نگاشت‌کاهش توسط گوگل در سال ۲۰۰۳ توسعه داده و ارائه شد و یک چارچوب نرم‌افزاری است که بستری امن و مقیاس‌پذیر برای توسعه کاربردهای توزیعی فراهم می‌کند. درواقع نگاشت‌کاهش مجموعه‌ای از توابع کتابخانه‌ای را در دل خود دارد که جزئیات و پیچیدگی اجرای همزمان را از دید کاربر پنهان می‌کند علاوه بر این مشکلاتی را از قبیل توازن بار، توزیع داده و تحمل خطا را کارسازی می‌کند. همانطور که در شکل ۲ نشان داده شده است در این روش، دو گام اصلی به نام‌های Reduce و Map وجود دارد [۱۶].



شکل۲. نمونه‌ای از چارچوب نگاشت‌کاهش [۱۶]

در مدل نگاشت‌کاهش ابتدا کل داده‌ها به بخش‌هایی تقسیم می‌شود که ما می‌توانیم برای پردازش این داده‌ها هرکدام از این بخش‌ها را به‌صورت جداگانه پردازش کرده و از خروجی آنها برای رسیدن به جواب نهایی بهره ببریم. روش کار به این‌صورت است که گره اصلی، ورودی را گرفته، و آن را به زیر مسائل کوچکتری تقسیم می‌کند. سپس آنها را بین گره‌هایی که وظیفه انجام کارها را دارند، توزیع می‌کند. ممکن است این نود نیز همین کار را تکرار کند که در این حالت یک ساختار چند سطحی داریم. در نهایت این زیر مسائل پردازش شده و پاسخ به گره اصلی ارسال می‌شود. این مرحله که معمولا ورودی آن خروجی مرحله قبل یعنی مرحله نگاشت می‌باشد روی داده‌های ورودی یکسری پردازش انجام می‌دهد که موجب تجمیع شدن داده‌ها می‌شود البته در بین دو مرحله نگاشت و کاهش گاها یکسری پردازش‌ها نیز روی داده‌ها صورت می‌گیرد و موجب سازماندهی‌تر شدن داده‌ها می‌گردد. این مرحله Shuffling نام دارد. مراحل انجام به‌صورت زیر است: گره اصلی که پاسخ‌ها و نتایج را دریافت کرد، آنها را برای ارائه خروجی، ترکیب می‌کند. در این میان ممکن است اعمالی مانند مرتب‌کردن، فیلترکردن، خلاصه‌کردن و یا تبدیل‌کردن، بر روی نتایج انجام دهد. این دو عمل اصلی، بر روی یک زوج مرتب (key, value) اعمال می‌شود. تابع نگاشت، یک زوج مرتب از داده را گرفته و به فهرستی از زوج مرتب‌ها تبدیل می‌کند سپس، چارچوب نگاشت‌کاهش، همه زوج‌ها با کلید یکسان را از همه فهرست‌ها جمع‌آوری کرده و آنها را باهم، یک گروه می‌کند. پس به ازای هر کلید تولیدشده، یک گروه ایجاد می‌شود. حال تابع کاهش، بر روی هر گروه اعمال می‌شود. در ادامه چارچوب نگاشت‌کاهش، یک لیست از (key, value) ها را به فهرستی از value ها تبدیل می‌کند [۱۷].

بر این اساس [۱۸]، K-Means را بر پایه نگاشت‌کاهش ارائه کرده است. در الگوریتم K-Means حساس‌ترین قسمت

محاسبه، محاسبه فاصله است. این الگوریتم در هر تکرار نیازمند (nk) محاسبه فاصله است که n تعداد اشیاء و k تعداد خوشه‌های ایجاد شده است. واضح است که محاسبه فاصله از یک شیء تا مرکز مرتبط با محاسبه فاصله بین سایر اشیاء با مراکز مربوطه نیست. بنابراین محاسبه فاصله بین اشیاء مختلف و مراکز می‌تواند به‌صورت مجزا و همزمان انجام شود. لذا در تابع نگاشت، هر نمونه را به نزدیکترین مرکز انتساب داده و در تابع کاهش، فرایند بروزرسانی مراکز جدید انجام می‌شود. [۱۹] تلاش‌هایی را برای بهبود معیار فاصله انجام داده است. معیار فاصله در این روش با تمرکز برروی مزایای ۳ روش فاصله اقلیدسی، فاصله DTW و فاصله SPDTW ارائه شده است. در [۲۰] یک الگوریتم فازی C-Means مبتنی بر نگاشت‌کاهش در هدوپ ارائه شده است. از آنجاییکه در قسمت قبل روش FCM معرفی گردید در این قسمت از بیان جزئیات آن خودداری می‌شود و در مقابل برای روشن‌تر شدن بیشتر روش ارائه شده، به قسمت‌های نگاشت و کاهش تمرکز می‌شود. الگوریتم معرفی شده MR(MapReaduce)-FCM نام دارد. در مرحله اول با ایجاد نمونه ماتریس عضویت به‌صورت تصادفی و به کمک کپی دیتاست HDFS به حافظه محلی، به محاسبه ماتریس مراکز خوشه‌ها براساس معادله ۴ پرداخته می‌شود. HDFS یک سیستم فایل توزیع شده برای نگهداری مقدار انبوه از اطلاعات در سرتاسر تعداد زیاد ماشین که در داخل یک کلاستر طراحی شده‌اند می‌باشد. از بلاک‌ها برای نگهداری فایل یا قسمتی از فایل استفاده می‌کند و از مدل یکبار نوشتن و چندبار خواندن برای دسترسی به داده پشتیبانی می‌کند.

$$\text{Purity} = \frac{1}{n}\sum_{j=1}^{k}\max_{i}\left(\left|L_i \cap C_j\right|\right) \qquad (۴)$$

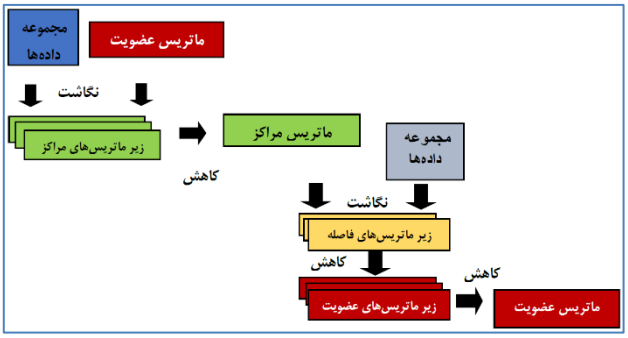
که در آن C_j شامل تمام نمونه‌های داده‌ای است که توسط الگوریتم خوشه‌بندی به خوشه j اختصاص داده شده است، n تعداد نمونه‌های داده در مجموعه داده‌ها، k تعداد خوشه‌هایی است که از فرآیند خوشه‌بندی تولید می‌شوند، L_i نشان‌دهنده تخصیص واقعی نمونه‌های داده است به خوشه i. الگوریتم این روش در الگوریتم ۱ نشان داده شده است.

Algorithm Main Procedure of MR-FCM Algorithm

Input: dataset
Output: purity
Randomly initialize membership matrix
while stopping condition is not met **do**
vertically merge data set with membership matrix and store in HDFS
Hadoop calls map and reduce jobs of first and then store locally
end while
calculate purity and write result to file

الگوریتم۱. شبه کد الگوریتم MR-FCM

همانطور که در الگوریتم ۱ دیده می‌شود این الگوریتم در هر مرحله کل دیتاست را به عنوان ورودی می‌گیرد و این‌کار را به تعداد دفعات مختلف تکرار می‌کند تا اینکه شرط توقف محقق گردد. شکل ۳ فلوچارت گام به گام هدوپ را نشان می‌دهد.



شکل ۳. فلوچارت گام به گام هدوپ

همانطور در شکل ۴ دیده می‌شود کار نگاشت‌کاهش در دو مرحله انجام می‌شود. در مرحله اول با ایجاد نمودن ماتریس تعلق به صورت تصادفی، و به کمک کل دیتاست به محاسبه ماتریس مراکز دسته ها می‌پردازد. سپس در مرحله دوم به کمک ماتریس مراکز دسته تولیدشده در مرحله قبل و کل دیتاست، به محاسبه ماتریس درجه عضویت‌ها پرداخته می‌شود. برای درک بهتر موضوع شبه کد این مراحل در الگوریتم‌های ۲، ۳، ۴ و ۵ نشان داده شده است.

Algorithm Map (key, value) of MapReduce job 1

Input: key: data record, value: data record values and membership matrix

Output: <key', value'> pair, where values' is the intermediate centroid matrix

For each key **do**
Calculate intermediate centroid matrix using equation 5 and store in value'
emit <key', value'> pair
End for

الگوریتم ۲. شبه کد الگوریتم نگاشت مرحله اول

Algorithm Map (key, value) of MapReduce job 1

Input: key: data record, value: intermediate centroid value

Output: <key', value'> pair, where values' is the centroid value

For each key **do**

Calculate centroid values by summing over intermediate centroid values and store in value'

emit <key', value'> pair

End for

الگوریتم ۳. شبه کد الگوریتم کاهش مرحله اول

Algorithm Map (key, value) of MapReduce job 2

Input: key: data record, value: data record values and centroid matrix

Output: <key', value'> pair, where value' is the intermediate membership matrix

For each key **do**

Calculate distances

Update intermediate membership matrix and store value'

emit <key', value'> pair

End for

الگوریتم ۴. شبه کد الگوریتم نگاشت مرحله دوم

Algorithm Reduce (key, value) of MapReduce job 2

Input: key: data record, value: intermediate membership matrix

Output: <key', value'> pair, where value' is the membership matrix

For each key **do**

Merge intermediate membership matrices and store in value'

emit <key', value'> pair

End for

الگوریتم ۵. شبه کد الگوریتم کاهش مرحله دوم

در ادامه شاهد آن هستیم که مرجع [۲۱] یک چارچوب هدوپ چندگره یکپارچه برای پیش‌بینی عوامل پرخطر دیابت با استفاده ازخوشه‌بندی فازی مبتنی بر نگاشت‌کاهش ارائه کرده است. علاوه بر موارد برشمرده شده، در بسیاری از کارهای اخیر، الگوریتم‌های خوشه‌بندی مانند K-prototypes، K-medoids و K-modes بوسیله نگاشت‌کاهش اصلاح شده‌اند که می‌توانید جزئیات آنها را در مرجع [۲۲] مشاهده نمایید.

۳. روش پیشنهادی

اگر چه همانطور که در پیشینه پژوهشی گفته شد ترکیب دو روش FCM و PCM، یک روش کارآمد جهت خوشه‌بندی داده‌های با خوشه‌های همپوشان و نویزی را تولید می‌کند ولی این روش محدود به تشخیص خوشه‌های کروی بوده و قادر به خوشه‌بندی کلان‌داده‌ها با ساختار پیچیده نمی‌باشد [۱۳]. در این بخش راهکاری برای خوشه‌بندی داده‌های غیرخطی جدایی‌پذیر و همپوشان ارائه می‌دهیم که نسبت به نویز و داده‌های پرت مقاوم است. برای لحاظ کردن خوشه‌های همپوشان و مقاوم در برابر نویز و داده‌های پرت با بهره‌گیری از ایده روش PFCM، تابع هدف پیشنهادی مدل‌سازی می‌شود. به‌منظور ارائه روش پیشنهادی در چارچوب خوشه‌بندی، تابع هدف PFCM با عبارت جریمه بهبود داده شد و برای تشخیص خوشه‌های خطی جدایی ناپذیر در ساختارهای پیچیده داده‌ای کلان، تابع هدف پیشنهادی را در معماری یادگیری چندهسته‌ای تعریف می‌نماییم، بدین‌گونه که خوشه‌بندی داده‌ها را در فضای غیرخطی ویژگی انجام می‌دهیم. از طرفی انتخاب و ترکیب توابع هسته جهت انجام یک خوشه‌بندی مبتنی بر هسته کارآمد امری بسیار حیاتی می‌باشد. متاسفانه در بسیاری از کاربردها یافتن ترکیب مناسبی از هسته‌ها امری دشوار می‌باشد. لذا روش پیشنهادی در این بخش ارائه می‌گردد تا موارد ذکر شده را در چارچوب یکپارچه خوشه‌بندی در نظر بگیرد. روش پیشنهادی با استفاده از مدل یادگیری چندهسته‌ای و تنظیم خودکار وزن هسته‌ها نسبت به توابع نامناسب یا ویژگی‌های نامرتبط ایمن می‌گردد. این امر سبب کاهش حساسیت روش پیشنهادی به انتخاب هسته ناکارآمد می‌گردد. معادله ۵ را که یک ترکیب خطی از M هسته پایه در ϕ برای نگاشت داده‌ها به فضای ویژگی باشد در نظر بگیرید:

$$\varphi(x)=w_1\varphi_1(x)+w_2\varphi_2(x)+...+w_M\varphi_M(x) \quad (۵)$$

در روش پیشنهادی کمینه‌سازی تابع هدف معادله ۶ جهت نیل به اهداف ذکر شده ارائه می‌گردد:

$$J_{\text{Proposed method}}(W,T,U,V) = \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}^{m} t_{ci}^{p} (\varphi(x_i)-v_c)^T(\varphi(x_i)-v_c) + \sum_{c=1}^{C}\mu_c \sum_{i=1}^{N} u_{ci}^{m}(1-t_{ci})^p \quad (۶)$$

$$+\alpha\left(\sum_{(i,j)\in M}\sum_{c=1}^{C}\sum_{\substack{l=1\\ l\neq c}}^{C} u_{ci}^{m}u_{lj}^{m}t_{ci}^{p}t_{lj}^{p} + \sum_{(i,j)\in C}\sum_{c=1}^{C} u_{ci}^{m}u_{cj}^{m}t_{ci}^{p}t_{cj}^{p}\right)$$

که M و C مجموعه محدودیت‌ها هستند. $V_c \in R^L$ نماینده خوشه c ام و $V \equiv [v_c]_{l\times c}$ یک ماتریس $L\times C$ است که هر ستون آن نماینده یک خوشه را نشان می‌دهد.

بردار $W=(w_1, w_2, \ldots, w_M)^T$ بردار وزن هسته‌ها است که شرط $\sum_{k=1}^{M} wk = 1$ را برآورده می‌سازد. $U \equiv [u_{ci}]_{c\times n}$ ماتریس عضویت فازی است که عنصر u_{ci} بیانگر درجه عضویت فازی داده x_i به خوشه c با در نظر گرفتن شرط $\sum_{c=1}^{C} u_{ci} = 1$ است. m کنترل‌کننده درجه فازی خوشه‌بندی است. $T \equiv [t_{ci}]_{c\times n}$ ماتریس عضویت امکان‌پذیری است که هر عنصر t_{ci} از آن بیانگر درجه عضویت امکان‌پذیری داده x_i به خوشه c است p نیز بیانگر فاکتور وزن درجه عضویت امکان‌پذیری است.

$(\varphi(x_i) - v_c)^T(\varphi(x_i) - v_c)$ بیانگر فاصله میان داده x_i و نماینده خوشه v_c در فضای ویژگی است. μ_c نیز پارامتر مقیاس است که مقدار مناسب برای آن به صورت معادله ۷ پیشنهاد شده است:

$$\mu_c = \frac{\sum_{i=1}^{N} u_{ci}^{m}t_{ci}^{p}(\varphi(x_i) - v_c)^T(\varphi(x_i) - v_c)}{\sum_{i=1}^{N} u_{ci}^{m}t_{ci}^{p}} \qquad (۷)$$

دو جمله اول از رابطه (W,T,U,V) $J_{\text{Proposed method}}$ روش PFCMرا به نوع بهبود یافته چندهسته‌ای از آن مدل می‌کند، به گونه‌ای که فشردگی خوشه‌ها در فضای ویژگی تضمین گردد. جمله اول برابر با مجموع مربعات فاصله داده‌ها تا نماینده خوشه هاست که با درجه‌های عضویت فازی و امکان‌پذیری به صورت وزن دار مدل شده است. جمله دوم تا حد ممکن سعی در بیشینه نمودن درجه‌های امکان‌پذیری برای فرار از پاسخ‌های بدیهی دارد. جمله سوم از رابطه جریمه را کنترل نموده و با مقدار α به عنوان درجه اهمیت نسبی نظارت وزن‌دهی شده است. این جمله متشکل از دو عبارت جریمه است. بخش اول از جمله سوم جریمه خوشه‌های متفاوت را با توجه به درجه عضویت متناظرشان جریمه می‌نماید. در مقابل بخش دوم از جمله سوم جریمه خوشه‌های یکسان با توجه به درجه عضویت متناظرشان جریمه می‌شود. با کمینه‌سازی این رابطه مجموع فاصله‌های درون خوشه‌ای افرازهای نهایی در فضای ویژگی تا حد ممکن کمینه می‌شود، قضیه زیر شروط لازم برای کمینه‌سازی تابع هدف رابطه را بیان می‌کند [۲۳].

قضیه: اگر به $U \equiv [u_{ci}]_{c\times n}$، $T \equiv [t_{ci}]_{c\times n}$ و $W \equiv [w_k]_{m\times 1}$ مقادیر زیر منتسب گردد آنگاه تابع هدف $J_{\text{Proposed method}}$ به یکی از کمینه‌های محلی خود همگرا می‌شود. روابط مورد نظر در معالات ۸ تا ۱۵ آمده است.

$$t_{ci} = \frac{1}{1 + \left(\frac{D_{ci}^2 + \alpha\left(S_{ci}^{M} + S_{ci}^{C}\right)}{\mu_c}\right)^{\frac{1}{p-1}}} \qquad (۸)$$

$$u_{ci} = \frac{1}{\sum_{k=1}^{C}\left(\frac{t_{ci}^{p-1}\left(D_{ci}^2 + \alpha\left(S_{ci}^{M} + S_{ci}^{C}\right)\right)}{t_{ki}^{p-1}\left(D_{ki}^2 + \alpha\left(S_{ki}^{M} + S_{ki}^{C}\right)\right)}\right)^{\frac{1}{m-1}}} \qquad (۹)$$

$$W_k = \frac{\frac{1}{y_k}}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_M}} \qquad (۱۰)$$

$$D_{ci}^2 = (\varphi(x_i) - v_c)^T(\varphi(x_i) - v_c) \qquad (۱۱)$$

که

$$S_{ci}^{M} = \sum_{(i,j)\in M}\sum_{\substack{l=1\\ l\neq c}}^{C} u_{lj}^{m}t_{lj}^{p}\,, \qquad (۱۲)$$

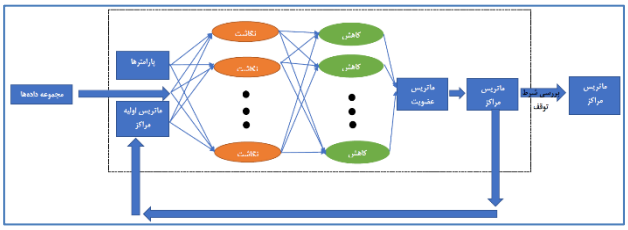
$$S_{ci}^{C} = \sum_{(i,j)\in C} u_{cj}^{m}t_{cj}^{p}\,, \qquad (۱۳)$$

$$Y_k = \sum_{c=1}^{C}\sum_{i=1}^{C} u_{ci}^{m}t_{ci}^{p}\theta_{ci}^{k}\,, \qquad (۱۴)$$

$$\theta_{ci}^{k} = K_k(x_i, x_i) - \frac{2\sum_{j=1}^{N} u_{cj}^{m}t_{cj}^{p}k_k(x_i,x_i)}{\sum_{j=1}^{N} u_{cj}^{m}t_{cj}^{p}} + \frac{\sum_{r=1}^{N}\sum_{s=1}^{N} u_{cr}^{m}t_{cr}^{p}u_{cs}^{m}t_{cs}^{p}k_k(x_r,x_s)}{\left(\sum_{r=1}^{N} u_{cr}^{m}t_{cr}^{p}\right)\left(\sum_{s=1}^{N} u_{cs}^{m}t_{cs}^{p}\right)} \qquad (۱۵)$$

طبق آنچه که در پیوست ارائه شده است، اثبات می‌شود که تابع هدف $J_{\text{Proposed method}}$ با ارائه خوشه‌بندی معناداری از داده‌ها به یکی از کمینه‌های محلی خود همگرا می‌شود [۲۳]. روش پیشنهادی در ۳ گام انجام می‌گردد. در گام اول ابتدا تعداد خوشه‌ها و نماینده خوشه‌ها به صورت تصادفی تعیین و داده‌ها به زیربخش‌هایی تقسیم می‌گردند. در گام دوم، پردازش زیربخش‌ها به طور همزمان و براساس چارچوب نگاشت‌کاهش طی دو مرحله انجام می‌شود. در مرحله نگاشت، داده‌های هر زیربخش به قسمت‌هایی تقسیم می‌شود در ادامه برای هر قسمت کلید میانی استخراج می‌گردد و در مرحله کاهش، ماتریس V که هر ستون آن

نماینده یک خوشه است و ماتریس U که ماتریس عضویت فازی است به‌دست می‌آیند. در گام سوم مراکز خوشه‌ها استخراج و در صورت نیاز گام دوم تا رسیدن به شرط توقف الگوریتم ادامه می‌یابد. شرط توقف، کمینه‌شدن تابع هدف و کوچکتر بودن تغییر مقدار عددی آن از حدآستانه مورد نظر می‌باشد. فلوچارت روش پیشنهادی در شکل ۴ نشان داده شده است.



شکل ۴. فلوچارت روش پیشنهادی

منطق کلی روش پیشنهادی به قرار زیر می‌باشد:

- ورودی: دیتاست، تعداد خوشه‌ها، بردار وزن هسته‌ها W، ماتریس نماینده خوشه‌ها V، کنترل‌کننده درجه فازی m، ماتریس عضویت امکان‌پذیری T، فاکتور وزن درجه عضویت امکان‌پذیری و μ_c پارامتر مقیاس.
- دریافت ورودی‌ها
- ایجاد تصادفی ماتریس نماینده خوشه‌ها V
- تقسیم تصادفی داده‌های ورودی به زیربخش‌ها در چارچوب نگاشت‌کاهش هدوپ

الف) در مرحله نگاشت، در هر گره تعدادی از نقاط به عنوان کلید و مراکز خوشه‌ها به صورت یک آرایه به عنوان مقادیر ارسال می‌شود و در آن گره فاصله نقطه از هر مرکز خوشه محاسبه و به‌صورت زوج‌هایی از کلید مقدار که کلید شامل مرکز خوشه و فاصله نقطه از آن و مقدار برابر خود نقطه است ارسال می‌شود. با استفاده از این تکنیک مقادیر ارسالی بر اساس مرکز خوشه و فاصله نقاط از آن ها مرتب شده و به کاهش‌دهنده ها ارسال می‌شوند.

ب) در مرحله کاهش، فاصله تمام نقاط از مرکز یک خوشه دریافت می‌شود و بر اساس پارامتری که تعیین می‌کنیم تعدادی از نقاط که فاصله کمتری از مرکز خوشه دارند و به صورت مرتب شده دریافت شده‌اند را انتخاب و مرکز خوشه جدید براساس آن چه در روش فازی گفته شد محاسبه می شود. مجددا می‌شود تمام مراکز خوشه جدید بدست آمده در کاهش دهنده‌ها را به نگاشت‌ها داد و همین عملیات را تا همگرایی مراکز خوشه‌ها ادامه داد.

ج) خروجی کاهش‌دهنده‌ها به منظور محاسبه V برای استفاده در محاسبات زیر بخش‌های بعدی، ترکیب می‌شود.

- ترکیب U های مربوط به تمام زیربخش‌ها و محاسبه U نهایی
- محاسبه بردار V
- بررسی شرط توقف
- پایان

بر این اساس شبه کد روش پیشنهادی در الگوریتم ۶ نشان داده شده است.

Inputs: dataset D
Number of clusters k
Weights of kernels W
Kernels of clusters V
Degree of fuzzification m
Probability of membership T
Scale parameter

Initialization:
Getting inputs
Initialization of V
Distributing data along nodes

While coverage kernels:

Map phase1:
Input: key: data record, value: intermediate centroid value
Output: <key', value'> pair, where values' is the centroid value

For each key do
Calculate centroid values by summing over intermediate centroid values and store in value'
emit <key', value'> pair
End for

Map phase 2:
Input: key: data record, value: data record values and centroid matrix
Output: <key', value'> pair, where value' is the intermediate membership matrix
For each key do
Calculate distances
Update intermediate membership matrix and store value'

emit <key', value'> pair
End for

Reduce phase:

Input: key: data record, value: intermediate membership matrix

Output: <key', value'> pair, where value' is the membership matrix

For each key do
Merge intermediate membership matrices and store in value'
emit <key', value'> pair
End for

End while

الگوریتم ۶. شبه کد روش پیشنهادی

۴. ملاحظات فنی و پیاده‌سازی

۴-۱- انتخاب هسته‌های پایه

انواع مختلفی از توابع هسته با تعداد متفاوت می‌توانند به عنوان هسته‌های پایه در آزمایش‌ها مورد استفاده قرار گیرند. در مسائل دنیای واقعی هیچ راهنمایی برای انتخاب مجموعه هسته مناسب وجود ندارد و مجموعه هسته‌های متفاوت کارایی متفاوتی از خوشه بندی را نتیجه می‌دهند. ما در روش پیشنهادی از سه دسته توابع هسته‌ای: توابع هسته حاصل از اطلاعات طیفی داده‌ها، توابع هسته گاوسی و توابع هسته چند جمله‌ای برای ساخت هسته‌های $\{K\}_{k=1}^{M}$ استفاده می‌شود. فرض کنید $X=[x_1,x_2,\dots,x_N]_{l\times N}$ یک ماتریس $l\times N$ که هر ستون آن نشان‌دهنده یک بردار داده از فضای R^l است باشد و $V=[v_1,v_2,\dots,v_N]_{N\times N}$ بردارهای ویژه هسته خطی X^TX باشد. در دسته اول از هسته‌ها، M_v ماتریس هسته $\{k_1^v,k_2^v,\dots,k_{M_v}^v\}$ به صورت زیر ساخته می شوند: $k_k^V=V_k^TV_k,\ k=1,2,\dots,M_v$ در دسته دوم M_g ماتریس هسته $\{k_1^v,k_2^v,\dots,k_{M_v}^v\}$ توسط نگاشت گاوسی به صورت معادله ۱۶ ساخته می‌شود:

$$k_k^V(x_i,x_j)=exp\left(\frac{(x_i-x_j)^T(x_i-x_j)}{2^{(M_g-k)}\sigma_X}\right),k=1,2,\dots,M_g \quad (۱۶)$$

که σ_X انحراف معیار $\binom{N}{2}$ جفت فاصله‌های نقاط مجموعه داده است. مضرب $2^{(M_g-k)}$ در مخرج کسر سبب تولید ماتریس‌های هسته با مقیاس‌های متفاوت می‌شود. در دسته آخر نیز از توابع هسته چند جمله‌ای به صورت معادله ۱۷ استفاده می‌گردد:

$$k_k^P(x_i,x_j)=\left(x_i^Tx_j-C\right)^T,k=1,2,\dots,M_P \quad (۱۷)$$

بنابراین $M=M_v+M_g+M_p$ ماتریس هسته و معادله ۱۸ را به عنوان هسته‌های پایه مورد استفاده در آزمایش‌ها خواهیم داشت.

$$\left\{k_1^v,k_2^v,\dots,k_{M_v}^v,k_1^g,k_2^g,\dots,k_{M_g}^g,k_1^P,k_2^P,\dots,k_{M_P}^P\right\} \quad (۱۸)$$

۴-۲- بررسی تابع هدف پیشنهادی از نگاه چندهسته‌ای

روش‌های مبتنی بر هسته از نگاشت داده‌ها به فضای ویژگی به منظور کشف روابط غیرخطی داده‌ها بهره می برند.

فرض کنید $\Phi=\{ø_1,ø_2,\dots,ø_M\}$ مجموعه ای از M نگاشت باشد به گونه‌ای که هر نگاشت $ø_k$ داده $x\in R^{\square}$ را به بردار L_k بعدی $ø_k(x)$ در فضای ویژگی نگاشت می‌کند. همچنین $\{k_1,k_2,\dots,k_M\}$ را هسته‌های متناظر با این نگاشت‌ها در نظر می‌گیریم به‌صورتی‌که در معادله ۱۹ داریم:

$$(x_i)^T ø_k(x_i)K_k(x_i,x_j)=ø_k \quad (۱۹)$$

ترکیب خطی این هسته‌ها به صورت معادله ۲۰ است:

$$\widehat{\emptyset}(x)=\sum_{k=1}^{M}w_k\,\emptyset(x) \quad (۲۰)$$

که $w_k\geq 0$ بیانگر وزن هسته i ام می باشد. از آنجا که همه نگاشت‌ها لزوما دارای بعد برابر نیستند، لذا ممکن است ترکیب خطی این نگاشت‌های ضمنی امکان‌پذیر نباشد. از این رو مجموعه جدیدی از نگاشت‌های مستقل $\varphi=\{\varphi 1,\varphi 2,\dots,\varphi M\}$ از نگاشت‌های اصلی Φ به صورت زیر معادله ۲۱ تعریف می‌شود:

$$\varphi_1(x)=\begin{bmatrix}\emptyset_1(x)\\0\\\vdots\\0\end{bmatrix},\ \varphi_2(x)=\begin{bmatrix}0\\\emptyset_2(x)\\\vdots\\0\end{bmatrix},\ \varphi_M(x)=\begin{bmatrix}0\\\vdots\\0\\\emptyset_M(x)\end{bmatrix}\in R^L \quad (۲۱)$$

هر یک از این نگاشت‌ها داده x را به یک بردار L بعدی تبدیل می کند که معادله ۲۲ به دست می آید:

$$L = \sum_{k=1}^{M} L_k \quad (۲۲)$$

ساختن نگاشت جدید از این طریق، یکسان بودن بعد ویژگی تمامی نگاشت‌ها را تضمین می‌کند، به‌گونه‌ای که ترکیب خطی آنها نیز خوش‌تعریف می‌باشد. نگاشت‌های تعریف شده تشکیل مجموعه‌ای از بردارهای عمود بر هم را می‌دهند به‌گونه‌ای که معادله ۲۳ به دست می آید:

$$\varphi_k(x_i)^T \varphi_{\acute{k}}(x_j) = \begin{cases} k_k(x_i, x_j) & k = \acute{k} \\ 0 & k \neq \acute{k} \end{cases} \quad (۲۳)$$

تابع هدف پیشنهادی سعی در یافتن ترکیب نامنفی خطی معادله ۲۴:

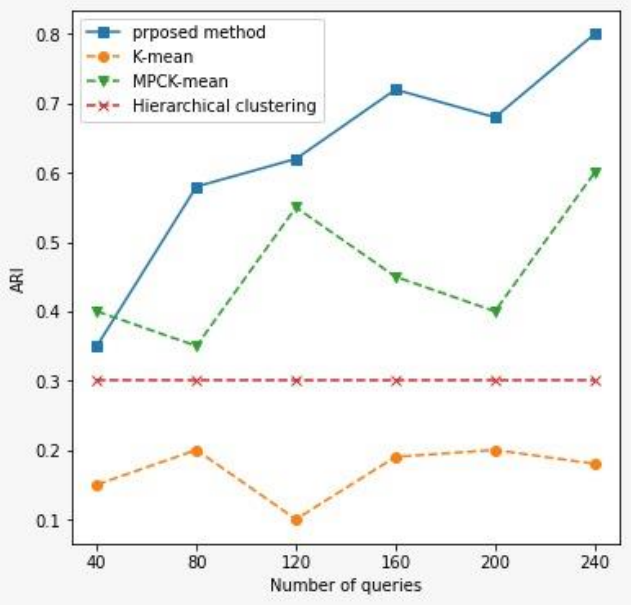
$$\varphi(x) = \sum_{k=1}^{M} w_k \varphi_k(x) \quad (۲۴)$$

از M هسته φ دارد تا بتواند داده‌ها را به بهترین فضای ویژگی نگاشت نماید. به همین دلیل در این تابع، هدف فاصله میان داده x_i و نماینده خوشه v_c به صورت $(\varphi(x_i)-v_c)^T(\varphi(x_i)-v_c)$ تعریف می‌شود.

۵. ارزیابی و تحلیل نتایج

به‌منظور ارزیابی، کارایی روش پیشنهادی در مقایسه با سه روش خوشه‌بندی KMeans [۱۹]، MPC-KMeans [۱۳] و سلسله مراتبی Hierarchical Clustering [۲۴] بررسی می گردد. در رابطه با دو روش اول، پیشتر در این مقاله صحبت به میان آمد اما توضیح مختصر در رابطه خوشه‌بندی سلسله مراتبی اینکه در این دسته از روش‌ها، داده‌ها با توجه به معیار فاصله به شیوه یک درخت سلسله مراتبی، سازماندهی می‌شوند. روش کار در خوشه‌بندی سلسله مراتبی به‌طور معمول براساس الگوریتم‌های حریصانه است. توضیح بیشتر اینکه خوشه‌بندی داده‌ها در یک ساختار سلسله مراتبی و با تولید یک نمودار درختی از خوشه‌ها به نام دندروگرام انجام می‌پذیرد. برش دندروگرام حاصل در هر سطح دلخواه سبب ایجاد خوشه‌های متفاوتی خواهد شد. روش‌های این دسته یا تقسیم‌کننده هستند یا تجمیعی. در روش تقسیم‌کننده، دندروگرام خوشه‌ها به روش بالا به پایین ساخته می‌شود. در این روش‌ها، نخست همه داده‌ها در یک خوشه واحد قرار می‌گیرند و سپس تقسیم خوشه‌ها معیارهای مناسب تا رسیدن به شرط همگرایی همراه با ساخت دندروگرام خوشه‌ها ادامه می‌یابد. برخلاف روش‌های تقسیم‌کننده، در روش‌های جمع‌کننده، دندروگرام خوشه‌ها به روش پایین به بالا ساخته می‌شود. اینگونه روش‌ها، هر داده را به عنوان یک خوشه در پایین‌ترین سطح در نظر می‌گیرند و ادغام خوشه را تا رسیدن به شرط توقف ادامه می دهند [۲۵]. BRICH یکی از روش‌های مطرح موجود در این دسته از روش‌های خوشه‌بندی می‌باشد [۲۶].

در مراجع [۲۷] و [۲۸]، ده تا از بهترین مجموعه‌های داده‌ای و پروژه‌هایی که برای مقاصد رگرسیون و خوشه‌بندی می‌باشد معرفی شده است. مجموعه داده کیفیت نوشیدنی، از مخزن یادگیری ماشین UCI، یکی از مجموعه داده‌های اشاره شده در این مراجع و سایر مراجع مشابه می‌باشد. این مجموعه داده از مجموعه‌های پر استفاده مخزن یادگیری UCI می‌باشد تا جاییکه تعداد بازدیدهای انجام گرفته از این مجموعه تا زمان ارائه مقاله حاضر، ۲۳۸۲۰۵۱ مورد و تعداد دانلود آن ۳۶۸۸۴ بار می‌باشد [۲۹]. مخزن یادگیری ماشین UCI مجموعه‌ای از پایگاه‌های داده، تئوری‌های دامنه و تولیدکننده‌های داده است که توسط جامعه یادگیری ماشین برای تحلیل تجربی الگوریتم‌های یادگیری ماشین استفاده می‌شود. در حال حاضر 623 نوع مجموعه داده در این مخزن ارائه می‌شود [۳۰]. آزمایش‌ها بر روی مجموعه داده UCI انجام شد. تعداد نمونه‌های مجموعه داده کیفیت نوشیدنی، ۴۸۹۸ عدد می‌باشد که اطلاعات ۱۲ ویژگی در آن مورد توجه قرار گرفته است. این ویژگی ها عبارتند از: اسیدیته ثابت، اسیدیته فرار، اسیدسیتریک، میزان شکر، کلریدها، دی اکسید گوگرد آزاد، دی اکسید گوگرد کل، غلظت، pH، سولفات‌ها، میزان ناخالصی و ویژگی آخر ویژگی کیفیت است که متغیر خروجی بوده و بر اساس داده‌های حسی می‌باشد. جهت کسب اطلاعات بیشتر در رابطه با ویژگی‌های این مجموعه داده می‌توانید به مرجع [۳۱] مراجعه بفرمایید. نتیجه آزمایش‌ها بر اساس جستارها و انحراف معیار ARI در شکل ۵ گزارش شده است. انحراف معیار ARI یکی از شاخص های پراکندگی است که نشان می‌دهد به‌طور میانگین داده‌ها چه مقدار از مقدار متوسط فاصله دارند. اگر انحراف معیار مجموعه‌ای از داده‌ها نزدیک به صفر باشد، نشانه آن است که داده‌های خوشه‌های متفاوت نزدیک به هم هستند و پراکندگی اندکی دارند در حالی که انحراف معیار بزرگ بیانگر پراکندگی قابل توجه داده‌ها می‌باشد. ماتریس‌های هسته $\{k_1^v, k_2^v, \dots, k_5^v, k_1^g, k_2^g, \dots, k_7^g\}$ در همه آزمایش‌ها به عنوان هسته‌های پایه در نظر گرفته شده‌اند.



شکل ۵. مقایسه کارایی روش پیشنهادی با سایر روش‌ها

تحلیل و مقایسه روش پیشنهادی با سایر روش‌ها براساس چه شاخص‌های میزان انحراف معیار و تعداد جستارها انجام پذیرفت. براساس آنچه که در نمودارهای شکل ۴ با بررسی رفتار سایر روش‌ها مشاهده می‌شود می‌توان دریافت که روش پیشنهادی از دیگر روش‌های خوشه‌بندی KMeans، MPC-KMeans و سلسله مراتبی کارآمدتر است. این امر نشان‌دهنده قابلیت اتکای بالای به روش پیشنهادی است. دلیل برتری روش پیشنهادی نسبت به روش KMeans را می‌توان در خاصیت غیرخطی روش پیشنهادی دانست که با بهره‌گیری از مدل یادگیری چندهسته‌ای در چارچوب نگاشت‌کاهش قادر به کشف روابط پیچیده بین داده ای می‌باشد در حالی‌که روش KMeans از فقدان این ویژگی رنج می‌برد.

در مقایسه روش پیشنهادی با روش سلسله مراتبی باید گفت که روش پیشنهادی کارایی بسیار بالاتری را نسبت به روش سلسله مراتبی از خود نشان داده است. لذا می‌توان نتیجه گرفت که روش سلسله مراتبی برای خوشه‌بندی داده‌های با ساختار پیچیده یا ابعاد بالا به اندازه کافی کارآمد نمی‌باشد. عدم مدل‌سازی روابط غیرخطی داده‌ها توسط روش سلسله مراتبی را می‌توان یکی از دلایل ناکارآمدی آن بیان نمود.

در مقایسه روش پیشنهادی با روش MPC-KMeans که به طور کل از دو روش KMeans و سلسله مراتبی نتایج بهتری ارائه داده است می توان گفت این برتری در مجموعه داده‌های با ابعاد بالاتر مشهودتر است. هر چند MPC-KMeans سعی در یادگیری متریک صحیح در حین خوشه‌بندی دارد، اما آزمایش‌ها نشان داده اند که کارایی این روش با افزایش تعداد جستارها به بیش از یک آستانه افزایش چندانی ندارد. همانگونه که در شکل ۵ نشان داده شده است با افزایش جستارها از یک آستانه، کارایی روش MPC-KMeans در یک سطح تقریباً ثابت و غیر افزایشی باقی می‌ماند. در مقابل در روش پیشنهادی شاهد روند افزایش کارایی به ازای افزایش تعداد جستار هستیم که این امر بر کارایی ویژگی‌های انتخابی و قابلیت یادگیری بالای روش پیشنهادی دلالت دارد.

تعیین مقدار مناسب برای ضرایب وزن m و p به عنوان یک مساله باز در حوزه خوشه‌بندی فازی مطرح است. بطور مثال در جدول۱ کارایی روش پیشنهادی با در نظر گرفتن مقادیر $m=p=3$ و $m=p=1.3$ نشان داده شده است. طبق این شکل فازی‌تر کردن خوشه‌بندی نتایج بهتری را ارائه داده است.

جدول ۱. کارایی روش پیشنهادی با ضرایب وزنی m=p=3 و m=p=1.3

ضرایب وزنی	انحراف معیار
m=p=1.3	0.95
m=p=3	0.98

در ادامه به دلیل انطباق ماهیت داده‌های مرتبط با مفاهیم زمانبندی انجام وظایف با ویژگیهای داده‌های مورد تمرکز این مقاله، روش پیشنهادی بر روی دیتاست شبیه‌ساز CloudSim نیز پیاده سازی شد. CloudSim بستری است که اخیرا برای شبیه‌سازی موارد مربوط به رایانش ابری، محاسبه تأخیرهای زمانی و زمانبندی انجام وظایف مورد توجه محققان است. داده مورد نیاز از طریق ارتباط با یکی از نویسندگان مرجع [۳۲] دریافت گردید. این داده‌ها شامل ویژگی‌ها و اطلاعات متنوعی بود که برخی از ویژگی‌های مهم آن عبارت بودند از:

- توپولوژی شبکه: اطلاعات مربوط به توپولوژی شبکه که نشان می‌دهد که چگونه سرورها و ابرها به یکدیگر متصل شده‌اند و چگونه ترافیک در شبکه جریان می‌یابد.
- سرورها و ماشین‌های مجازی: مشخصات سرورها و ماشین‌های مجازی است که نماینده‌های منابع محاسباتی در محیط ابری هستند. که شامل تعداد هسته‌های پردازشی، ظرفیت حافظه، دیسک، پهنای باند شبکه و سایر ویژگی‌های مربوط به سخت‌افزار است.
- کاربران و درخواست‌ها: اطلاعات کاربران و درخواست‌هایی است که آنها برای استفاده از خدمات ابری ارسال می‌کنند که شامل

نوع درخواست، زمانبندی، مقدار منابع درخواستی و سایر ویژگی‌های مرتبط با کاربران است.

- الگوهای ترافیک: الگوهای ترافیک مختلف است که کاربران در حین استفاده از سرویس‌های ابری تولید می‌کنند که شامل الگوهای مصرف منابع، الگوهای تغییرات ترافیک و الگوهای بارگذاری مختلف باشند.
- زمانبندی و رویدادها: اطلاعات مربوط به زمانبندی فعالیت‌ها و رویدادهای مختلف در محیط ابری است که شامل زمانبندی درخواست‌ها، زمانبندی تغییرات ترافیک، زمانبندی ایجاد و حذف سرورها و ماشین‌های مجازی و سایر رویدادهای مرتبط است.

مرجع [۳۳] تاریخچه کاملی از انواع شبیه‌سازهای حوزه‌های اینترنت اشیاء، رایانش‌های ابری، لبه و مه را معرفی کرده است که از جمله آنها می‌توان به iFogSim اشاره کرد [۳۴].

پس از انجام آزمایشها، مشاهده شد که نتایج به دست آمده همروندی و تشابه زیادی با نتایج بدست آمده بر روی مجموعه داده قبلی را دارد. از اینرو از ذکر آنها در این مقاله خودداری می‌شود. این نتایج بیانگر کارایی و کاربردی بودن روش پیشنهاد شده در این مقاله می‌باشد.

۶. بحث و نتیجه‌گیری

دراین مقاله روشی مبتنی بر خوشه‌بندی فازی‌چندهسته‌ای برای خوشه‌بندی کلان‌داده‌ها بوسیله مدل نگاشت‌کاهش هدوپ ارائه گردید. همپوشانی خوشه‌ها و قدرت تعمیم منطق فازی در مواجهه با داده‌های نویزدار و پرت دلیل اصلی توجه و تمرکز این مقاله به خوشه‌بندی فازی بوده است. کارهای انجام شده در این زمینه معرفی‌گردید و نقاط برتری آنها عنوان شد. در ادامه روش پیشنهادی در چارچوب نگاشت‌کاهش هدوپ معرفی شد. هدوپ ما را قادر ساخت تا به جای تعامل با سیستم عامل و پردازنده، با یک کلاستر منطقی از پردازش‌ها و گره‌های انباره داده تعامل داشته باشیم. دو جزء مهم هدوپ HDFS که از پتابایت داده پشتیبانی می‌کند و نگاشت‌کاهش توسعه‌پذیر که نتایج را به صورت دسته‌ای محاسبه می‌کند می‌باشند. برای تشخیص خوشه‌های خطی‌جدایی‌ناپذیر در ساختارهای پیچیده کلان داده‌ای، تابع هدف پیشنهادی در معماری یادگیری چندهسته‌ای در نظرگرفته شد و در چارچوب نگاشت‌کاهش هدوپ پیاده‌سازی گردید. روش پیشنهادی با تنظیم خودکار وزن هسته‌ها، و در نظرگرفتن عبارت جریمه، نسبت به توابع نامناسب یا ویژگی‌های نامرتبط ایمن‌گردید. اینکار سبب کاهش حساسیت روش پیشنهادی به انتخاب هسته ناکارآمد گردید. در ادامه کارایی روش‌پیشنهادی در مقایسه با ۳ روش خوشه‌بندی KMeans، MPC-KMeans و سلسله مراتبی بررسی‌گردید. برای انجام آزمایشها از دو مجموعه داده استفاده گردید. در ابتدا دیتاست مورد استفاده، مجموعه داده کیفیت نوشیدنی از مخزن داده یادگیری ماشین UCI انتخاب گردید. UCI مجموعه‌ای از پایگاه‌های داده، تئوری‌های دامنه و تولیدکننده‌های داده است که مورد توجه بسیاری از محققان این حوزه می‌باشد و این مجموعه داده یکی از ده مجموعه داده برتر پرتواتر توصیه شده جهت عملیات رگرسیون و خوشه‌بندی می باشد. مجموعه داده دوم، مجموعه داده شبیه‌ساز CloudSim بود که از طریق ارتباط با یکی از نویسندگان مرجع معرفی شده دریافت گردید. CloudSim بستری است که اخیرا برای شبیه‌سازی موارد مربوط به رایانش ابری، محاسبه تأخیرهای زمانی و زمانبندی انجام وظایف مورد توجه محققان است. نتایج به دست آمده از آزمایشها بر روی هر دو مجموعه داده از همروندی و تشابه زیادی برخودار بودند که خود دلالت بر کارایی و کاربردی بودن روش پیشنهاد شدی این مقاله دارد.

تعیین مقدار مناسب برای پارامتر کنترل‌کننده درجه فازی خوشه بندی m و پارامتر وزن درجه عضویت امکان‌پذیری p، هنوز به عنوان یک مساله باز در حوزه خوشه‌بندی فازی مطرح است. از اینرو در این مقاله آزمایش‌هایی نیز در خصوص تعیین مقادیر مناسب برای این پارامترها انجام گرفت. نتایج حاصله بیانگر این موضوع است که فازی‌تر کردن خوشه‌بندی نتایج بهتری را ارائه می‌دهد. در آزمایش‌های انجام‌گرفته کارایی بالای روش پیشنهادی در مقایسه با سایر روش‌های خوشه‌بندی KMeans، MPC-KMeans و سلسله مراتبی بسیار مشهود بود که در بخش قبلی مورد تحلیل گردید. از مهمترین دلایل روش پیشنهادی در خوشه بندی کلان داده‌ها، به توانایی مدل‌سازی روابط غیرخطی داده‌ها با استفاده از مدل یادگیری چندهسته‌ای، تعیین مقادیر مناسب برای پارامترهای فازی‌سازی و امکان‌پذیری، و ارائه الگوریتم در مدل نگاشت‌کاهش می‌توان اشاره کرد.

۷. پیشنهادهایی برای کارهای آینده

در پایان موارد زیر به عنوان پیشنهادهایی برای کارهای آینده ارائه می‌گردد:

۱- از آنجا که رویکرد پیشنهادی مبتنی بر یادگیری متریک خطی است یکی از محدودیت‌های مدل می‌تواند شرایطی باشد که در آن فضای مساله خطی نباشد. در این شرایط با استفاده از کرنل

های غیرخطی یا روش های متریک خطی چندگانه می‌توان مساله را حل کرد.

۲- وجود داده‌های دارای نویز و یا داده‌های پرت می‌تواند در عملکرد مدل پیشنهادی موثر باشد که استفاده از انواع روش های کشف اینگونه از داده‌ها و یا مدل‌هایی که مقاومت بیشتری نسبت به این گونه داده‌ها دارند و یا تغییر برخی از هایپر پارامترهای مدل می‌توان این مشکل را تا حدودی برطرف کند.

۳- از آنجا که تعیین مناسب مرکز برای خوشه‌ها مساله مهمی است، استفاده از بعضی روش‌های متاهیوریستیک مثل الگوریتم ژنتیک یا ازدحام ذرات می‌تواند در تعیین بهینه یا سریعتر مرکز خوشه‌ها کمک بسزایی داشته باشد.

۴- استفاده از سایر معیارهای شباهت به جای فاصله اقلیدسی مثل معیار فاصله ماهالانوبیس یا منهتن ممکن است در برخی از مسایل منجر به نتایج بهتری در عملکرد مدل شود.

۵- همانطور که پیشتر دیدیم، تعیین نوع و پارامترهای مناسب برای هسته از جمله مواردی است که نقش مهمی در کارایی این مدل دارد. لذا جستجوی مناسب در فضای مربوط به این مقادیر و یافتن مقدار بهینه برای آنها از جمله مسائلی است که می‌تواند در کارهای پیش رو مد نظر قرار بگیرد.

مراجع

[1] S.M. Razavi, M. Kashani, S. Paydar, **"**Big Data Fuzzy C-Means Algorithm based on Bee Colony Optimization using an Apache Hbase", Journal of Big Data, Vol. 8, Article Number: 64, 2021.

[2] S. Sinha, "Hadoop Ecosystem: Hadoop Tools for Crunching Big Data", edureka, https://www.edureka.co/blog/hadoop-ecosystem, 2022.

[3] S. Landest, T. khoshgoftaar, A.N. Richter, "A Survey of Open Source Tools for Machine Learning with Big Data in the Hadoop Ecosystem**"**, Journal of Big Data, Vol. 2, No.1, 2015.

[4] X. Liu, X. Zhu, M. Li, L. Wang, E. zhu, T. Liu, M. Kloft, D. Shen, J. Yin, W. Gao**,** **"**Multiple Kernel k-Means with Incomplete Kernels**"**, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 42, No. 5, pp.1191-1204, 2020.

[5] R. K. Sanodiya, S. Saha, J. Mathew, "A Kernel Semi-Supervised Distance Metric Learning with Relative Distance: Integration with a MOO Approach**"**, Expert Systems with Applications, Elsevier, Vol. 125, pp. 233-248, 2019.

[6] M. Soleymani Baghshah, S. Bagheri Shouraki, "Efficient Kernel Learning from Constraints and Unlabeled Data**"**, 20th International Conference on Pattern Recognition, Istanbul, Turkey, pp. 3364-3367, 2010.

[7] S. Zhu, D. Wang, T. Li, "Data Clustering with Size Constraints**"**, Knowledge-Based Systems, Elsevier, Vol. 23, pp. 883-889, 2010.

[8] L. A. Maraziotis, "A Semi-Supervised Fuzzy Clustering Algorithm Applied to Gene Expression Data", Pattern Recognition, Elsevier, Vol. 45, pp. 637-648, 2014.

[9] J. Bezdek, R. Ehrlich, W. Full, **"**FCM: the Fuzzy C-Means Clustering Algorithm", Computers & Geosciences, Elsevier Vol. 10, Issue. 2-3, pp. 191-203, 1984.

[10] O. Ozdemir, A. Kaya, "Comparison of FCM, PCM, FPCM and PFCM Algorithms in Clustering Methods", Afyon Kocatepe University Journal of Science and Engineering, pp. 92-102, 2019**.**

[11] M. A. Lopez Felip, T. J. Davis, T. D. Frank, J.A. Dixon, **"**A Cluster Phase Analysis for Collective Behavior in Team Sports", Human Movement Science, Elsevier, Vol. 59, pp. 96-111, 2018.

[12] F. Hai Jun, W. Xiao Hong, M. Han Ping, W. Bin, **"**Fuzzy Entropy Clustering using Possibilistic Approach", Advanced in Control Engineering and Information Science, Elsevier, Procedia Engineering Vol. 15, pp.1993-1997, 2011.

[13] M. Bouzbida, L. Hassine, A. Chaari, "Robust Kernel Clustering Algorithm for Nonlinear System Identification", Hindawi, Mathematical Problems in Engineering, pp. 1-11, 2017.

[14] T.H. Sardar, Z. Ansari, **"**MapReduce-based Fuzzy C-means Algorithm for Distributed Document Clustering", Journal of The

Learning Approach**”**, Pattern Recognition, Elsevier, Vol. 48, Issue. 3, pp. 935-967, 2015.

[24] H. Hassani, M. Kalantari, C. Beneki, “Comparative Assessment of Hierarchical Clustering Methods for Grouping in Singular Spectrum Analysis”, AppliedMath, Vol. 1, No.1, pp. 18-36, 2021.

[25] S.A. Elavarasi, D.J. Akilandeswari, D.B. Sathiyabhama, **“**A Survey on Partition Clustering Algorithms**”**, International Journal of Enterprise Computing and Business Systems, Vol.1, pp.1–14, 2011.

[26] T. Zhang, R. Ramakrishnan, M. Livny, **“**Birch: an Efficient Data Clustering Method for Very Large Databases”, SIGMOD Record, Vol.25, No.2, pp.103–114, 1996.

[27] Top 10 Regression Datasets and Projects, https://www.interviewquery.com/p/regression-datasets-and-projects.

[28] 10 Open Datasets for Linear Regression, https://www.telusinternational.com/articles/10-open-datasets-for-linear-regression, 2021.

[29] UCI, Machine Learning Repository, Center for Machine Learning and Intelligent Systems, “Win Quality Data Set”, https://archive.ics.uci.edu/ml/datasets/wine+quality, Site visit: 2023.

[30] UCI, Machine Learning Repository, **“**Center for Machine Learning and Intelligent Systems”, University of California, School of Information and Computer Science: Irvine, CA, USA, https://archive.ics.uci.edu/, Site Visit: 2023.

[31] C. Swafford, **“**Red Wine Quality Analysis”, https://rpubs.com/cswaff7/775970, 2021.

[32] M. A. Ala’anzy, M. Othman, Z. M. Hanapi, M. A. Alrshah, “Locust Inspired Algorithm for Cloudlet Scheduling in Cloud Computing Environments**”**, Sensors, Vol. 21, No. 21, 19 Pages, 2021.

[33] R. Mahmud, S. Pallewatta, M. Goudarzi, R. Buyya, “iFogSim2: An Extended iFogSim Simulator for Mobility, Clustering, and Microservice Management in Edge and Fog Computing Environments”, The University Institution of Engineers (India): Series B, Vol. 103, No. 1, pp.131-142, 2022.

[15] Q. Yu, Z. Ding, **“**An Improved Fuzzy C-Means Algorithm based on MapReduce”**,** 8th International Conference on Biomedical Engineering and Informatics (BMEI), pp. 634-638, 2015.

[16] J. Dean, S. Ghemawat, “MapReduce: Simplified Data Processing on Large Clusters”, Sixth Symposium on Operating System Design and Implementation, San Francisco, CA, pp. 137-150, 2004.

[17] L. Jiamin and F. Jun, “A Survey of MapReduce based Parallel Processing Technologies”, China Communications, Vol. 11, No. 14, pp. 146–155, 2014.

[18] W. Zhao, H. Ma, Q. He, **“**Parallel K-Means Clustering based on MapReduce, in Cloud Computing”, IEEE International Conference on Cloud Computing, pp. 674-679, Part of the Lecture Notes in Computer Science book series (LNCS, volume 5931), 2009.

[19] H. Bei, Y. Mao, W. Wang, X. Zhang, **“**Fuzzy Clustering Method Based on Improved Weighted Distance”, Mathematical Problem in Engineering, Vol. 5, Hindawi, 2021.

[20] S.A.Ludwig, “MapReduce-based Fuzzy C-Means Clustering Algorithm: Implementation and Scalability”, International Journal of Machine Learning and Cybernetics, pp. 923-934, Copyright owner: Springer-Verlag Berlin Heidelberg, 2015.

[21] J. Ramisingh, V. Bhuvaneswari, **“**An Integrated Multi-Node Hadoop Framework to Predict High-Risk Factors of Diabetes Mellitus using a Multilevel MapReduce based Fuzzy Classifier (MMR-FC) and Modified DBSCAN Algorithm”, Applied Soft Computing, Vol. 108, 2021.

[22] T.H. Sardar, Z. Ansari, **“**Partition based clustering of large datasets using MapReduce framework: An analysis of recent themes and directions”, Future Computing and Informatics Journal, Vol. 3, No. 2, pp. 247-261, 2018.

[23] A. A. Abin, H. Beigy, “Active Constrained Fuzzy Clustering: A Multiple Kernels

of Melbourne, Journal of Systems and Software, Vol. 190, 2022.

[34] H. Gupta, A.V. Dastjerdi, S.K. Ghosh, R. Buyya, "iFogSim: A toolkit for Modeling and Simulation of Resource Management Techniques in the Internet of Things, Edge and Fog Computing Environments", Cloud and Fog Computing, Volume 47, Issue 9, Pages 1275-1296, 2017.

پیوست- اثبات همگرایی تابع هدف

در اینجا هدف نهایی یافتن همزمان وزن‌های $w \equiv [w_k]_{M\times 1}$، درجه‌های عضویت امکان‌پذیری $T \equiv [t_{ci}]_{C\times N}$ ، درجه‌های عضویت فازی $U \equiv [u_{ci}]_{C\times N}$ و نمایندگان خوشه $V \equiv [u_c]_{L\times C}$ است به‌گونه‌ای که تابع هدف کمینه شود. از راهکار بهینه‌سازی متناوب به‌منظور کمینه‌سازی $J_{Proposed\ method}$ استفاده می‌شود. بدین منظور یک تابع انرژی جدید تعریف می‌شود که به ازای هر شرط $\sum_{c=1}^{C} u_{ci} = 1$ متغیر لاگرانژ λ_i و به ازای شرط $\sum_{k=1}^{M} wk = 1$ متغیر لاگرانژ β را در نظر می‌گیرد. به منظور ساده سازی فرمول‌ها از D_{ci} برای نمایش فاصله بین داده x_i و نماینده خوشه v_c استفاده می‌شود که داریم:

$(D_{ci}^2 = (\varphi(x_i) - v_c)^T(\varphi(x_i) - v_c))$. بنابراین از تابع هدف تابع لاگرانژین زیر نتیجه می‌شود.

$$J_{Proposed\ method}(w,T,U,V,\lambda,\beta) = \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}^m t_{ci}^p + \sum_{c=1}^{C}\mu_c \sum_{i=1}^{N} u_{ci}^m (1-t_{ci})^p + \alpha\left(\sum_{(i,j)\in M}\sum_{c=1}^{C}\sum_{\substack{l=1\\ l\neq c}}^{C} u_{ci}^m u_{lj}^m t_{ci}^p t_{lj}^p + \sum_{(i,j)\in C}\sum_{c=1}^{C} u_{ci}^m u_{cj}^m t_{ci}^p t_{cj}^p\right) + \sum_{i=1}^{N}\lambda_i\left(\sum_{C=1}^{C} u_{ci} - 1\right) + 2\beta\left(\sum_{K=1}^{M} w_k - 1\right) \quad (۱)$$

کمینه‌سازی تابع انرژی (۱) مقدار کمینه تابع هدف $J_{Proposed\ method}$ را نتیجه می‌دهد. بهینه‌سازی درجه‌های عضویت امکان پذیری $T \equiv [t_{ci}]_{C\times N}$ ، درجه‌های عضویت فازی $U \equiv [u_{ci}]_{C\times N}$ و وزن‌های $w \equiv [w_k]_{M\times 1}$ توسط سه لم زیر بیان می‌شود.

لم ۱: اگر مقادیر وزن های $w \equiv [w_k]_{M\times 1}$ ، درجه های عضویت فازی $U \equiv [u_{ci}]_{C\times N}$ و نمایندگان خوشه $V \equiv [u_c]_{L\times C}$ ثابت باشند آنگاه مقدار بهینه درجه‌های عضویت امکان پذیری $T \equiv [t_{ci}]_{C\times N}$ برابر است با:

$$t_{ci} = \frac{1}{1+\left(\frac{D_{ci}^2+\alpha\left(s_{ci}^M+S_{ci}^c\right)}{\mu_c}\right)^{\frac{1}{p-1}}} \quad \forall c,i \quad (۲)$$

اثبات: برای یافتن مقادیر بهینه درجه‌های امکان‌پذیری T، ابتدا وزن‌های w، درجه‌های عضویت فازی U و نمایندگان خوشه V را ثابت در نظر گرفته می‌شود. با فرض مقادیر ثابت برای وزن‌ها، درجه‌های عضویت فازی و نمایندگان خوشه، مقادیر فاصله‌ها نیز ثابت خواهند بود. با مشتق‌گیری تابع لاگرانژین (۱) نسبت به درجه های عضویت امکان‌پذیری و قراردادن آن برابر با صفر، به ازای هر درجه عضویت امکان‌پذیری t_{ci} داریم:

$$\frac{\partial J(w,T,U,\lambda,\beta)}{\partial t_{ci}} = p u_{ci}^m t_{ci}^{p-1} D_{ci}^2 - \mu_c p u_{ci}^m (1-t_{ci})^{p-1} + \alpha\left[p u_{ci}^m t_{ci}^{p-1}\left[\underbrace{\sum_{(i,j)\in M}\sum_{\substack{l=1\\ l\neq 1}}^{C} u_{lj}^m t_{lj}^p}_{s_{ci}^M} + \underbrace{\sum_{(i,j)\in C} u_{cj}^m t_{cj}^p}_{s_{ci}^c}\right]\right] = 0 . \quad (۳)$$

با ساده سازی جبری رابطه بالا خواهیم داشت :

$$\mu_c(1-t_{ci})^{p-1} = t_{ci}^{p-1}\left(D_{ci}^2 + \alpha\left(s_{ci}^M + S_{ci}^c\right)\right) \quad (۴)$$

بنابراین

$$\left(\frac{1-t_{ci}}{t_{ci}}\right)^{p-1} = \frac{D_{ci}^2+\alpha\left(s_{ci}^M+S_{ci}^c\right)}{\mu_c} \quad (۵)$$

و

$$\frac{1}{t_{ci}} = 1 + \left(\frac{D_{ci}^2+\alpha\left(s_{ci}^M+S_{ci}^c\right)}{\mu_c}\right)^{\frac{1}{p-1}} \quad (۶)$$

$\Rightarrow$

از اینرو جواب بهینه برای t_{ci} به صورت زیر حاصل می‌گردد.

$$t_{ci} = \frac{1}{1+\left(\frac{D_{ci}^2+\alpha\left(s_{ci}^M+S_{ci}^c\right)}{\mu_c}\right)^{\frac{1}{p-1}}} \quad (۷)$$

و بدین ترتیب لم اثبات می‌گردد.

لم ۲: اگر مقادیر وزن‌های $w \equiv [w_k]_{M\times 1}$، درجه‌های عضویت امکان‌پذیری $T \equiv [t_{ci}]_{C\times N}$ و نمایندگان خوشه $V \equiv [u_c]_{L\times C}$ ثابت باشند آنگاه مقدار بهینه درجه‌های عضویت فازی $U \equiv [u_{ci}]_{C\times N}$ برابر است با :

$$u_{ci} = \frac{1}{\sum_{k=1}^{C}\left(\frac{t_{ci}^{p-1}\left(D_{ci}^2+\alpha\left(s_{ci}^M+S_{ci}^c\right)\right)}{t_{ki}^{p-1}\left(D_{ki}^2+\alpha\left(s_{ki}^M+S_{ki}^c\right)\right)}\right)^{\frac{1}{m-1}}} \quad \forall c,i \quad (۸)$$

اثبات: برای یافتن مقادیر بهینه درجه‌های فازی U، ابتدا وزن‌های w، درجه‌های عضویت امکان‌پذیری T و نمایندگان خوشه V را ثابت در نظر می‌گیریم. با فرض مقادیر ثابت برای وزن‌ها، درجه های عضویت امکان‌پذیری و نمایندگان خوشه، مقادیر فاصله‌ها نیز ثابت خواهند بود. با مشتق‌گیری از تابع لاگرانژین (۱) نسبت به

درجه‌های عضویت فازی و قرار دادن آن برابر با صفر، به ازای هر درجه عضویت فازی u_{ci} داریم:

$$\frac{\partial J(w,T,U,\lambda,\beta)}{\partial u_{ci}} = m u_{ci}^{m-1} t_{ci}^{p} D_{ci}^{2} + \mu_c m u_{ci}^{m-1}(1-t_{ci})^{p} + \alpha\left[m u_{ci}^{m-1} t_{ci}^{p}\left[\underbrace{\sum_{(i,j)\in M}\sum_{\substack{l=1\\ l\neq i}}^{C} u_{lj}^{m} t_{lj}^{p}}_{s_{ci}^{M}} + \underbrace{\sum_{(i,j)\in C} u_{cj}^{m} t_{cj}^{p}}_{s_{ci}^{C}}\right]\right] - \lambda_i = 0. \quad (۹)$$

با ساده‌سازی رابطه بالا خواهیم داشت:

$$m u_{ci}^{m-1}\left(t_{ci}^{p-1}\left(D_{ci}^{2} + \alpha(s_{ci}^{M} + S_{ci}^{c})\right) + \mu_c(1-t_{ci})^{p}\right) = \lambda i \quad (۱۰)$$

و

$$\Rightarrow u_{ci} = \left(\frac{\lambda i}{m\left(t_{ci}^{p}\left(D_{ci}^{2}+\alpha(s_{ci}^{M}+S_{ci}^{c})\right)+\mu_c(1-t_{ci})^{p}\right)}\right)^{\frac{1}{m-1}} \quad (۱۱)$$

از طرف دیگر از رابطه (۴) داریم:

$$\mu_c(1-t_{ci})^{p} = t_{ci}^{p-1}(1-t_{ci})\left(D_{ci}^{2} + \alpha\left(s_{ci}^{M} + S_{ci}^{c}\right)\right) \quad (۱۲)$$

رابطه (۱۱) با استفاده از رابطه (۱۲) به صورت زیر ساده می‌شود.

$$u_{ci} = \left(\frac{\lambda i}{m\left(t_{ci}^{p}\left(D_{ci}^{2} + \alpha(s_{ci}^{M} + S_{ci}^{c})\right) + t_{ci}^{p-1}(1-t_{ci})\left(D_{ci}^{2} + \alpha(s_{ci}^{M} + S_{ci}^{c})\right)\right)}\right)^{\frac{1}{m-1}} \quad (۱۳)$$

با ساده‌سازی رابطه بالا جواب بهینه برای u_{ci}را به‌صورت زیر داریم:

$$u_{ci} = \left(\frac{\lambda i}{m\left(t_{ci}^{p-1}\left(D_{ci}^{2}+\alpha(s_{ci}^{M}+S_{ci}^{c})\right)\right)}\right)^{\frac{1}{m-1}} \quad (۱۴)$$

با استفاده از شرط $\sum_{k=1}^{C} u_{ki} = 1$ می توان λ را به‌صورت زیر از رابطه بالا حذف نمود.

$$\sum_{k=1}^{C} u_{ki} = \sum_{k=1}^{C}\left(\frac{\lambda i}{m\left(t_{ci}^{p-1}\left(D_{ci}^{2}+\alpha(s_{ci}^{M}+S_{ci}^{c})\right)\right)}\right)^{\frac{1}{m-1}} = 1 \quad (۱۵)$$

$$\left(\frac{\lambda i}{m}\right)^{\frac{1}{m-1}} = \frac{1}{\sum_{k=1}^{C}\left(\frac{1}{t_{ki}^{p-1}\left(D_{ki}^{2}+\alpha(s_{ki}^{M}+s_{ki}^{c})\right)}\right)^{\frac{1}{m-1}}} \quad (۱۶)$$

با جایگزینی رابطه (۱۶) در رابطه (۱۴) می‌توان جواب بسته برای درجه‌های عضویت فازی بهینه را به‌صورت زیر بدست آورد.

$$u_{ci} = \frac{1}{\sum_{k=1}^{C}\left(\frac{t_{ci}^{p-1}\left(D_{ci}^{2}+\alpha(s_{ci}^{M}+s_{ci}^{c})\right)}{t_{ki}^{p-1}\left(D_{ki}^{2}+\alpha(s_{ki}^{M}+s_{ki}^{c})\right)}\right)^{\frac{1}{m-1}}} \quad (۱۷)$$

که اثبات لم کامل می‌گردد.

در لم های ۱ و ۲ مقادیر بهینه درجه‌های عضویت امکان‌پذیری T و درجه‌های عضویت فازی U با فرض ثابت‌بودن مقادیر وزن‌های w و نمایندگان خوشه V بدست آمده است. لم زیر مقادیر بهینه وزن‌ها را برای ترکیب هسته‌ها نتیجه می‌دهد.

لم ۳: اگر مقادیر درجه‌های عضویت امکان‌پذیری $T \equiv [t_{ci}]_{C\times N}$ و درجه‌های عضویت فازی $U \equiv [u_{ci}]_{C\times N}$ ثابت باشند آنگاه مقدار بهینه وزن‌های $w \equiv [w_k]_{M\times 1}$ برابر است با:

$$W_k = \frac{\frac{1}{y_k}}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_M}} \quad \forall k. \quad (۱۸)$$

اثبات: برای یافتن مقادیر بهینه نمایندگان خوشه و وزن‌ها، ابتدا درجه‌های عضویتw، امکان‌پذیری T و درجه عضویت فازیU ثابت فرض می‌شوند. با مشتق‌گیری از تابع لاگرانژین (۱) نسبت به v_c و قراردادن آن برابر با صفر خواهیم داشت:

$$\frac{\partial J(w,T,U,\lambda,\beta)}{\partial v_c} = -2\sum_{i=1}^{N} u_{ci}^{m} t_{ci}^{p}(\varphi(x_i) - v_c) = 0 \quad (۱۹)$$

از آنجا که مقادیر T و U معلوم است می توان مقادیر بهینه v_c را به صورت رابطه بسته زیر محاسبه نمود.

$$v_c = \frac{\sum_{i=1}^{N} u_{ci}^{m} t_{ci}^{p}\varphi(x_i)}{\sum_{i=1}^{N} u_{ci}^{m} t_{ci}^{p}} \quad (۲۰)$$

از آنجا که این نمایندگان خوشه در فضای ویژگی هستند و ممکن است دارای بعد بی‌نهایت باشند لذا محاسبه مستقیم نمایندگان خوشه‌ها امکان‌پذیر نمی‌باشد. خوشبختانه در بهینه‌سازی $J_{\text{Proposed method}}(w,T,U,V,\lambda,\beta)$ می‌توان مقادیر درجه‌های عضویت امکان‌پذیری، درجه‌های عضویت فازی و وزن‌های بهینه را بدون محاسبه نمایندگان خوشه‌ها تعیین نمود که در ادامه بیان می‌شود. بنابراین سعی می‌شود تا نیاز به محاسبه مستقیم نمایندگان خوشه v_c را در تابع انرژی $J_{\text{Proposed method}}(w,T,U,V,\lambda,\beta)$ برطرف شود. همانگونه که پیشتر بیان شد فاصله داده x_iو نماینده خوشه v_c در فضای ویژگی به‌صورت زیر محاسبه می‌شود:

$$\begin{aligned} D_{ci}^{2} &= (\varphi(x_i) - v_c)^{T}(\varphi(x_i) - v_c) \\ &= \varphi(x_i)^{T}\varphi(x_i) - 2\varphi(x_i)^{T}v_c + {v_c}^{T}v_c \\ &= \varphi(x_i)^{T}\varphi(x_i) - 2\varphi(x_i)^{T}\left(\frac{\sum_{j=1}^{N} u_{cj}^{m} t_{cj}^{p}\varphi(x_j)}{\sum_{j=1}^{N} u_{cj}^{m} t_{cj}^{p}}\right) \\ &+ \left(\frac{\sum_{r=1}^{N} u_{cr}^{m} t_{cr}^{p}\varphi(x_r)}{\sum_{r=1}^{N} u_{cr}^{m} t_{cr}^{p}}\right)^{T}\left(\frac{\sum_{s=1}^{N} u_{cs}^{m} t_{cs}^{p}\varphi(x_s)}{\sum_{s=1}^{N} u_{cs}^{m} t_{cs}^{p}}\right) \\ &= \sum_{k=1}^{M} w_k^{2} k_k(x_i,x_i) - \frac{2\sum_{j=1}^{N}\sum_{k=1}^{M} w_k^{2} u_{cj}^{m} t_{cj}^{p} k_k(x_i,x_i)}{\sum_{j=1}^{N} u_{cj}^{m} t_{cj}^{p}} \\ &+ \frac{\sum_{r=1}^{N}\sum_{s=1}^{N}\sum_{k=1}^{M} w_k^{2} u_{cr}^{m} t_{cr}^{p} u_{cs}^{m} t_{cs}^{p} k_k(x_r,x_s)}{\left(\sum_{r=1}^{N} u_{cr}^{m} t_{cr}^{p}\right)\left(\sum_{s=1}^{N} u_{cs}^{m} t_{cs}^{p}\right)} \end{aligned}$$

$$= \sum_{k=1}^{M} w_k^2 \underbrace{\left(k_k(x_i,x_i) - \frac{2\sum_{j=1}^{N} u_{cj}^m t_{cj}^p k_k(x_i,x_i)}{\sum_{j=1}^{N} u_{cj}^m t_{cj}^p} + \frac{\sum_{r=1}^{N}\sum_{s=1}^{M} u_{cr}^m t_{cr}^p u_{cs}^m t_{cs}^p k_k(x_r,x_s)}{\left(\sum_{r=1}^{N} u_{cr}^m t_{cr}^p\right)\left(\sum_{s=1}^{N} u_{cs}^m t_{cs}^p\right)} \right)}_{\vartheta_{ci}^k}$$

$$= \sum_{k=1}^{M} w_k^2 \vartheta_{ci}^k \qquad (۲۱)$$

عدم نیاز به محاسبه مستقیم نمایندگان خوشه v_c در محاسبه D_{ci} در رابطه (۲۱) نشان داده شده است. از اینرو تابع انرژی رابطه (۱) به شکل زیر بازنویسی می‌شود.

$$J_{\text{Proposed method}}(w,T,U,V,\lambda,\beta) = \sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}^m t_{ci}^p \sum_{k=1}^{M} w_k^2 \vartheta_{ci}^k + \sum_{c=1}^{C} \mu_c \sum_{i=1}^{N} u_{ci}^m (1-t_{ci})^p + \alpha\left(\sum_{(i,j)\in M}\sum_{c=1}^{C}\sum_{\substack{l=1\\ l\neq c}}^{C} u_{ci}^m u_{lj}^m t_{ci}^p t_{lj}^p + \sum_{(i,j)\in C}\sum_{c=1}^{C} u_{ci}^m u_{cj}^m t_{ci}^p t_{cj}^p\right) + \sum_{i=1}^{N}\lambda_i\left(\sum_{C=1}^{C} u_{ci} - 1\right) + 2\beta\left(\sum_{K=1}^{M} w_k - 1\right) \qquad (۲۲)$$

با فرض ثابت بودن درجه‌های عضویت امکان‌پذیری و فازی، و مشتق‌گیری از رابطه (۲۲) نسبت به w_k و برابر با صفر قرار دادن آن داریم:

$$J_{\text{Proposed method}}(w,T,U,V,\lambda,\beta) = 2\left(\underbrace{\sum_{c=1}^{C}\sum_{i=1}^{N} u_{ci}^m t_{ci}^p \vartheta_{ci}^k}_{y_k}\right) w_k - 2\beta = 0 \qquad (۲۳)$$

$$\Rightarrow w_k = \frac{\beta}{y_k} \qquad (۲۴)$$

از آنجا که $\sum_{K=1}^{M} w_k = 1$ لذا داریم:

$$\sum_{K=1}^{M} w_k = \beta\left(\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_M}\right) = 1 \qquad (۲۵)$$

و

$$\Rightarrow \beta = \frac{1}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_M}} \qquad (۲۶)$$

با جایگذاری رابطه (۲۶) در رابطه (۲۴) می‌توان مقادیر وزن بهینه را به‌صورت زیر بدست آورد:

$$W_k = \frac{\frac{1}{y_k}}{\frac{1}{y_1} + \frac{1}{y_2} + \cdots + \frac{1}{y_M}} \qquad (۲۷)$$

و بدین ترتیب اثبات لم تکمیل می شود.

با استفاده از لم‌های ۱ ، ۲ و ۳ می‌توان همگرایی روش معرفی شده را توسط سه لم زیر نتیجه گرفت:

لم ۴ : فرض کنید کنید $\tau(T) = J_{\text{Proposed method}}$ باشد که در آن $T \equiv [t_{ci}]_{C\times N}$ ، $\sum_{c=1}^{C} u_{ci} = 1$ ثابت بوده و شرط $U \equiv [u_{ci}]_{C\times N}$ (به ازای $i=1,2,\ldots,N$) را برآورده می‌کند. همچنین $w \equiv [w_k]_{M\times 1}$ ثابت بوده و به ازای $1 \le c \le C$ و $1 \le i \le N$ نامساوی های $D_{ci}^2 > 0$ ، $m > 1$ ، $p > 1$ برقرار است. آنگاه T بهینه محلی $\tau(T)$ است اگر و فقط اگر t_{ci} (به ازای $c=1,2,\ldots,C$ و $i=1,2,\ldots,N$) طبق رابطه اول قضیه محاسبه گردد.

اثبات: شرط لازم برای بهینگی محلی در لم ۱ ثابت شده است. برای اثبات کافی است، ماتریس هسیان $H(\tau(T))$ از $\tau(T)$ با استفاده از رابطه (۲۱) به صورت زیر محاسبه شود.

$$h_{fg.ci}(T) = \frac{\partial}{\partial t_{.fg}}\left[\frac{\partial \tau(T)}{\partial t_{.ci}}\right] = \begin{cases} p(p-1)u_{ci}^m\left(t_{ci}^{p-2}D_{ci}^2 + \mu_c(1-t_{ci})^{p-2} + \alpha t_{ci}^{p-2}\left(s_{ci}^M + s_{ci}^C\right)\right), \\ \quad \text{if } f = c, g = i \\ \text{در غیر اینصورت } 0 \end{cases} \qquad (۲۸)$$

در رابطه (۲۸) ، $h_{fg.ci}(T)$ یک ماتریس قطری است که به ازای تمام $1 \le c \le C$ و $1 \le i \le N$ ، مقادیر t_{ci} و u_{ci} به کمک رابطه‌های دوم و سوم قضیه محاسبه می‌شوند.

همچنین داریم $0 < t_{ci} < 1$ ، $u_{ci} > 0$ ، $m > 1$ ، $D_{ci}^2 > 0$ ، $\mu_c > 0$ و $\alpha > 0$. با توجه به موارد ذکر شده نتیجه می‌شود که ماتریس هسیان فوق یک ماتریس مثبت معین است. بنابراین داریم:

$$J_{\text{Proposed method}}(U^T,T^{T+1},W^T) \le J_{\text{Proposed method}}(U^T,T^T,W^T)$$

و بدین ترتیب شرط کافی رابطه اول قضیه برای کمینگی محلی $\tau(T)$ اثبات می‌شود.

لم ۵ : فرض کنید $\tau(U) = J_{\text{Gol Function}}$ باشد که در آن $U \equiv [u_{ci}]_{C\times N}$ و شرط $\sum_{c=1}^{C} u_{ci} = 1$ (به ازای $i=1,2,\ldots,N$) را برآورده می‌کند، $T \equiv [t_{ci}]_{C\times N}$ ثابت بوده، $w \equiv [w_k]_{M\times 1}$ ثابت بوده و به ازای $1 \le c \le C$ و $1 \le i \le N$ نامساوی $D_{ci}^2 > 0$ ، $m > 1$ ، $p > 1$ برقرار است. آنگاه U بهینه محلی $\tau(U)$ است اگر و فقط اگر u_{ci} (به ازای $c=1,2,\ldots,C$ و $i=1,2,\ldots,N$) طبق رابطه دوم قضیه محاسبه گردد.

اثبات: شرط لازم برای بهینگی محلی در لم ۲ ثابت شده است. برای اثبات کافی بودن آن، ماتریس هسیان $H(\tau(T))$ از $\tau(T)$ با استفاده از رابطه (۱) به صورت زیر محاسبه می‌شود.

$$h_{fg.ci}(U) = \frac{\partial}{\partial u_{.fg}}\left[\frac{\partial \tau(U)}{\partial u_{.ci}}\right] = \begin{cases} (m-1)u_{ci}^{m-2}\left(t_{ci}^p D_{ci}^2 + \mu_c(1-t_{ci})^p + \alpha t_{ci}^p\left(s_{ci}^M + s_{ci}^C\right)\right), & \text{If } f = c, g = i \\ \text{در غیر اینصورت } 0 \end{cases} \qquad (۲۹)$$

در رابطه (۲۹) ، $h_{fg.ci}(U)$ یک ماتریس قطری است که به ازای تمام $1 \le c \le C$ و $1 \le i \le N$ ، مقادیر t_{ci} و u_{ci} به کمک رابطه‌های اول و دوم قضیه محاسبه می‌شوند. همچنین داریم $0 < t_{ci} < 1$ ، $u_{ci} > 0$ ، $m > 1$ ، $D_{ci}^2 > 0$ ، $\mu_c > 0$

و $\alpha > 0$. با توجه به موارد ذکر شده نتیجه می‌شود که ماتریس هسیان فوق یک ماتریس مثبت معین است. بنابراین داریم:
$J_{\text{Proposed method}}(U^T,T^{T+1},W^T) \leq J_{\text{Proposed method}}(U^T,T^T,W^T)$ و بدین ترتیب شرط کافی رابطه دوم قضیه برای کمینگی محلی $\tau(U)$ اثبات می‌شود.

لم ۶ : فرض کنید $\tau(W) = J_{\text{Proposed method}}$ باشد که در آن $w \equiv [w_k]_{M\times 1}$ و شرط $\sum_{k=1}^{M} w_k = 1$ را برآورده می کند، $U \equiv [u_{ci}]_{C\times N}$ ثابت بوده و شرط $\sum_{c=1}^{C} u_{ci} = 1$ (به ازای i=1,2,…,N) را برآورده می‌کند. $T \equiv [t_{ci}]_{C\times N}$ ثابت بوده و به ازای $1 \leq c \leq C$ و $1 \leq i \leq N$ نامساوی های $D_{ci}^2 > 0$ ، $m > 1$ ، $p > 1$ برقرار است. آنگاه W بهینه محلی $\tau(W)$ است اگر و فقط اگر w_k (به ازای k=1,2,…,M) طبق رابطه سوم قضیه محاسبه گردد.

اثبات: شرط لازم برای بهینگی محلی در لم ۳ ثابت شده است. برای اثبات کافی بودن آن، ماتریس هسیان $H(\tau(T))$ از $\tau(T)$ با استفاده از رابطه (۲۲) به صورت زیر محاسبه می‌شود:

$$h_{f,k}(W) = \frac{\partial}{\partial w_f}\left[\frac{\partial \tau(W)}{\partial w_k}\right] = \begin{cases} 2y_k & \text{if } f = k \\ 0, & \text{otherwise} \end{cases} \quad (۳۰)$$

در رابطه (۳۰)، $h_{f,k}(W)$ یک ماتریس قطری است که به ازای تمام $1 \leq c \leq C$ و $1 \leq i \leq N$، مقادیر t_{ci} و u_{ci} به کمک رابطه‌های اول و دوم قضیه محاسبه می‌شوند. همچنین داریم $\mu_c > 0$ ، $D_{ci}^2 > 0$ ، $m > 1$ ، $u_{ci} > 0$ ، $0 < t_{ci} < 1$ و $\alpha > 0$. با توجه به موارد ذکر شده نتیجه می‌شود که ماتریس هسیان فوق یک ماتریس مثبت معین است. بنابراین داریم:
$J_{\text{Proposed method}}(U^T,T^{T+1},W^T) \leq J_{\text{Proposed method}}(U^T,T^T,W^T)$ و بدین ترتیب شرط کافی رابطه سوم قضیه برای کمینگی محلی $\tau(w)$ اثبات می‌شود.

اثبات همگرایی قضیه: شرط‌های لازم برای کمینگی محلی تابع هدف $J_{\text{Proposed method}}(w,T,U,V)$ در لم‌های ۱ ، ۲ و ۳ اثبات شده است. در لم ۱ کمینگی محلی این تابع هدف به فرض ثابت بودن درجه‌های عضویت فازی و وزن‌ها در هر گام اثبات شده است. در لم ۲ کمینگی محلی این تابع هدف با فرض ثابت بودن درجه های عضویت امکان‌پذیری و وزن‌ها در هر گام اثبات شده است. لم ۳ نیز کمینگی محلی این تابع هدف را با فرض ثابت بودن درجه های عضویت فازی و امکان‌پذیری اثبات نموده است. با استفاده از این سه لم، درستی
$J_{\text{Gol Function}}(U^T,^{T+1},W^T) \leq J_{\text{Gol Function}}(U^T,T^T,W^T)$
اثبات می‌شود. بنابراین همگرایی تابع هدف به اثبات می‌رسد.