

# ارائه روشی جدید برای خودکار سازی آستانه گیری در خوشه بندی بخردانه

محمد یوسف نژاد<sup>۱</sup>، دانشجوی دکتری؛ علی ریحانیان<sup>۲</sup>، دانشجوی دکتری؛ بهروز مینایی بیدگلی<sup>۳</sup>، دانشیار

۱- دانشکده علوم و تکنولوژی کامپیوتر - دانشگاه هوا و فضای نانجینگ - نانجینگ - چین - myousefnezhad@nuaa.edu.cn

۲- دانشکده مهندسی برق و کامپیوتر - دانشگاه تبریز - تبریز - ایران - ali.reihanian@gmail.com

۳- دانشکده مهندسی کامپیوتر - دانشگاه علم و صنعت ایران - تهران - ایران - b\_minai@iust.ac.ir

**چکیده:** در سال های اخیر، پژوهشگران، روش های مکاشفه ای مبتنی بر نظریه خرد جمعی را به منظور ارزیابی و انتخاب نتایج به دست آمده از خوشه بندی های پایه پیشنهاد کردند. در این روش ها، نتایج خوشه بندی با استفاده از معیارهای پراکندگی، استقلال و عدم تمرکز ارزیابی شده و با آستانه گیری از ارزیابی ها، نتایج به دست آمده انتخاب و ترکیب می شوند. هدف این مقاله، ارائه روشی جهت تخمین خودکار مقادیر بهینه آستانه، بر اساس ویژگی های اصلی داده در روش خوشه بندی بخردانه می باشد. علاوه بر آن، در این مقاله، به منظور اندازه گیری پراکندگی، معیاری جدید با عنوان همگونی بر اساس معیار APMM ارائه می شود. همچنین، جهت محاسبه استقلال به عنوان وزنی در ترکیب نتایج اولیه، روش انباشت مدارک وزن دار ارائه می شود. مقایسه نتایج تجربی به دست آمده بر روی چندین مجموعه داده استاندارد با سایر روش های خوشه بندی (ترکیبی)، نشان می دهد که روش پیشنهادی این مقاله از کارایی مناسبی برخوردار است.

**واژه های کلیدی:** خوشه بندی ترکیبی، خوشه بندی مبتنی بر انتخاب، خوشه بندی بخردانه، آستانه گیری خودکار در خوشه بندی، معیار همگونی.

## Proposing a New Framework for Automation of Thresholding in Wisdom of Crowds Cluster Ensemble Selection

M. Yousefnezhad<sup>1</sup>, PhD Candidate; A. Reihanian<sup>2</sup>, PhD Candidate; B. Minaei-Bidgoli<sup>3</sup>, Associate Professor

1- Faculty of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China, Email: myousefnezhad@nuaa.edu.cn

2- Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran, Email: ali.reihanian@gmail.com

3- School of Computer Engineering, Iran University of Science and Technology, Tehran, Iran, Email: b\_minai@iust.ac.ir

**Abstract:** Recently, researchers proposed heuristic frameworks which are based on the Wisdom of Crowds in order to evaluate and select the basic results. In these methods, basic results are evaluated by diversity, independency and decentralization metrics. Then, the evaluated results are selected by thresholding, and combined by a consensus function. This paper aims to propose a method for automatic evaluation of the optimized threshold values based on the basic features of the input data in WOCCE. Also, Uniformity, a metric which is based on APMM, is introduced for calculating the diversity of two basic clustering results. Furthermore, Weighted Evidence Accumulation Clustering (WEAC), a new method for considering independency as a weight in the process of combining the basic results, is introduced in this paper. The experimental results indicate that the proposed method has higher efficiency in comparison with the results of other cluster ensemble methods.

**Keywords:** Cluster ensemble, Cluster ensemble selection, Wised clustering, Automatic thresholding in clustering, Uniformity.

تاریخ ارسال مقاله: ۱۳۹۶/۱۲/۰۷

تاریخ اصلاح مقاله: ۱۳۹۷/۰۳/۲۴

تاریخ پذیرش مقاله: ۱۳۹۷/۰۶/۱۳

نام نویسنده مسئول: بهروز مینایی بیدگلی

نشانی نویسنده مسئول: ایران - تهران - میدان رسالت - خیابان هنگام - خیابان دانشگاه - دانشگاه علم و صنعت ایران - دانشکده مهندسی کامپیوتر.

## ۱- مقدمه

تحلیل خوشه‌بندی، نقش مهمی را در حوزه‌های علمی مختلف مانند داده‌کاوی، یادگیری ماشین، بازشناسی الگو، هوش تجاری و خوشه‌بندی اسناد ایفا می‌کند [۲-۱، ۷]. خوشه‌بندی، وظیفه کاوش الگوهای پنهان در داده‌های بدون برچسب را بر عهده دارد [۸]. به‌عنوان کاربردهایی از عمل خوشه‌بندی، می‌توان به استفاده از آن به‌منظور پیش‌بینی درخواست آتی کاربر در وب [۳]، ادغام تصاویر چند فوکوسه [۴]، اثربخشی بسط پرس‌وجو [۵] و حل مسائل بهینه‌سازی پویا [۶] اشاره کرد.

خوشه‌بندی ترکیبی، به‌عنوان یک ابزار قدرتمند برای تحلیل داده‌ها ظاهر شده است [۹]، که به خاطر پیچیدگی مسئله و ضعف روش‌های خوشه‌بندی پایه، می‌توان از روش‌های مبتنی بر آن بهره برد. در سال‌های اخیر، خوشه‌بندی ترکیبی در پژوهش‌های علمی مختلف به کار گرفته شده است که می‌توان به استفاده از آن جهت شناسایی خوشه‌های شکل گرفته به‌صورت مصنوعی<sup>۱</sup> [۱۰]، از بین بردن نویز<sup>۲</sup> در تصاویر پزشکی [۱۱] و دسته‌بندی نیمه نظارتی<sup>۳</sup> [۱۲] اشاره کرد.

در پژوهش‌های جدیدی که پیرامون خوشه‌بندی ترکیبی صورت پذیرفته، کیفیت نتایج اولیه خوشه‌بندی و پراکندگی<sup>۴</sup> در نتایج اولیه، توجه بسیاری را به خود جلب کرده است. اما پاسخ به بعضی سوالات در این زمینه، همچنان با ابهامات زیادی روبروست. خوشه‌بندی ترکیبی مبتنی بر انتخاب<sup>۵</sup>، روشی است که در آن از زیرمجموعه منتخب از نتایج اولیه، برای ترکیب و ساخت نتایج نهایی استفاده می‌شود [۸، ۱۳-۱۹].

خوشه‌بندی ترکیبی مبتنی بر انتخاب با سه چالش عمده همراه می‌باشد. یکی از این چالش‌ها، به‌کارگیری یک استراتژی تولید مناسب می‌باشد. با وجود این که یک استراتژی مناسب به‌منظور تولید نتایج الگوریتم‌های خوشه‌بندی پایه می‌تواند به‌طور چشم‌گیری بر روی عملکرد خوشه‌بندی ترکیبی مبتنی بر انتخاب تأثیر بگذارد، اما هیچ‌گونه استراتژی مشخصی در تحقیق‌های صورت گرفته قبلی در این زمینه ارائه نشده است [۲۰]. برخی از این تحقیق‌ها [۱۹، ۲۳-۲۱]، ابتدا هر یک از مولفه‌های خوشه‌بندی ترکیبی مبتنی بر انتخاب را به‌طور جداگانه اجرا می‌کنند (یعنی ابتدا تمامی نتایج پایه را تولید می‌کنند)، سپس به ارزیابی این نتایج می‌پردازند و ... درحالی که باقی تحقیق‌ها [۱۴، ۲۴]، هر یک از اجزای خوشه‌بندی ترکیبی مبتنی بر انتخاب را به‌طور تدریجی اجرا می‌کنند. به این مفهوم که نتیجه اولین الگوریتم خوشه‌بندی پایه را تولید کرده و آن را ارزیابی می‌کنند و ... از آنجایی که روش دوم، از نتایج ارزیابی در هر گام برای بهبود کیفیت نتایج تولید شده در گام‌های بعدی استفاده می‌کند، به آن، روش با مکانیزم بازخورد<sup>۶</sup> گفته می‌شود.

چالش دوم در خوشه‌بندی ترکیبی مبتنی بر انتخاب، ارزیابی می‌باشد. معیار اطلاعات متقابل نرمال شده (NMI)<sup>۷</sup> به‌عنوان متداول‌ترین معیار پراکندگی که در خوشه‌بندی ترکیبی مبتنی بر انتخاب استفاده

می‌شود، مشکل تقارن دارد [۱۴، ۲۱، ۲۲]. اگرچه برخی تحقیق‌ها، روش‌های جایگزینی مانند APMM [۲۱] و MAX [۲۲] را برای حل این مشکل ارائه دادند، اما روش پیشنهادی آن‌ها تنها برای ارزیابی پراکندگی بین یک خوشه و یک افزایش به‌کار گرفته می‌شود. از آنجایی که استفاده از روش‌های ذکر شده به‌منظور ارزیابی دو افزایش، موجب افزایش پیچیدگی زمانی می‌شود، لازم است که یک معیار جدید که بتواند به‌طور مستقیم پراکندگی بین دو افزایش را ارزیابی کند، ارائه شود. چالش سوم در خوشه‌بندی ترکیبی مبتنی بر انتخاب، بحث آستانه‌گیری می‌باشد. در عمل، یافتن مقادیر بهینه آستانه کار سختی بوده و از طرف دیگر، عملکرد خوشه‌بندی ترکیبی مبتنی بر انتخاب به‌طور قابل ملاحظه‌ای به مقادیر آستانه بستگی خواهد داشت [۲۰].

تعداد قابل توجهی از الگوریتم‌ها در علوم کامپیوتر، با الهام از طبیعت شکل گرفته‌اند. به‌عنوان مثال، شبکه‌های عصبی مصنوعی به‌عنوان روش‌های یادگیری، یا الگوریتم ژنتیک به‌عنوان روش بهینه‌سازی، از این گروه‌اند. این الگوریتم‌ها، با الهام از طبیعت، به‌عنوان رویکردی نو به‌منظور تولید نتایج دقیق<sup>۸</sup>، پایدار<sup>۹</sup> و مستحکم<sup>۱۰</sup>، در مسائل پیچیده به‌کار گرفته شدند. محاسبات جمعی<sup>۱۱</sup>، یک رویکرد جدید در علوم کامپیوتر می‌باشد که بر اساس برخی از نظریات موجود در علوم اجتماعی شکل گرفته است. نظریه خرد جمعی<sup>۱۲</sup> یکی از این نظریات می‌باشد که یک روش مستحکم را به‌منظور تولید نتایج دقیق در محاسبات جمعی توضیح می‌دهد.

نظریه خرد جمعی اولین بار توسط سورویکی معرفی شد. سورویکی در کتاب خود بیان می‌کند که یک جمع می‌تواند یک مسئله را بهتر از اکثر اعضای آن جمع به‌صورت انفرادی، حل کند. مطابق تعریف این کتاب، یک جمعیت به هر گروهی از افراد اطلاق می‌شود که شرایط چهارگانه پراکندگی، استقلال، عدم تمرکز و روش ترکیب آراء را دارند و می‌توانند به‌طور جمعی تصمیمی بگیرند و یا مسئله‌ای را حل کنند.

یکی از دلایلی که سورویکی در خصوص چرایی کارکرد نظریه خرد جمعی مطرح می‌کند این است که نظر هر فرد، دو عنصر را در درون خود دارد: اطلاعات صحیح و اطلاعات غلط. اطلاعات صحیح (از آن رو که صحیح‌اند) هم‌جهت‌اند و بر روی یکدیگر انباشه می‌شوند، اما خطاها در جهات مختلف و غیرهمسو عمل می‌کنند، لذا تمایل به حذف یکدیگر دارند، در نتیجه پس از جمع نظرات آنچه که می‌ماند اطلاعات صحیح است [۱۳، ۱۴، ۱۸، ۱۹، ۲۵].

با کمی جستجو، می‌توان به بسیاری از مفاهیم جدید در علوم مختلف دست یافت که از نظریه خرد جمعی به‌عنوان یک منبع اساسی استفاده کرده‌اند، مانند روش Delphi در مدیریت، سرمایه‌گذاری جمعی<sup>۱۳</sup> و ... اخیراً، این نظریه در علوم کامپیوتر به‌منظور بهینه‌سازی منابع در شبکه‌های حسگر بی‌سیم [۲۶] استفاده شده است. علاوه بر این، تحقیق‌های گوناگونی در حوزه یادگیری نظارت شده [۳۲-۲۷] و یادگیری بدون نظارت [۱۹، ۲۱] انجام شده‌اند که از نظریه خرد جمعی به‌منظور ارائه روش‌های جدید استفاده کرده‌اند. این تحقیق‌ها تصریح

۱. تنوع آراء - هر فرد باید به طور جداگانه اطلاعاتی از موضوع مورد نظر داشته باشد، حتی اگر اطلاعات مزبور غلط و مخدوش باشند.
۲. استقلال آراء - نظر افراد باید به طور مستقل و بدون تأثیر گرفتن از یک فرد یا گروه مشخص شکل گیرد.
۳. عدم تمرکز - افراد باید توانایی شخصی سازی و نتیجه گیری مبتنی بر دانش محلی خود را داشته باشند.
۴. مکانیزم ترکیب - باید مکانیزمی وجود داشته باشد که بتوان توسط آن، نظرات افراد را با یکدیگر ترکیب کرده و به یک نظر جمعی تبدیل نمود.

## ۲-۲- خوشه بندی ترکیبی

ایده اصلی خوشه بندی اطلاعات، جدا کردن نمونه‌ها از یکدیگر و قرار دادن آنها در گروه‌های شبیه به هم می‌باشد. به این معنی که نمونه‌های شبیه به هم باید در یک گروه قرار بگیرند و با نمونه‌های گروه‌های دیگر حداکثر تفاوت را دارا باشند [۱۹، ۳۳]. در واقع، خوشه بندی داده‌ها یک ابزار ضروری برای یافتن گروه‌ها در داده‌های بدون برچسب است [۱۷]. از آن جایی که اکثر روش‌های خوشه بندی پایه روی جنبه‌های خاصی از داده‌ها تأکید می‌کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد می‌باشند. به همین دلیل، نیازمند روش‌هایی هستیم که بتوانند با استفاده از ترکیب این روش‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه تری را تولید کنند. هدف اصلی خوشه بندی ترکیبی، جستجوی نتایج بهتر و مستحکم‌تر با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشه بندی اولیه است [۱۶، ۱۷]. خوشه بندی ترکیبی می‌تواند جواب‌های بهتری از نظر استحکام<sup>۱۶</sup>، نو بودن<sup>۱۷</sup>، پایداری<sup>۱۸</sup> و انعطاف پذیری<sup>۱۹</sup> نسبت به روش‌های پایه ارائه دهد [۱۵، ۱۷، ۳۴، ۳۵]. به طور خلاصه خوشه بندی ترکیبی شامل دو مرحله اصلی زیر می‌باشد [۱۴، ۱۵]:

۱. تولید نتایج متفاوت از خوشه بندی‌ها، به عنوان نتایج خوشه بندی اولیه، با اعمال روش‌های مختلف؛ که این مرحله را، مرحله ایجاد تنوع یا پراکندگی می‌نامند.
۲. ترکیب نتایج به دست آمده از خوشه بندی‌های متفاوت اولیه، به منظور تولید خوشه نهایی؛ که این کار توسط تابع توافقی<sup>۲۰</sup> (الگوریتم ترکیب کننده) انجام می‌شود.

## ۳- کارهای انجام شده

روش‌های خوشه بندی ترکیبی سعی می‌کنند تا با ترکیب افرازهای مختلف تولید شده از روش‌های خوشه بندی پایه، یک افراز مستحکم از داده‌ها تولید کنند [۱۷، ۳۸-۳۶]. در اکثر تحقیقات اخیر، همه افرازها با وزن برابر در ترکیب نهایی حاضر می‌شوند و همه خوشه‌های موجود در افرازها نیز با وزن برابر در ترکیب نهایی شرکت می‌کنند [۱۷، ۳۹] و یک معیار برای انتخاب از میان ترکیب‌های ممکن ارائه شده که مبتنی بر کیفیت کلی یک خوشه بندی است. برای این کار، آن‌ها میزان ثبات

می‌کنند که معمولاً، استفاده از نظریه خرد جمعی منجر به عملکرد بهتر و پایداری بیشتری خواهد شد.

همان‌طور که تصریح شد، در سال‌های اخیر، روش‌های مبتنی بر نظریه خرد جمعی، در حوزه یادگیری بدون نظارت پیشنهاد شده‌اند. این روش‌ها، به منظور ارزیابی و انتخاب نتایج به دست آمده از خوشه بندی‌های پایه، ارائه شده‌اند که ما آن‌ها را در این مقاله با عنوان "خوشه بندی بخردانه" مطرح می‌کنیم. در این روش‌ها، با استفاده از معیارهای پراکندگی، استقلال و عدم تمرکز، نتایج خوشه بندی پایه ارزیابی شده و سپس با آستانه گیری از این ارزیابی‌ها، نتایج به دست آمده انتخاب و ترکیب می‌شوند [۱۳، ۱۴، ۱۸، ۱۹]. اگرچه در این روش‌ها، مقادیر تعیین شده در فرآیند آستانه گیری، تأثیر قابل توجهی در کارایی و زمان اجرای الگوریتم دارند، ولی تا به حال هیچ روشی جهت تخمین این مقادیر ارائه نشده است.

هدف این مقاله، ارائه روشی جهت تخمین خودکار مقادیر آستانه به صورت بهینه و بر اساس ویژگی‌های اصلی داده ورودی می‌باشد. علاوه بر آن، در این مقاله، جهت اندازه گیری پراکندگی دو خوشه بندی پایه، معیاری جدید تحت عنوان همگونی<sup>۱۴</sup>، بر اساس معیار APMم ارائه شده است. همچنین، جهت حذف آستانه گیری، معیار استقلال به عنوان وزنی در ترکیب نتایج اولیه در نظر گرفته می‌شود. بدین منظور، روشی جدید تحت عنوان روش انباشت مدارک وزن دار<sup>۱۵</sup> ارائه می‌شود. نتایج تجربی به دست آمده بر روی چندین مجموعه داده استاندارد نشان می‌دهد که روش پیشنهادی این مقاله، به طور مؤثری نتایج نهایی را بهبود می‌بخشد. همچنین، مقایسه نتایج به دست آمده با سایر روش‌های خوشه بندی ترکیبی نشان از کارایی بالای روش پیشنهادی دارد.

در ادامه مقاله، ابتدا در بخش دوم به بررسی پیش زمینه‌های مورد نیاز پرداخته شده و در بخش سوم، کارهای انجام شده در این زمینه مرور می‌شوند. سپس در بخش چهارم، مدل پیشنهادی این مقاله ارائه می‌شود و در بخش پنجم به ارزیابی و بررسی فواید و مشکلات مدل پیشنهادی پرداخته می‌شود. در نهایت در بخش ششم، نتایج حاصل از این مقاله و خط و مشی کارهای آتی بیان می‌شوند.

## ۲- پیش زمینه

### ۲-۱- نظریه خرد جمعی

نظریه خرد جمعی که اولین بار سورویکی در کتابی با همین نام ارائه داده است، عنوان می‌کند که یک جمع می‌تواند مسئله را بهتر از اکثر اعضای گروه حل کند. مک‌کی اشاره می‌کند که همه جمعیت‌ها (گروه‌ها) بخردانه نیستند. یک مثال روشن از این قضیه، بازار سهام است که جمعیت به سمت حباب بازار هدایت می‌شود. بنابراین، ابتدا باید فهمید که تحت چه شرایطی خرد جمعی می‌تواند اثرگذار باشد [۱۴، ۱۹، ۲۵]. سورویکی، چهار شرط اساسی زیر را برای تمایز جمعیت بخردانه از یک جمعیت غیرعقل پیشنهاد می‌دهد [۱۴، ۲۵]:

این روش، شامل رویکردی نوین جهت تشکیل ماتریس همبستگی بدون نیاز به تمامی نتایج خوشه‌های خوشه‌بندی اولیه می‌باشد [۲۱، ۲۲].

یوسف‌نژاد و همکاران، اولین بار مفهوم استقلال دو الگوریتم خوشه‌بندی را معرفی، و تاثیرات آن بر کارایی نتیجه نهایی خوشه‌بندی را بررسی کردند. در این روش، دو الگوریتم غیر هم‌نام، کاملاً مستقل و درجه استقلال الگوریتم‌های هم‌نام، بر اساس پارامترهای اساسی آن الگوریتم محاسبه می‌شوند. برای مثال، آن‌ها مقادیر تصادفی اولیه مراکز خوشه در الگوریتم k-means را به‌عنوان پارامترهای اساسی مؤثر برای این الگوریتم فرض کردند [۱۸، ۱۹].

علیزاده و همکاران، از مفهوم ذکر شده (استقلال) برای تعریف معیار استقلال در خوشه‌بندی بخردانه استفاده کردند. در این روش، با تغییر روش کار الگوریتم Linkage، معیاری تحت عنوان Likeness برای محاسبه درجه استقلال الگوریتم با استفاده از ماتریس پارامترهای اساسی آن الگوریتم، معرفی شد [۱۴].

در پژوهشی دیگر، علیزاده با استفاده از معیار پایداری، روشی دیگر برای محاسبه درجه استقلال دو الگوریتم هم‌نام پیشنهاد داد [۱۳].

خوشه‌بندی بخردانه، اولین بار توسط علیزاده و همکاران با استفاده از بازتعریف و نگاشت مفهوم خردمندی - از تعاریف کتاب سورویکی - به مسائل خوشه‌بندی، معرفی شد [۱۴]. در روش ارائه شده توسط علیزاده و همکاران، نتایج الگوریتم‌های پایه در فرآیند اجرای الگوریتم، تولید و سپس بر اساس معیارهای استقلال و پراکندگی، ارزیابی می‌شوند و در صورت تأیید، به مجمع خردمند وارد می‌شوند.

در فرآیند ارزیابی نتایج به‌دست آمده در روش‌های ارائه شده در [۱۳، ۱۴، ۱۸، ۱۹]، همیشه مقادیر ارزیابی با یک سطح آستانه که توسط کاربر به‌صورت دستی وارد می‌شود، مقایسه خواهند شد. با اینکه این مقادیر آستانه تأثیر بسزایی بر کارایی و زمان اجرای الگوریتم دارند، هیچ‌کدام از روش‌های پیشین، راهکاری مناسب برای محاسبه و تخمین این مقادیر آستانه ارائه نکردند. هدف این مقاله، بررسی عوامل مؤثر بر روی این مقادیر و ارائه روشی به‌منظور تخمین مقدار بهینه آن‌ها است.

#### ۴- مدل پیشنهادی

جهت ارائه روش پیشنهادی، ما ابتدا بر اساس تعاریف مطرح شده در بخش قبل (کارهای انجام شده)، یک تعریف پایه در مورد چهار عنصر سازنده خرد جمعی یعنی پراکندگی، استقلال، عدم تمرکز و مکانیزم ترکیب مناسب، ارائه می‌دهیم و سپس، راهکار پیشنهادی این مقاله را بر اساس این تعریف پایه بیان خواهیم کرد.

نو بودن و پایداری در نتایج اولیه، از مهم‌ترین خواص هستند که این مقاله به دنبال رسیدن به آن است. در اینجا، نو بودن یعنی رسیدن به افراز جدیدی در خوشه‌بندی‌های اولیه که تازه‌حال، در سایر نتایج به این حالت نرسیده‌ایم که این امر، کمک بسزایی در کشف الگوهای جدید (دانش ضمنی) از داده می‌کند. پایداری نیز تضمین می‌کند که با تکرار مکرر یک روش روی یک داده، به نتایج مشابهی خواهیم رسید.

بین افراز ترکیبی و افرازهای پایه را در نظر گرفتند و با استفاده از یک قاعده ترکیبی ثابت، یک معیار شباهت دو به دو را روی فضای ویژگی‌های چند بعدی به‌کار بردند.

عظیمی و همکاران [۴۰] از مفهوم پراکندگی برای هوشمند نمودن خوشه‌بندی ترکیبی استفاده کردند. این روش، به‌صورت پویا اقدام به انتخاب زیرمجموعه بهینه‌ای از نتایج اولیه در ترکیب نهایی می‌کند. نتایج تجربی به‌دست آمده نیز بیانگر این مورد می‌باشند، که ترکیب خوشه‌بندی‌های اولیه با بیشترین، کمترین و میزان متوسطی از تطبیق با خوشه‌بندی ترکیبی اولیه، نتیجه بهتری را به ترتیب، در مجموعه داده‌های راحت، سخت و متوسط می‌دهد. روش فوق در هر مجموعه داده سعی می‌کند تا نتایج خوشه‌بندی اولیه‌ای که موجب منحرف شدن نتایج نهایی می‌شود را از ترکیب نهایی خارج کند و به این ترتیب، خوشه‌بندی‌های ترکیبی اولیه‌ای را که دارای دقت<sup>۲۱</sup> نسبتاً مناسبی هستند، وارد ترکیب نهایی کند [۴۲-۴۰].

چندین روش اعتبارسنجی خوشه، مبتنی بر ایده استفاده از پایداری پیشنهاد شده است [۴۱-۴۴]. بن‌هور و همکاران نیز روشی برای محاسبه پایداری ارائه کردند که بر مبنای شباهت بین نمونه‌ها در خوشه‌بندی‌های مختلف عمل می‌کند. در این روش، ماتریس همبستگی با استفاده از روش بازنمونه‌برداری به‌دست می‌آید [۴۳].

فرد و جین یک روش خوشه‌بندی ترکیبی ارائه کردند که در آن با استفاده از معیار پایداری خوشه، شباهت دو به دو آموزش داده می‌شود. در این روش، به‌جای استفاده از معیارهای ارزیابی مبتنی بر افراز نهایی، افرازهای حاصل از الگوریتم‌های پایه، در نواحی مختلف از فضای ویژگی چند بعدی مورد ارزیابی قرار می‌گیرند [۳۹].

فرن و لین، روشی برای خوشه‌بندی ترکیبی پیشنهاد کردند که از زیرمجموعه موثرتری از افرازهای اولیه در ترکیب نهایی استفاده می‌کند. در این روش، اگرچه تعداد اعضای شرکت‌کننده در ترکیب نهایی کمتر از یک خوشه‌بندی ترکیبی کامل<sup>۲۲</sup> است، به دلیل انتخاب افرازهای با کارایی بالاتر، نتایج نهایی بهبود می‌یابند. پارامترهایی که در این روش مورد توجه قرار گرفتند، عبارتند از: کیفیت و پراکندگی [۲۳]. در این روش، از معیار مجموع اطلاعات متقابل نرمال شده (SNMI<sup>۲۳</sup>) برای یک افراز در مقایسه با افرازهای دیگر ترکیب برای اندازه‌گیری کیفیت یک افراز استفاده شده است. همچنین در این روش، معیار اطلاعات متقابل نرمال شده (NMI) (بین تمام افرازهای موجود در ترکیب) برای اندازه‌گیری پراکندگی لازم برای ترکیب به کار رفته است [۲۳]. فرن و لین نشان دادند که روش پیشنهادی آن‌ها نسبت به خوشه‌بندی ترکیبی کامل و یا روش انتخاب تصادفی، از کارایی بهتری برخوردار است [۲۳].

علیزاده و همکاران [۲۲]، روشی جهت انتخاب خوشه بر اساس معیار پایداری ارائه دادند. در این روش، به معرفی معیار APMM و روش ماکزیمم، جهت رفع مشکل تقارن در معیار NMI پرداخته شده است.

چالش، برخی تحقیق‌ها به منظور ارائه روش‌هایی برای تعیین مقدار اولیه مناسب در این الگوریتم، صورت گرفتند [۱۴، ۴۶، ۴۷]. بر اساس این تحقیق‌ها، توجه به مقادیر اولیه و انتخاب آن‌ها به صورت هدفمند و مناسب (نه به صورت تصادفی)، می‌تواند عملکرد الگوریتم را بهبود بخشیده و باعث افزایش دقت آن شود.

همچنین، ذکر این نکته ضروری است که وقتی در این مقاله بحث استقلال به میان می‌آید، به هیچ عنوان استقلال مطلق (یعنی کاملاً مستقل بودن دو الگوریتم)، مد نظر نیست، چرا که دو الگوریتمی که به نوعی از هم مستقلند (با یک درجه استقلال فرضی)، حداقل در متریک مورد استفاده با هم اشتراکاتی دارند. به عبارت دیگر، اگر به تحلیل میزان مستقل بودن دو الگوریتم بپردازیم، خواهیم دید که این‌ها در بعضی خصیصه‌ها (مانند متریک و یا نمایش داده‌ها) با یکدیگر مشترک و در بعضی دیگر (مانند تابع هدفی که باید آن را بیشینه یا کمینه کنند)، از یکدیگر مستقل هستند. در نتیجه، نمی‌توان گفت که فرضاً دو الگوریتم، کاملاً مستقل و یا کاملاً وابسته هستند، بلکه می‌توان بیان کرد که دو الگوریتم با چه درجه استقلال از هم مستقلند.

خوشه‌بندی بخردانه، ماتریسی از مقادیر اولیه را به عنوان عامل محرک الگوریتم، بر اساس روش کار هر الگوریتم، در نظر می‌گیرد (به عنوان مثال، مقدار تصادفی اولیه مراکز خوشه‌ها در الگوریتم‌هایی مثل k-means و FCM، یا پارامترهای داخلی الگوریتم‌ها همانند ماتریس فاصله در روش‌های Spectral و هر نوع مقادیر اولیه‌ای که می‌توانند روش کار الگوریتم‌ها را تغییر دهند، که به این مقادیر اولیه، پارامترهای اساسی الگوریتم گفته می‌شود). بدیهی است، چون روش حل مسئله در هر الگوریتم ثابت است، اگر مقادیر ثابت بمانند، جواب‌های نهایی الگوریتم پایه یکی خواهد بود یا به عبارتی، نتایج هر الگوریتم پایه به مقادیر این ماتریس وابسته می‌باشد. از این رو بر اساس تعریف استقلال، درجه استقلال دو الگوریتم به صورت شبه کد الگوریتم ۱ محاسبه می‌شود [۱۹]:

```
Function BIndependency (C1, C2, P1, P2)
If type of cluster C1 and C2 is equal then
Distance-Matrix is distance between P1 and P2
Do until Distance-Matrix is not null
Find minimum cell of Distance-Matrix
Store cell in Temp-Array
Remove Row and Column of founded cell
Create new Distance-Matrix
End loop
Return average of Temp-Array
Else
Result = 1
End If
End Function
```

الگوریتم ۱: محاسبه درجه استقلال دو الگوریتم خوشه‌بندی [۱۴، ۱۸]

[۱۹]

در الگوریتم ۱، C1 و C2 دو الگوریتم خوشه‌بندی می‌باشند که قرار است درجه استقلال آنها با هم مقایسه شود. همچنین P1 و P2 به ترتیب، ماتریس‌های پارامترهای اساسی این دو الگوریتم می‌باشند. در صورت

در صورتی که فقط رسیدن به پایداری نتایج نهایی خوشه‌بندی ترکیبی مهم باشد، ممکن است این دو خصوصیت (نو بودن و پایداری) در خلاف راستای هم‌دیگر عمل کنند، به این معنی که: "هر چقدر نو بودن جواب‌ها در تکرار مکرر یک خوشه‌بندی بیشتر باشد، جواب‌های غیر پایداری مشاهده می‌شود و هر چقدر پایداری بیشتری مد نظر باشد، خیلی از جواب‌های نو از دست رفته و نهایتاً درصد پیش‌بینی الگوی درست کمتری مشاهده خواهد شد [۸، ۱۹-۱۳، ۳۸-۳۶]."

#### ۴-۱- محاسبه درجه استقلال دو الگوریتم

در خوشه‌بندی بخردانه، استقلال دو الگوریتم خوشه‌بندی به این صورت تعریف می‌شود که [۱۴، ۱۹]: "نتیجه خوشه‌بندی پایه نباید متأثر از نتایج دیگر خوشه‌بندی‌های پایه باشد. این تأثیر می‌تواند در سطح نوع الگوریتم (گروه) یا پارامترهای مؤثر در نتایج یک الگوریتم خاص (افراد) باشد." در بیشتر روش‌های خوشه‌بندی ترکیبی، جهت ایجاد پراکندگی و رسیدن به نتایج نوتر و انعطاف‌پذیرتر، از تکرار مکرر یک الگوریتم پایه خوشه‌بندی (برای مثال k-means) روی داده، بهره گرفته می‌شود. در این الگوریتم‌ها، عموماً جهت ایجاد نتایج متفاوت، در بخشی از روش حل مسئله از مقادیر قابل برنامه‌ریزی یا تصادفی استفاده می‌شود. برای مثال در k-means، مقادیر اولیه مراکز خوشه‌ها یا مقدار k یا تعداد دفعات تکرار الگوریتم، جزء این پارامترها می‌باشد. لازم به ذکر است که برخی از الگوریتم‌ها همانند Linkage، که با تکرار مکرر بر روی یک داده، همیشه یک جواب معین را تکرار می‌کنند (معمولاً از مولد اعداد تصادفی استفاده نمی‌کنند)، شامل این قانون نمی‌شوند و معمولاً در ساخت نتایج اولیه خوشه‌بندی ترکیبی، از هر یک از انواع آن فقط یک‌بار استفاده می‌شود [۱۷، ۲۱، ۳۷، ۴۵].

طبق تعریف ذکر شده استقلال در خوشه‌بندی بخردانه، استفاده از روش بالا جهت ایجاد پراکندگی، باعث انتشار خطا در نتیجه نهایی می‌شود [۱۳، ۱۴، ۱۸]. به عنوان مثالی از استقلال، اگر ما دو نتیجه مشابه از الگوریتم‌های k-means و FCM داشته باشیم، آنگاه چون روش‌های حل مسئله (توابع هدف) در این دو الگوریتم با هم متفاوت بوده و نسبت به هم مستقل هستند، نتایج این دو خوشه‌بندی با اینکه مشابه ولی مستقل و قابل اتکا می‌باشند. همچنین، به عنوان مثالی از انتشار خطا، اگر دو خوشه‌بندی پایه که هر دو با k-means انجام شده‌اند، دارای نتایج مشابه باشند و پارامترهای تاثیرگذار در الگوریتم k-means (برای مثال نقاط تصادفی اولیه مراکز خوشه‌ها) با هم برابر بوده و یا اختلاف ناچیزی با یکدیگر داشته باشند، آنگاه این دو خوشه‌بندی به علت استفاده از روش مشابه، به همدیگر وابسته می‌باشند. یعنی اگر در هر دو، یک پارامتر تغییر کند، نتایج باز هم مشابه خواهند بود.

در این جا، باید به ذکر نکته‌ای در ارتباط با تأثیر انتخاب مقادیر اولیه بر نتایج نهایی در برخی روش‌های خوشه‌بندی بپردازیم. به عنوان مثال، اگر الگوریتم k-means ده‌ها بار بر روی یک داده ثابت و با یک مقدار دهی اولیه ثابت اعمال شود، نتیجه یکسان خواهد بود. برای حل این



ما در این مقاله، از معیار APMM برای محاسبه مقدار پراکندگی استفاده می‌کنیم، چراکه این معیار، هم سریع‌تر از NMI می‌باشد و هم مشکل تقارن ندارد [۱۳، ۱۴، ۱۸، ۱۹، ۲۱]. در این روش، برای محاسبه تراکم خوشه  $C_i$  از رابطه (۲) استفاده می‌کنیم:

$$AAPMM(C_i) = \frac{1}{M} \sum_{j=1}^M APMM(C_i, P_j^{b*}) \quad (2)$$

در رابطه (۲) پارامتر  $P_j^{b*}$  نشان دهنده ژامین افراز از مجموعه مرجع است. همچنین تابع APMM در رابطه (۲) را از رابطه (۳) محاسبه می‌کنیم:

$$APMM(C, P) = \frac{-2n_c \log\left(\frac{n}{n_c}\right)}{n_c \log\left(\frac{n_c}{n}\right) + \sum_{i=1}^{k_p} n_i^p \log\left(\frac{n_i^p}{n}\right)} \quad (3)$$

در رابطه (۳)، پارامتر  $n$ ، تعداد کل نمونه‌های خوشه  $C$  می‌باشد. همچنین،  $n_c$  تعداد نمونه‌های مشترک بین خوشه  $C$  و افراز  $P$  می‌باشد. علاوه بر این،  $K_p$  تعداد خوشه‌های موجود در افراز  $P$  بوده و  $n_i^p$  تعداد نمونه‌های خوشه  $i$ -ام در افراز  $P$  می‌باشد. از آنجایی که AAPMM فقط تراکم یک خوشه را محاسبه می‌کند، برای محاسبه یک خوشه‌بندی، در این مقاله مطابق رابطه (۴)، از معیاری جدیدی تحت عنوان همگونی استفاده می‌شود:

$$Uniformity(P) = \max_{i=1}^M AAPMM(C_i) \quad (4)$$

در رابطه (۴)، معیار همگونی، تراکم یک افراز که در اینجا نتیجه یک خوشه‌بندی پایه است را محاسبه می‌کند که در آن،  $M$  تعداد کل خوشه‌ها می‌باشد. روش‌های پیشین، از میانگین AAPMM خوشه‌ها برای ارزیابی تراکم استفاده کرده‌اند که این امر، باعث افزایش احتمال ورود خوشه‌ها با پراکندگی کمتر و در نتیجه، کاهش دقت الگوریتم می‌شود. از آنجایی که خروجی معیار همگونی بین ۰ و ۱ می‌باشد، طبق رابطه (۵)، پراکندگی نهایی هر افراز برابر با تفاضل ۱ از مقدار به دست آمده این معیار می‌باشد:

$$DIV(P) = 1 - Uniformity(P) \quad (5)$$

این مقاله، از ترکیب روش‌های عظیمی و همکاران [۴۰-۴۲] و علیزاده [۱۳]، روشی را جهت تعیین مقدار بهینه برای آستانه‌گیری پراکندگی به صورت خودکار ارائه می‌دهد.

همان‌طور که پیش‌تر نیز اشاره شد، عظیمی و همکاران، داده را به سه نوع راحت، متوسط و سخت دسته‌بندی کردند. همچنین آنها اثبات کردند که به ترتیب برای هر یک از انواع راحت، متوسط و سخت داده بهتر است نتایج با ۳۳ درصد پراکندگی پایین، میانی و بالا، پس از ارزیابی انتخاب شوند [۴۰-۴۲]. از این‌رو، در این مقاله به جای استفاده از یک مقدار ثابت به‌عنوان آستانه پراکندگی، پس از ارزیابی پراکندگی نتایج اولیه، از بازه  $0.33 \leq DIV \leq 0.66$  برای انتخاب نتایج اولیه در داده‌های راحت، از بازه  $0.33 \leq DIV \leq 0.66$  برای انتخاب نتایج اولیه در داده‌های

یکی نبودن نوع دو الگوریتم، استقلال آنها برابر با ۱ محاسبه می‌شود که به معنی کاملاً مستقل می‌باشد، در غیر این صورت، ماتریس  $n \times n$  فاصله Max-Distance بر اساس  $P_1$  و  $P_2$  تشکیل می‌شود. در اینجا،  $n$  برابر با حداکثر اندازه ماتریس‌های  $P_1$  و  $P_2$  است. لازم به ذکر است که در این حالت، هر چه فاصله (در این مقاله ما از فاصله اقلیدسی استفاده کردیم، ولی می‌توان از هر معیار فاصله دیگری استفاده کرد) بیشتر باشد، درجه استقلال بهتری به دست خواهد آمد. از این‌رو در حلقه الگوریتم ۱ می‌بایست هر بار، مقدار حداقل، پیدا شده و در Temp-Array نگهداری شود و سطر و ستونی که در آن، این مقدار وجود دارد، حذف شده و برای ماتریس جدید به وجود آمده، مجدداً همین کار تکرار شود. نهایتاً مقدار درجه استقلال، میانگین مقادیر حداقل‌ها در ماتریس فاصله خواهد بود. خروجی الگوریتم ۱ همواره یک مقدار بین ۰ و ۱ خواهد بود که در آن، ۱ به معنی کاملاً مستقل و ۰ به معنی کاملاً وابسته می‌باشد.

در روش‌های پیشین، از میانگین مقادیر BIndependency برای محاسبه درجه استقلال هر الگوریتم استفاده می‌شد. چون این مقادیر، شامل تعداد زیادی ۱ (استقلال الگوریتم‌های غیر هم‌نام) می‌باشند، درجه استقلال همیشه برابر با یک عدد نزدیک به ۱ محاسبه می‌شد. در این مقاله، مطابق با رابطه (۱)، مقدار بیشینه مقادیر محاسبه شده BIndependency، فقط در الگوریتم‌های هم‌نام به‌عنوان درجه استقلال الگوریتم در نظر گرفته می‌شود تا تأثیر آن بر روی کیفیت نتایج نهایی حفظ شود:

$$IND(C) = \max_{i=1}^M BIndependence(C, C_i) \quad (1)$$

در روش‌های پیشین، پس از محاسبه درجه استقلال، آن را با مقدار آستانه قابل برنامه‌ریزی  $iT$  که توسط کاربر در ابتدای اجرای الگوریتم وارد می‌شد، مقایسه و در صورت تأیید، نتیجه آن را وارد مجمع بخردانه می‌کردیم [۱۴، ۱۸، ۱۹]. در این مقاله، از درجه استقلال الگوریتم به‌عنوان وزنی در ترکیب نتایج انتخاب‌شده، استفاده می‌کنیم. در این‌روش، هم یک مرحله از اجرای الگوریتم حذف می‌شود که این امر، موجب کاهش زمان اجرای الگوریتم می‌شود و هم دیگر نیازی به آستانه‌گیری از معیار استقلال نخواهیم داشت. در بخش مکانیزم ترکیب نتایج اولیه، روشی جدید تحت عنوان انباشت مدارک وزن‌دار را جهت استفاده از درجه استقلال الگوریتم به‌عنوان وزن‌های نتایج اولیه انتخاب شده معرفی خواهیم کرد.

#### ۴-۲- محاسبه پراکندگی نتایج اولیه

در خوشه‌بندی بخردانه، پراکندگی نتایج اولیه را به این صورت تعریف می‌کنیم که: "هر الگوریتم خوشه‌بندی پایه، باید به‌طور جداگانه و بدون واسطه، به داده‌های مسئله دسترسی داشته و آن‌ها را تحلیل و خوشه‌بندی کند، حتی اگر نتایج آن غلط باشد." در اینجا، نتایج غلط، موجب کشف عدم تنوع و جلوگیری از تکرار یک جواب خاص خواهد شد.

۳- مقدار آستانه  $cT$ ، ضریب عدم تمرکز نامیده می شود که از آن به عنوان ضریب برای تعداد خوشه های الگوریتم پایه استفاده می شود. در این روش، تعداد خوشه ها در خوشه بندی پایه از  $k$  تا  $k \times cT$  متغیر می باشد.

در این مقاله، دو شرط اول عیناً استفاده شده است، ولی به جای تعیین مقدار آستانه  $cT$ ، از روشی که اخیراً توسط علیزاده ارائه شد، برای تولید پراکندگی لازم استفاده شده است. در این روش، تعداد خوشه های هر الگوریتم از بازه  $k-2, \dots, k+2$  به شرطی که مقدار انتخاب شده در این بازه کمتر از عدد ۲ نباشد، تعیین می شود که در آن،  $k$ ، تعداد کلاس داده در نتیجه نهایی می باشد [۱۳].

#### ۴-۴- مکانیزم ترکیب نتایج اولیه

در خوشه بندی بخردانه، خوشه های انتخاب شده توسط ماتریس همبستگی، با هم ترکیب شده و نتیجه نهایی را تولید می کنند. در روش های پیشین، از روش انباشت مدارک (EAC) استفاده می شد که در آن، نتایج  $m$  خوشه بندی روی داده های نمونه برداری شده، در ماتریس همبستگی  $n \times n$  ذخیره می شوند. هر داده ورودی از این ماتریس در روش انباشت مدارک، به صورت رابطه (۸) محاسبه می شود [۳۷]:

$$C(i, j) = \frac{n_{i,j}}{m_{i,j}} \quad (8)$$

در رابطه (۸)،  $n_{i,j}$  تعداد دفعاتی است که جفت نمونه های  $i$  و  $j$  با هم در یک خوشه گروه بندی شده اند و  $m_{i,j}$  تعداد نمونه برداری هایی است که هر دوی این جفت نمونه ها به طور هم زمان در آن ظاهر شده اند [۳۷]. با توجه به نحوه محاسبه پارامتر  $n_{i,j}$ ، می توان گفت شمارش تعداد نمونه ها، یعنی هر نتیجه، با وزن مساوی و برابر با ۱، در جواب نهایی شرکت می کند. در این مقاله، دیدگاه جدیدی در مورد رابطه (۸) مطرح می شود، که در آن، به جای فرض کردن وزن یکسان و برابر با ۱ برای هر نتیجه تولید شده از الگوریتم های شرکت کننده در مجمع بخردانه، از درجه استقلال آن الگوریتم ها که طبق رابطه (۱) محاسبه می شود و مقداری بین ۰ و ۱ خواهد بود، استفاده می شود. از این رو، رابطه (۸) به صورت زیر اصلاح می شود:

$$C(i, j) = \frac{\sum IND_{Alg_p}}{m_{i,j}} \quad (9)$$

در رابطه (۹)،  $IND_{Alg_p}$  درجه استقلال الگوریتمی است که جفت نمونه های  $i$  و  $j$  با هم در یک خوشه، گروه بندی کرده است. این روش که ما به عنوان روش انباشت مدارک وزن دار می شناسیم، با تأثیر مستقیم استقلال بر روی ماتریس همبستگی، باعث تغییر شکل دندروگرام ترکیب نتایج شده و کیفیت نتیجه نهایی را بهبود می بخشد.

متوسط و از بازه  $0.66 \leq DIV$  برای انتخاب نتایج اولیه در داده های سخت استفاده می شود.

برای ارزیابی داده ها به منظور دسته بندی آنها در سه مجموعه مذکور، این مقاله از روش علیزاده [۱۳] که به نوعی توسعه یافته تر از روش عظیمی و همکاران می باشد، استفاده کرده است. در این روش، مطابق رابطه (۶) از معیار سادگی به عنوان روشی جهت تعیین نوع داده استفاده شده است [۱۳، ۲۲]:

$$Simplicity(D) = \frac{1}{B} \sum_{i=1}^B Stability(P_i) \quad (6)$$

در رابطه (۶)،  $B$  تعداد کلیه افزایش های تولید شده در فرآیند خوشه بندی می باشد و پایداری افزایش  $i$ -ام مطابق رابطه (۷) محاسبه می شود [۱۳، ۲۲]:

$$Stability(P) = \frac{1}{M} \sum_{i=1}^M APMM(C_i, P) \quad (7)$$

در رابطه (۷)،  $M$  تعداد خوشه هر افزایش می باشد و  $APMM$  از رابطه (۳) محاسبه می شود. مطابق تعریف علیزاده، اگر نتایج سادگی داده ای کمتر از ۰.۵ باشد، آن داده را سخت، اگر نتایج سادگی آن داده بین ۰.۵ و ۰.۵۵ باشد، آن داده را متوسط و در صورت به دست آمدن نتایج سادگی دیگری برای آن داده، آن را راحت در نظر می گیریم [۱۳، ۲۲].

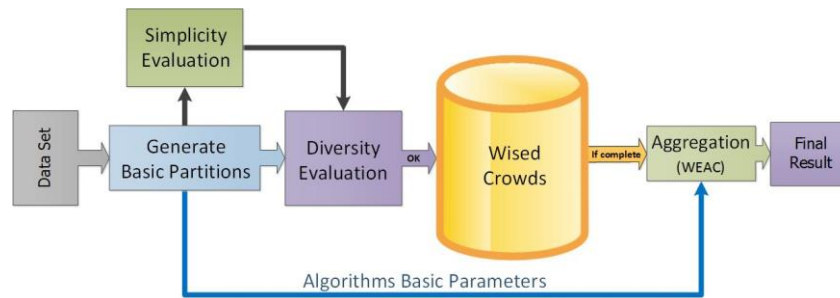
#### ۴-۳- عدم تمرکز در تولید نتایج اولیه

در معیار کیفی عدم تمرکز، مطابق تعاریف خوشه بندی بخردانه، دو واژه اصلی مطرح می شود: شخصی سازی و نتیجه گیری بر اساس دانش محلی.

شخصی سازی در خوشه بندی بخردانه به این صورت تعریف می شود [۱۳، ۱۴، ۱۸، ۱۹]: "شخصی سازی در خوشه بندی ترکیبی یعنی باید الگوریتم پایه برای رسیدن به نتایج مؤثرتر، به هر طریقی که مناسب تر است داده های ورودی را تحلیل کند که این کار، می تواند به صورت تغییر ابعاد داده یا ادغام داده ها و ... صورت پذیرد."

مطابق تعاریف خوشه بندی بخردانه، تحلیل بر اساس دانش محلی به این صورت تعریف می شود [۱۳، ۱۴، ۱۸، ۱۹]: "در خوشه بندی بخردانه، هر الگوریتم پایه جهت رسیدن به نتایج مؤثرتر، باید به طور آزادانه بهترین انتخاب را در پارامترهای اساسی خود داشته باشد." مطابق تعاریف بالا، علیزاده و یوسف نژاد، سه شرط زیر را برای تحقق شرایط عدم تمرکز در خوشه بندی بخردانه تعریف کردند [۱۳، ۱۸]:

- ۱- تعداد الگوریتم های پایه شرکت کننده در تولید نتایج اولیه باید بیشتر از یک الگوریتم باشد.
- ۲- روش ورود نتایج یک الگوریتم پایه به مجمع بخردانه باید طوری باشد که نتایج نهایی تحت تأثیر خطاهای آن قرار نگیرد، یا به عبارتی، نباید روش تصمیم گیری در مورد جواب نهایی، متمرکز باشد.



شکل ۱: چهارچوب خوشه‌بندی بخردانه با تخمین مقادیر آستانه به صورت خودکار

#### ۵-۴- جمع‌بندی

شکل ۱ فرآیند اجرای الگوریتم خوشه‌بندی بخردانه را در روش پیشنهادی این مقاله به تصویر می‌کشد. همان‌طور که در شکل ۱ مشاهده می‌شود، ابتدا نتایج خوشه‌بندی‌های پایه، تولید شده، آنگاه بر اساس رابطه (۶)، سادگی داده چک شده، سپس بر اساس رابطه (۵)، پراکندگی نتایج اولیه، ارزیابی شده و نتایج مناسب وارد مجمع بخردانه می‌شوند. در بخش ترکیب، با استفاده از رابطه (۹)، عناصر ماتریس همبستگی بر اساس درجه استقلال الگوریتم‌هایشان تشکیل شده و با استفاده از الگوریتم سلسله مراتبی اتصال میانگین، نتایج خوشه‌بندی نهایی تشکیل می‌شود [۱۳، ۱۴، ۱۶، ۱۸، ۱۹]. الگوریتم ۲، شبه‌کد فرآیند شکل ۱ را نمایش می‌دهد:

```
Function Wised-Clustering (Data, k, n)
  Initiate nCrowd = 0
  Do until nCrowd is less than n
    Generate a new basic clustering result
    Calculate result Simplicity
    Evaluate Diversity of result
    If result was acceptable then
      Add result to wised crowds
      nCrowd++
    End If
  End Do
  Create Co-Association matrix based on WEAC
  Create final result based on Co-Association matrix
End Function
```

الگوریتم ۲: شبه‌کد خوشه‌بندی بخردانه با آستانه‌گیری خودکار

در الگوریتم ۲، پارامتر ورودی  $n$  برابر با تعداد اعضای مجمع بخردانه [۱۳، ۱۴، ۱۸، ۱۹]،  $k$  برابر با تعداد خوشه‌های نتیجه نهایی مسئله،  $Data$  برابر با داده ورودی و  $nCrowd$  برابر با تعداد اعضای فعلی انتخاب شده در مجمع بخردانه می‌باشد. همان‌طور که پیش‌تر نیز به آن اشاره شد، در این فرآیند، به‌منظور تولید نتیجه نهایی از روی ماتریس همبستگی، از الگوریتم سلسله مراتبی اتصال میانگین استفاده می‌کنیم، چراکه تأثیر بسزایی بر روی کارایی الگوریتم و دقت نتیجه به‌دست آمده دارد [۸، ۱۶-۱۳، ۱۸، ۱۹، ۲۲، ۴۱، ۴۳، ۴۴]. مجموعه داده استاندارد می‌باشد که در آزمایش‌های تجربی صورت پذیرفته در این پژوهش، مورد استفاده قرار گرفته‌اند.

#### ۵- ارزیابی

##### ۵-۱- مجموعه داده‌ها

در ادامه، جدول ۱ مشاهده می‌شود که این جدول، شامل مشخصات ۱۴

جدول ۱: مجموعه داده‌ها

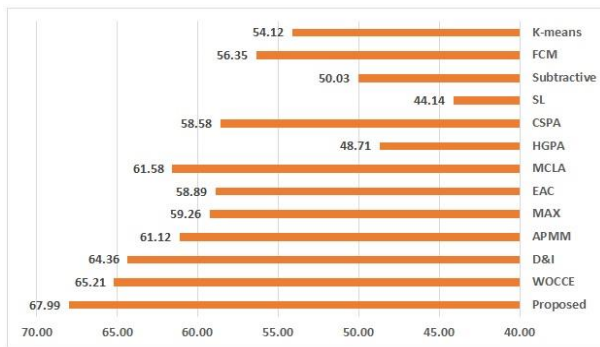
شماره	نام مجموعه داده	تعداد ویژگی‌ها	تعداد کلاس‌ها	تعداد نمونه‌ها
۱	Ionosphere	۳۴	۲	۳۵۱
۲	Optdigits	۶۲	۱۰	۵۶۲۰
۳	Pendigits	۱۶	۱۰	۱۰۹۹۲
۴	Statlog	۳۶	۷	۶۴۳۵
۵	Wine	۱۳	۲	۱۷۸
۶	Yeast	۸	۱۰	۱۴۸۴
۷	Half Ring	۲	۲	۴۰۰
۸	Iris	۴	۳	۱۵۰
۹	Balance Scale	۴	۳	۶۲۵
۱۰	Breast Cancer	۹	۲	۶۸۳
۱۱	Bupa	۶	۲	۳۴۵
۱۲	Galaxy	۴	۷	۳۲۳
۱۳	Glass	۹	۶	۲۱۴
۱۴	SA Heart	۹	۲	۴۶۲

برای انجام آزمایش‌ها، سعی شده است که مجموعه داده‌ها از لحاظ تعداد کلاس‌ها، تعداد ویژگی‌ها و همچنین تعداد نمونه‌ها از حداکثر تنوع برخوردار باشند. در نتیجه نتایج آزمایش‌ها تا حد ممکن مستحکم و قابل تعمیم خواهد بود. جدول ۱، اطلاعات مختصری از این مجموعه داده‌ها در اختیار می‌گذارد. برای اطلاعات بیشتر در مورد هر کدام از این مجموعه داده‌ها، به [۲۲، ۴۸] رجوع کنید. نتایج آزمایش‌ها بر روی ویژگی‌های نرمال شده از این مجموعه داده‌ها گزارش شده است.

##### ۵-۲- الگوریتم‌های خوشه‌بندی پایه

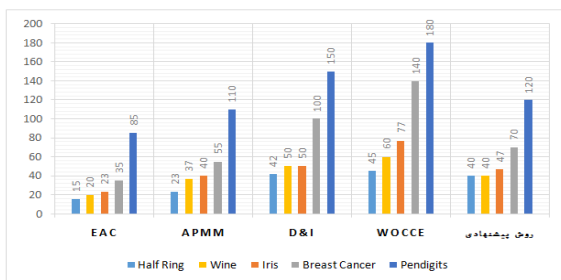
همان‌طور که در بخش عدم تمرکز به آن اشاره شد، یکی دیگر از الزامات روش خوشه‌بندی بخردانه، استفاده از الگوریتم‌های پایه غیرهم‌نام در فرآیند تولید نتایج اولیه است. این مقاله عیناً از همان الگوریتم‌های روش‌های پیشین جهت تولید نتایج اولیه استفاده کرده است [۱۴، ۱۸]. جدول ۲، فهرست این الگوریتم‌ها را نشان می‌دهد.





شکل ۲: میانگین دقت الگوریتم‌های خوشه‌بندی مورد آزمایش

مطابق با شکل ۲، با این که روش ترکیب کامل (EAC)، تمامی افزایش‌ها را نسبت به سایر روش‌های ترکیبی مبتنی بر انتخاب داراست. این مسئله در داده‌های بزرگ بهتر خود را نشان می‌دهد، به طوری که مطابق با جدول ۳، در سه مجموعه داده Optdigits، Pendigits و Statlogs، تشخیص الگوی صحیح با ترکیب تمامی نتایج، بسیار سخت حاصل می‌شود. از طرف دیگر، چهارچوب غیرمتمرکز خوشه‌بندی بخردانه که در شکل ۱ ارائه شده است، یکی از مهم‌ترین دلایل بهتر کارکردن الگوریتم پیشنهادی این مقاله در داده‌های نوع سخت می‌باشد، چون در سایر روش‌های مبتنی بر انتخاب، تغییرات در مقادیر ارزیابی معیار پراکندگی، قبل و پس از انتخاب خوشه، بر روی دقت نتیجه نهایی تأثیر بسیاری می‌گذارد. علاوه بر آن، می‌توان به تأثیرات سایر معیارهای پیشنهادی در خوشه‌بندی بخردانه، از جمله معیار استقلال نیز اشاره کرد. از طرفی دیگر، با اینکه در روش‌های WOCCE، APMM و MAX نیز از معیارهایی مشابه با روش پیشنهادی این مقاله جهت ارزیابی نتایج اولیه استفاده شده است، ولی روش پیشنهادی، به علت تخمین درست مقادیر آستانه، بهتر از آن روش‌ها عمل کرده است. قابل ذکر است، که روش پیشنهادی در مجموع تقریباً سه درصد بهتر از روش WOCCE عمل کرده است، ولی در این روش دیگر نیاز به تعیین مقادیر آستانه نیست. همچنین مطابق نتایج شکل ۳، زمان اجرای الگوریتم پیشنهادی این مقاله نسبت به سایر روش‌های خوشه‌بندی بخردانه به صورت چشمگیری کاهش پیدا کرده است و تقریباً نزدیک به روش APMM، که روشی تک معیاره است، می‌باشد.



شکل ۳: مقایسه زمان اجرای روش‌های استفاده کننده از نظریه خرد جمعی

جدول ۲: الگوریتم‌های پایه [۱۴، ۱۸]

شماره	نام الگوریتم
۱	K-Means
۲	Fuzzy C-Means
۳	Median K-Flats
۴	Gaussian Mixture
۵	Subtractive Clustering
۶	Single-Linkage Euclidean
۷	Single-Linkage Hamming
۸	Single-Linkage Cosine
۹	Average-Linkage Euclidean
۱۰	Average-Linkage Hamming
۱۱	Average-Linkage Cosine
۱۲	Complete-Linkage Euclidean
۱۳	Complete-Linkage Hamming
۱۴	Complete-Linkage Cosine
۱۵	Ward-Linkage Euclidean
۱۶	Ward-Linkage Hamming
۱۷	Ward-Linkage Cosine
۱۸	Spectral clustering using a sparse similarity matrix
۱۹	Spectral clustering using Nystrom method with orthogonalization
۲۰	Spectral clustering using Nystrom method without orthogonalization

### ۵-۳- نتایج آزمایش

روش پیشنهادی در محیط MATLAB 7.1 پیاده‌سازی و مورد آزمایش قرار گرفته است و نتایج آزمایش‌ها، روی میانگین ۱۰ بار اجرای مستقل برنامه، گزارش شده‌اند. عملکرد روش‌های مختلف خوشه‌بندی، با استفاده از فرایند بازبرچسب‌گذاری بین خوشه‌های به دست آمده و کلاس‌های واقعی و مقایسه آنها، محاسبه شده است [۱۴، ۱۸]. جدول ۳، عملکرد روش‌های مختلف را در مقایسه با روش پیشنهادی این مقاله نشان می‌دهد.

همان‌طور که در جدول ۳ مشاهده می‌شود، روش پیشنهادی در اکثر موارد بهتر از سایر روش‌ها عمل کرده است. همچنین، در برخی از داده‌ها همانند (Glass، Ionosphere و Yeast)، با اینکه شرایط کاملاً بهبود نیافته، اختلاف نتایج با بهترین روش، مقدار کمی می‌باشد. با اینکه روش Subtractive بهترین نتیجه را روی داده Ionosphere ایجاد کرده است ولی در سایر داده‌ها دقت مناسبی نداشته است. این مسئله، مثال مناسبی برای نشان دادن وابستگی کارایی الگوریتم‌های خوشه‌بندی پایه به جنبه‌های خاص از مجموعه داده، و همچنین دلیلی محکم جهت به کارگیری روش‌های ترکیبی می‌باشد.

شکل ۲، میانگین دقت الگوریتم‌های خوشه‌بندی را در آزمایش‌های تجربی جدول ۳ نشان می‌دهد.

جدول ۳: مقایسه دقت (Accuracy) روش‌های مختلف خوشه‌بندی بر حسب درصد. ستون دقت، مربوط به میانگین (Mean) و انحراف از معیار (Std) دقت حاصل از ۱۰ بار اجرای هر یک از الگوریتم‌های مورد آزمایش، بر روی هر یک از مجموعه داده‌های جدول ۱، می‌باشد.

شماره مجموعه داده	دقت	الگوریتم‌های خوشه‌بندی ترکیبی مبتنی بر انتخاب											الگوریتم‌های خوشه‌بندی پایه			
		Proposed Method [۱۴]	WOCCE [۱۴]	D&I [۱۹]	APMM [۲۱]	MAX [۲۲]	EAC [۳۷]	MCLA [۱۷]	HGPA [۱۷]	CSPA [۱۷]	Single-Linkage	Subtractive	FCM	K-means		
۱	میانگین	۷۳/۶۷	۷۰/۵۲	۶۹/۲۱	۷۰/۹۴	۶۴/۴۸	۶۷/۸۰	۷۱/۲۲	۵۸/۴۰	۷۰/۴۸	۶۴/۳۸	۷۷/۰۰	۶۷/۸۰	۶۵/۵۱		
	انحراف	۰/۳۴۱	۰/۱۳۲	۰/۷۴۰	۰/۱۳۰	۰/۹۱۴	۱/۱۱۸	۰/۲۱۰	۱/۳۷۸	۰/۱۲۱	۱/۳۰۴	۰/۰۱۲	۰/۹۷۴	۱/۳۴۲		
۲	میانگین	۷۸/۵۶	۷۷/۱۶	۷۷/۵۹	۷۷/۱۰	۷۶/۱۱	۴۸/۱۲	۷۷/۱۵	۶۴/۷۷	۷۵/۲۱	۱۰/۲۸	۴۷/۷۲	۳۸/۳۳	۴۷/۲۳		
	انحراف	۰/۶۹۲	۰/۲۱۰	۰/۶۹۰	۰/۸۴۱	۰/۶۵۰	۰/۵۰۳	۰/۴۵۲	۰/۱۹۸	۰/۶۴۳	۲/۲۰۲	۱/۳۱۲	۰/۹۲۱	۰/۲۴۱		
۳	میانگین	۶۴/۱۳	۵۸/۶۸	۵۹/۸۷	۴۷/۴۰	۵۷/۰۲	۴۳/۹۰	۵۸/۶۲	۴۷/۵۵	۵۸/۳۲	۱۰/۴۶	۱۰/۴۰	۳۶/۷۷	۴۰/۹۷		
	انحراف	۰/۴۲۰	۰/۱۸۰	۰/۸۱۰	۰/۶۹۹	۰/۵۲۱	۰/۴۳۰	۰/۷۳۰	۰/۳۳۱	۰/۹۰۳	۳/۹۲۰	۰/۹۵۶	۱/۰۲۰	۱/۶۹۰		
۴	میانگین	۵۷/۷۶	۵۵/۷۷	۵۵/۴۶	۵۴/۸۸	۵۴/۲۳	۴۳/۹۶	۵۵/۷۱	۵۲/۹۴	۵۴/۲۳	۲۳/۸۰	۲۳/۸۰	۴۹/۹۱	۴۰/۸۹		
	انحراف	۰/۵۹۱	۰/۷۱۹	۰/۱۸۰	۰/۵۲۸	۰/۱۴۰	۰/۸۱۷	۰/۳۴۲	۰/۴۹۱	۰/۹۵۶	۱/۵۶۲	۰/۹۳۴	۲/۱۴۰	۱/۸۳۱		
۵	میانگین	۷۴/۴۶	۷۱/۳۴	۷۰/۱۹	۶۴/۶۰	۶۹/۱۷	۷۰/۵۶	۷۰/۲۲	۶۲/۳۶	۶۷/۴۱	۳۷/۶۴	۶۷/۲۳	۷۱/۳۴	۶۵/۷۳		
	انحراف	۰/۱۴۱	۰/۵۴۲	۰/۲۴۰	۰/۲۳۱	۰/۷۸۹	۰/۸۹۰	۰/۳۳۰	۰/۲۰۲	۰/۹۳۱	۱/۳۲۱	۰/۹۱۰	۰/۷۸۵	۰/۵۱۰		
۶	میانگین	۳۰/۱۲	۳۲/۷۶	۳۱/۹۲	۳۱/۰۶	۳۲/۴۰	۳۱/۷۴	۱۷/۵۶	۱۵/۲۳	۱۴/۰۰	۲۹/۷۳	۳۱/۲۰	۲۹/۹۸	۳۱/۱۹		
	انحراف	۰/۴۶۲	۰/۲۶۸	۰/۸۳۰	۰/۲۴۵	۰/۱۲۴	۰/۲۳۴	۰/۹۶۲	۰/۳۱۰	۰/۳۱۰	۱/۱۸۶	۱/۵۷۰	۰/۳۴۱	۰/۶۹۲		
۷	میانگین	۹۷/۸۰	۸۷/۲۰	۸۶/۳۸	۸۰/۰۰	۷۸/۴۸	۷۷/۱۷	۷۴/۵۰	۵۰/۰۰	۷۴/۵۰	۷۴/۵۰	۸۶/۰۰	۷۸/۰۰	۷۵/۷۵		
	انحراف	۰/۳۱۰	۰/۱۴۰	۰/۹۲۰	۰/۴۲۰	۰/۵۹۰	۰/۹۲۰	۰/۳۸۰	۰/۳۱۰	۰/۸۵۰	۱/۱۹۰	۱/۶۲۰	۰/۹۲۱	۱/۳۴۰		
۸	میانگین	۹۷/۳۰	۹۲/۰۰	۸۹/۴۳	۷۴/۱۱	۷۲/۸۹	۹۶/۰۰	۸۹/۳۴	۴۸/۶۶	۸۵/۳۴	۶۸/۰۰	۵۵/۳۰	۸۲/۶۶	۶۵/۳۰		
	انحراف	۰/۶۲۰	۰/۵۹۰	۰/۷۲۰	۰/۲۵۰	۰/۳۲۰	۱/۹۳۰	۰/۶۶۰	۰/۹۲۰	۰/۱۹۰	۰/۹۹۰	۱/۱۱۰	۱/۰۴۰	۱/۴۶۰		
۹	میانگین	۵۸/۶۴	۵۴/۸۸	۵۴/۹۲	۵۲/۶۵	۵۲/۱۰	۵۲/۰۰	۵۱/۳۶	۴۱/۲۸	۵۱/۸۴	۴۶/۴۰	۴۵/۳۲	۴۴/۰۰	۴۰/۳۲		
	انحراف	۰/۵۸۰	۰/۶۱۰	۰/۴۲۰	۰/۶۳۰	۰/۹۱۰	۰/۲۸۰	۰/۷۱۰	۰/۱۱۰	۰/۸۲۰	۰/۸۲۰	۱/۹۲۰	۱/۶۷۰	۲/۱۷۰		
۱۰	میانگین	۹۸/۰۰	۹۶/۹۲	۹۶/۳۷	۹۶/۰۴	۷۵/۷۲	۹۵/۰۲	۹۶/۰۵	۵۰/۳۷	۸۰/۹۷	۶۵/۱۵	۶۵/۰۰	۹۶/۴۳	۹۳/۷۰		
	انحراف	۰/۱۴۰	۰/۷۷۰	۰/۲۹۰	۰/۱۸۸۰	۰/۴۰۰	۰/۳۹۰	۰/۵۲۰	۰/۲۳۰	۰/۲۷۰	۱/۰۴۰	۱/۷۵۰	۱/۳۱۰	۱/۹۴۰		
۱۱	میانگین	۶۰/۸۳	۵۷/۰۲	۵۶/۳۲	۵۵/۰۷	۵۶/۱۷	۵۵/۱۸	۵۵/۳۶	۵۰/۷۲	۵۶/۲۳	۵۷/۶۸	۵۳/۹۷	۵۰/۱۰	۵۴/۴۹		
	انحراف	۰/۱۲۰	۰/۴۶۰	۰/۹۱۰	۰/۲۸۰	۰/۲۸۰	۰/۵۱۰	۰/۱۷۰	۰/۸۶۰	۰/۵۹۰	۱/۳۳۰	۰/۲۵۰	۰/۸۳۰	۰/۹۸۰		
۱۲	میانگین	۳۷/۱۸	۳۵/۸۸	۳۴/۶۱	۳۳/۷۲	۳۲/۷۸	۳۱/۹۵	۲۸/۴۸	۳۱/۲۷	۲۹/۴۱	۲۵/۰۷	۲۹/۷۲	۳۴/۹۸	۳۰/۰۳		
	انحراف	۰/۶۷۰	۰/۸۱۰	۰/۵۲۰	۰/۳۶۰	۰/۶۹۰	۰/۹۲۰	۰/۴۲۰	۰/۹۰۰	۰/۳۸۰	۱/۱۶۰	۰/۸۲۰	۱/۴۸۰	۰/۴۸۰		
۱۳	میانگین	۵۱/۰۰	۵۱/۸۲	۵۰/۷۴	۴۷/۱۹	۴۴/۱۷	۴۵/۹۳	۵۱/۴۰	۴۱/۱۲	۳۸/۷۸	۳۶/۴۴	۳۶/۴۴	۴۷/۱۹	۴۲/۰۵		
	انحراف	۰/۷۸۰	۰/۹۲۰	۰/۳۴۰	۰/۲۱۰	۰/۴۷۰	۰/۶۳۰	۰/۸۵۰	۱/۳۵۰	۱/۸۲۰	۱/۴۸۰	۱/۲۵۰	۰/۸۷۰	۱/۳۷۰		
۱۴	میانگین	۷۲/۴۷	۶۸/۷۰	۶۷/۹۹	۷۰/۹۱	۶۳/۹۶	۶۵/۱۹	۶۲/۵۴	۵۰/۹۳	۵۸/۴۲	۶۵/۱۵	۶۷/۲۶	۶۳/۴۱	۶۴/۵۱		
	انحراف	۰/۲۵۰	۰/۴۶۰	۰/۷۴۰	۰/۴۲۰	۰/۷۲۰	۱/۹۲۰	۰/۵۷۰	۱/۹۳۰	۰/۷۴۰	۰/۹۲۰	۱/۱۷۰	۰/۶۱۰	۰/۷۴۰		

در انتها، به بررسی نتایج آزمایش‌های انجام شده به منظور تحلیل اثر نویز بر روی عملکرد روش پیشنهادی می‌پردازیم. در این آزمایش‌ها، از مجموعه داده‌های Optrdigits و Pendigits استفاده شد، چرا که این دو مجموعه داده، نسبتاً دارای تعداد ویژگی‌ها و نمونه‌های زیادی می‌باشند. شکل ۴ عملکرد روش پیشنهادی، WOCCE، D&I، APMM و MAX را بر روی مجموعه داده‌های ذکر شده با درصد‌های متفاوتی از نویز، نشان می‌دهد. به منظور اضافه کردن نویز، برخی از ویژگی‌های مجموعه داده‌های مذکور به طور تصادفی تغییر داده شدند. با توجه به شکل ۴، می‌توان دریافت که در آزمایش‌های انجام شده در این قسمت، روش پیشنهادی نتایج پایدارتری را تولید کرده است.

همان‌طور که پیش‌تر گفته شد، نتایج آزمایش‌های انجام شده در این پژوهش، که در جدول ۳ قابل مشاهده می‌باشند، بیان‌گر عملکرد بهتر روش پیشنهادی در مقایسه با دیگر روش‌ها هستند. اما برای حصول اطمینان از این امر که عملکرد روش پیشنهادی، از نظر آماری نیز با دیگر روش‌ها متفاوت است، از آزمون فریدمن<sup>۲۴</sup> پیاده‌سازی شده در نرم افزار IBM SPSS Statistics 22 استفاده شد. نتایج تحلیل‌ها، بیان‌گر تفاوت آماری قابل توجهی بین عملکرد روش‌های مورد مقایسه می‌باشند ( $p < 0.01$ ,  $df = 12$  و  $95.372642 = \text{مجذور کای}$ ). همچنین نتیجه رتبه‌بندی فریدمن، که در جدول ۴ قابل مشاهده است، بیان‌گر این مورد می‌باشد که روش پیشنهادی، بهترین عملکرد را، با کسب میانگین رتبه ۱۲/۲۹، در بین دیگر روش‌ها از خود نشان داده است.

بخردانه، مقادیر آستانه تعیین شده برای معیارهای ارزیابی، تأثیر قابل توجهی در کارایی و زمان اجرای الگوریتم دارند و در روش‌های پیشین، هیچ رویکردی جهت تعیین این مقادیر ارائه نشده است. در این مقاله رویکردی جدید، جهت تخمین خودکار مقادیر آستانه‌ای به صورت بهینه، بر اساس ویژگی‌های اصلی داده ورودی معرفی شده است. علاوه بر این، به منظور اندازه‌گیری پراکندگی دو خوشه‌بندی پایه، معیاری جدید تحت عنوان همگونی بر اساس معیار APMM ارائه شده که موجب حفظ کیفیت ارزیابی نتایج اولیه می‌شود. همچنین روش جدیدی تحت عنوان روش انباشت مدارک وزن دار، به منظور در نظر گرفتن استقلال به عنوان وزنی در ترکیب نتایج اولیه، ارائه شده است.

نتایج تجربی روش پیشنهادی مقاله بر روی ۱۴ مجموعه داده مختلف و متنوع نشان می‌دهند که این روش، نسبت به روش‌های متداول پایه، ترکیبی و بخردانه، برتری قابل ملاحظه‌ای دارد. همچنین، بررسی‌ها نشان می‌دهند که اگرچه روش پیشنهادی، از زیرمجموعه کوچکی از نتایج خوشه‌بندی‌های اولیه استفاده می‌کند، اما به خاطر مؤثر بودن این زیرمجموعه‌ها و همچنین حذف خوشه‌ها با کیفیت پایین و تکراری که تأثیر منفی روی میزان همبستگی واقعی نمونه‌ها دارند، نتیجه نهایی آن حتی از ترکیب کامل (EAC) هم بهتر می‌باشد.

جدول ۴: میانگین رتبه به دست آمده روش‌های حاضر در آزمایش‌ها، بر

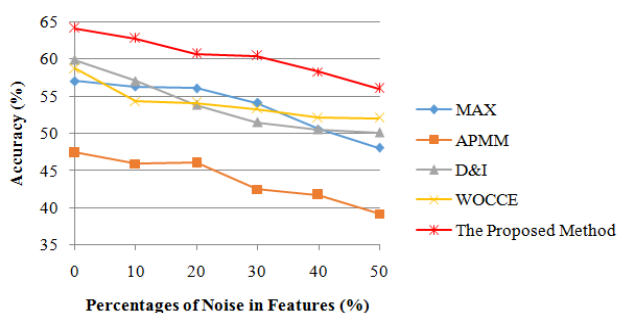
اساس آزمون فریدمن

نام الگوریتم	میانگین رتبه
Proposed Method	۱۲/۲۹
WOCCE	۱۱/۴۶
D&I	۱۰/۲۹
APMM	۸/۰۴
MAX	۷/۰۴
EAC	۷/۲۵
MCLA	۷/۵۴
HGPA	۳/۰۰
CSPA	۵/۵۷
Single-Linkage	۳/۵۴
Subtractive	۴/۹۳
FCM	۵/۷۵
K-means	۴/۳۲

## ۶- نتیجه

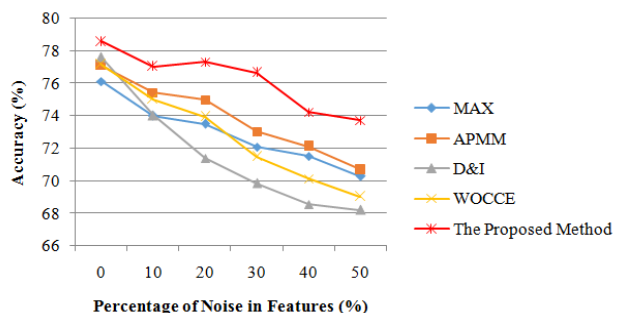
در این مقاله، روشی جدید بر اساس خوشه‌بندی بخردانه، برای حل مسائل خوشه‌بندی پیشنهاد شده است. از آنجایی که در خوشه‌بندی

Pendigits



ب: عملکرد روش پیشنهادی، WOCCE، D&I، APMM و MAX با اعمال آن‌ها به مجموعه داده Pendigits با درصدهای متفاوتی از نویز

Optdigits



الف: عملکرد روش پیشنهادی، WOCCE، D&I، APMM و MAX با اعمال آن‌ها به مجموعه داده Optdigits با درصدهای متفاوتی از نویز

شکل ۴: عملکرد روش پیشنهادی، WOCCE، D&I، APMM و MAX با اعمال آن‌ها به دو مجموعه داده Pendigits و Optdigits با درصدهای متفاوتی از نویز

[۳] سیامک عبدالله‌زاده، محمدعلی بالافر و لیلی محمدخانلی، «استفاده از خوشه‌بندی و مدل مارکوف جهت پیش‌بینی درخواست آتی کاربر در وب»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۵، شماره ۳، صفحه ۹۶-۸۹، پاییز ۱۳۹۴.

[۴] یوکابد صدیقی، علی آقاگل‌زاده و مهدی ازوجی، «ادغام تصاویر چندفوکوسه با استفاده از همدوسی فاز و خوشه‌بند K-means»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۵، شماره ۴، صفحه ۱۲۷-۱۱۷، زمستان ۱۳۹۴.

[۵] رضا خدایی، محمدعلی بالافر و سیدناصر رضوی، «اثربخشی بسط پرس‌وجو مبتنی بر خوشه‌بندی اسناد شبه‌بازخورد با الگوریتم K-

## مراجع

- [۱] سمیرا رفیعی و پرهام مرادی، «بهبود عملکرد الگوریتم خوشه‌بندی فازی سی-مینز با وزن‌دهی اتوماتیک و محلی ویژگی‌ها»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۶، شماره ۲، صفحه ۸۶-۷۵، تابستان ۱۳۹۵.
- [۲] علیرضا سردار و رمضان هاونگی، «بهبود عملکرد الگوریتم خوشه‌یابی خودکار تصاویر رنگی به کمک پیش‌پردازش با شبکه عصبی خودسازمانده»، مجله مهندسی برق دانشگاه تبریز، جلد ۴۷، شماره ۳، صفحه ۱۰۸۲-۱۰۷۳، پاییز ۱۳۹۶.

- [21] H. Alizadeh, B. Minaei-Bidgoli, and H. Parvin, "Cluster ensemble selection based on a new cluster stability measure," *Intelligent Data Analysis*, vol. 18, pp. 389-408, 2014.
- [22] H. Alizadeh, H. Parvin, and S. Parvin, "A framework for cluster ensemble based on a max metric as cluster evaluator," *IAENG International Journal of Computer Science*, vol. 39, pp. 10-19, 2012.
- [23] X. Z. Fern and W. Lin, "Cluster ensemble selection," *Statistical Analysis and Data Mining*, vol. 1, pp. 128-141, 2008.
- [24] A. K. Jain, A. Topchy, M. H. Law, and J. M. Buhmann, "Landscape of clustering algorithms," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, 2004, pp. 260-263.
- [25] J. Surowiecki, "The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business," *Economies, Societies and Nations*, vol. 296, 2004.
- [26] D. Yang, G. Xue, X. Fang, and J. Tang, "Crowdsourcing to smartphones: Incentive mechanism design for mobile phone sensing," in *Proceedings of the 18th Annual International Conference on Mobile Computing and Networking*, 2012, pp. 173-184.
- [27] L. Baker and D. Ellison, "The wisdom of crowds—ensembles and modules in environmental modelling," *Geoderma*, vol. 147, pp. 1-7, 2008.
- [28] B. Miller, P. Hemmer, M. Steyvers, and M. D. Lee, "The wisdom of crowds in rank ordering problems," in *9th International Conference on Cognitive Modeling*, 2009.
- [29] M. Steyvers, B. Miller, P. Hemmer, and M. D. Lee, "The wisdom of crowds in the recollection of order information," in *Advances in Neural Information Processing Systems*, 2009, pp. 1785-1793.
- [30] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds," in *Advances in Neural Information Processing Systems*, 2010, pp. 2424-2432.
- [31] D. P. Williams, "Underwater mine classification with imperfect labels," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 4157-4161.
- [32] S. K. Yi, M. Steyvers, M. Lee, and M. Dry, "Wisdom of the crowds in minimum spanning tree problems," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2010.
- [33] K. Faceli, A. C. De Carvalho, and M. C. De Souto, "Multi-objective clustering ensemble," *International Journal of Hybrid Intelligent Systems*, vol. 4, pp. 145-156, 2007.
- [34] H. G. Ayad and M. S. Kamel, "Cumulative voting consensus method for partitions with variable number of clusters," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 160-173, 2008.
- [35] A. Topchy, A. K. Jain, and W. Punch, "Combining multiple weak clusterings," in *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, 2003, pp. 331-338.
- [36] H. G. Ayad and M. S. Kamel, "Cluster-based cumulative ensembles," in *International Workshop on Multiple Classifier Systems*, 2005, pp. 236-245.
- [6] مجید محمدپور و حمید پروین، «الگوریتم ژنتیک آشوب گونه مبتنی بر حافظه و خوشه بندی برای حل مسائل بهینه سازی پویا»، *مجله مهندسی برق دانشگاه تبریز*، جلد ۴۶، شماره ۳، صفحه ۳۱۸-۲۹۹، پاییز ۱۳۹۵.
- [7] X. Wu, T. Ma, J. Cao, Y. Tian, and A. Alabdulkarim, "A comparative study of clustering ensemble algorithms," *Computers & Electrical Engineering*, vol. 68, pp. 603-615, 2018.
- [8] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys (CSUR)*, vol. 31, pp. 264-323, 1999.
- [9] F. Yang, T. Li, Q. Zhou, and H. Xiao, "Cluster ensemble selection with constraints," *Neurocomputing*, vol. 235, pp. 59-70, 2017.
- [10] L. Bai, J. Liang, and Y. Guo, "An ensemble clusterer of multiple fuzzy k-means clusterings to recognize arbitrarily shaped clusters," *IEEE Transactions on Fuzzy Systems*, 2018.
- [11] J. Bai, S. Song, T. Fan, and L. Jiao, "Medical image denoising based on sparse dictionary learning and cluster ensemble," *Soft Computing*, pp. 1-7, 2017.
- [12] V. Berikov, N. Karaev, and A. Tewari, "Semi-supervised classification with cluster ensemble," in *Engineering, Computer and Information Sciences (SIBIRCON), 2017 International Multi-Conference on*, 2017, pp. 245-250.
- [13] H. Alizadeh, *Cluster Ensemble Selection Based on Mathematical and Social Optimization Methods* (in Persian), PhD Thesis. Iran University of Science and Technology, 2014.
- [14] H. Alizadeh, M. Yousefnezhad, and B. M. Bidgoli, "Wisdom of Crowds cluster ensemble," *Intelligent Data Analysis*, vol. 19, pp. 485-503, 2015.
- [15] A. Fred and A. Lourenço, "Cluster ensemble methods: from single clusterings to combined solutions," in *Supervised and Unsupervised Ensemble Methods and Their Applications*, ed: Springer, 2008, pp. 3-30.
- [16] A. L. Fred and A. K. Jain, "Data clustering using evidence accumulation," in *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 2002, pp. 276-280.
- [17] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, vol. 3, pp. 583-617, 2002.
- [18] M. Yousefnezhad, *Cluster Ensemble Selection Based on the Wisdom of Crowds* (in Persian), MSc Thesis, Mazandaran University of Science and Technology, 2013.
- [19] M. Yousefnezhad, H. Alizadeh, and B. Minaei-Bidgoli, "New cluster ensemble selection method based on diversity and independent metrics (in Persian)," in *5th Conference on Information and Knowledge Technology (IKT'13)*, 2013, pp. 22-24.
- [20] M. Yousefnezhad and D. Zhang, "Weighted spectral cluster ensemble," in *Data Mining (ICDM), 2015 IEEE International Conference on*, 2015, pp. 549-558.

- CAMP 2006. International Workshop on, 2006, pp. 119-123.
- [43] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Biocomputing 2002*, ed: World Scientific, 2001, pp. 6-17.
- [44] T. Lange, V. Roth, M. L. Braun, and J. M. Buhmann, "Stability-based validation of clustering solutions," *Neural Computation*, vol. 16, pp. 1299-1323, 2004.
- [45] P.-Y. Mok, H. Huang, Y. Kwok, and J. Au, "A robust adaptive clustering analysis method for automatic identification of clusters," *Pattern Recognition*, vol. 45, pp. 3017-3033, 2012.
- [46] K. Arai and A. R. Barakbah, *Hierarchical K-means: an algorithm for centroids initialization for K-means*, Reports of the Faculty of Science and Engineering, vol. 36, pp. 25-31, 2007.
- [47] D. Pelleg and A. W. Moore, "X-means: Extending k-means with efficient estimation of the number of clusters," in *Icml*, 2000, pp. 727-734.
- [48] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz. *{UCI} Repository of machine learning databases*, 1998.
- [37] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 835-850, 2005.
- [38] L. I. Kuncheva and S. T. Hadjitodorov, "Using diversity in cluster ensembles," in *Systems, Man and Cybernetics, 2004 IEEE International Conference on*, 2004, pp. 1214-1219.
- [39] A. L. Fred and A. K. Jain, "Learning pairwise similarity for data clustering," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 2006, pp. 925-928.
- [40] J. Azimi, J. Maani, and N. Mozayyeni, "Improved Clustering Ensembles (in Persian)," presented at the 11th International CSI Computer Conference (CSICC06), 2006.
- [41] J. Azimi and M. Analoui, "Distinguishing Marginal Samples to Improve Clustering Ensembles (in Persian)," presented at the 11th International CSI Computer Conference (CSICC06), 2006.
- [42] J. Azimi, M. Mohammadi, and M. Analoui, "Clustering ensembles using genetic algorithm," in *Computer Architecture for Machine Perception and Sensing*, 2006.

## زیر نویس‌ها

<sup>14</sup> Uniformity

<sup>15</sup> Weighted Evidence Accumulation Clustering (Weighted EAC)

<sup>16</sup> Robustness

<sup>17</sup> Novelty

<sup>18</sup> Stability

<sup>19</sup> Flexibility

<sup>20</sup> Consensus Function

<sup>21</sup> Accuracy

<sup>22</sup> Full Ensemble

<sup>23</sup> Sum of Normalized Mutual Information

<sup>24</sup> Friedman Test

<sup>25</sup> Chi-Square

<sup>1</sup> Arbitrarily shaped clusters

<sup>2</sup> Denoising

<sup>3</sup> Semi-Supervised Classification

<sup>4</sup> Diversity

<sup>5</sup> Cluster Ensemble Selection

<sup>6</sup> Feedback mechanism

<sup>7</sup> Normalized Mutual Information

<sup>8</sup> Accurate

<sup>9</sup> Stable

<sup>10</sup> Robust

<sup>11</sup> Crowd Computing

<sup>12</sup> The Wisdom of Crowds

<sup>13</sup> Crowd Sourcing