انتخاب خوشه‌های اولیه به کمک الگوریتم‌های هوشمند برای مشارکت در خوشه‌بندی ترکیبی

محمدجواد حسین‌پور[1] حمید پروین[2]

۱- دانشکده مهندسی کامپیوتر، دانشکده آزاد اسلامی، واحد استهبان، استهبان، فارس، ایران
hoseinpoor@iauest.ac.ir

۲- استادیار- دانشکده مهندسی کامپیوتر، دانشکده آزاد اسلامی، واحد نورآباد ممسنی، نورآباد، فارس، ایران
باشگاه پژوهشگران جوان و نخبگان، واحد نورآباد ممسنی، دانشگاه آزاد اسلامی، نوراباد ممسنی، ایران
parvin@alumni.iust.ac.ir

چکیده: **به علت بدون ناظر بودن مسئله خوشه‌بندی انتخاب الگوریتمی خاص جهت خوشه‌بندی یک مجموعه ناشناس امری پر خطر و معمولا شکست خورده می‌باشد. به خاطر پیچیدگی مسئله و ضعف روش‌های خوشه‌بندی پایه، امروزه اکثر مطالعات به سمت روش‌های خوشه‌بندی ترکیبی هدایت شده است. پراکندگی در نتایج اولیه یکی از مهم‌ترین عواملی است که می‌تواند در کیفیت نتایج نهایی اثرگذار باشد. همچنین، کیفیت نتایج اولیه نیز عامل دیگری است که در کیفیت نتایج حاصل از ترکیب موثر است. هر دو عامل در تحقیقات اخیر خوشه‌بندی ترکیبی مورد توجه قرار گرفته‌اند. در اینجا یک چارچوب برای بهبود کارایی خوشه‌بندی پیشنهاد شده است که مبتنی بر استفاده از زیرمجموعه‌ای از خوشه‌های اولیه می‌باشند. انتخاب این زیرمجموعه نقش حیاتی در کارایی مجمع دارد. این انتخاب به کمک دو روش هوشمند انجام می‌گیرد. ایده‌های اصلی در روش‌های پیشنهادی برای انتخاب زیرمجموعه‌ای از خوشه‌ها، استفاده از خوشه‌های پایدار با الگوریتم‌های جستجوی هوشمند می‌باشند. برای ارزیابی خوشه‌ها، از معیار پایداری مبتنی بر اطلاعات متقابل استفاده شده است. در آخر نیز خوشه‌های انتخاب شده را به کمک چندین روش ترکیب نهایی با هم جمع می‌کنیم. نتایج تجربی روی چندین مجموعه داده استاندارد نشان می‌دهد که روش‌های پیشنهادی می‌توانند به طور موثری همچنین روش ترکیب کامل را بهبود دهند.**

کلمات کلیدی: **خوشه‌بندی ترکیبی، ارزیابی خوشه، اطلاعات متقابل، زیرمجموعه‌ای از نتایج اولیه، الگوریتم ژنتیک، الگوریتم نورد شبیه‌سازی شده، خوشه‌بندی انباشت مدارک، ماتریس همبستگی.**

تاریخ ارسال مقاله: ۱۳۹۳/۵/۱
تاریخ پذیرش مشروط مقاله: ۱۳۹۴/۹/۱۲
تاریخ پذیرش مقاله: ۱۳۹۵/۳/۱
نام نویسنده‌ی مسئول: حمید پروین
نشانی نویسنده‌ی مسئول: ایران – فارس – نوراباد ممسنی – دانشگاه آزاد اسلامی – دانشکده‌ی مهندسی کامپیوتر

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol13 No.2 Summer 2016

۱- مقدمه

ایده اصلی خوشه‌بندی اطلاعات، جداکردن 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 [۵] و [۶]. در واقع خوشه‌بندی داده‌ها یک ابزار ضروری برای یافتن گروه‌ها در داده‌های بدون برچسب است [۷]. حداقل ۵ دلیل اصلی برای اهمیت خوشه‌بندی وجود دارد.

۱- جمع‌آوری و برچسب‌گذاری یک مجموعه بزرگ از الگوهای نمونه می‌تواند بسیار باارزش باشد.

۲- ممکن است ما به دنبال‌کردن در جهت معکوس علاقمند باشیم. یعنی آموزش با مقدار زیاد داده‌های بدون برچسب و سپس تنها استفاده از ناظر برای برچسب‌گذاری خوشه‌های پیدا شده. این می‌تواند برای کاربردهای داده‌کاوی[۱] بزرگ که محتویات یک پایگاه داده از قبل شناخته شده نیست، مناسب باشد.

۳- در خیلی از کاربردها مشخصه‌های الگوها می‌توانند به آهستگی با زمان تغییر کنند، مثل رده‌بندی[۲] خودکار مواد غذایی با تغییر فصل. اگر این تغییرات بتواند با یک رده‌بند[۳] به صورت بدون ناظر[۴] رهگیری[۵] شود، عملکرد بهتری می‌تواند به دست آید.

۴- می‌توانیم از روش‌های بدون‌ناظر (خوشه‌بندی) برای پیداکردن و استخراج ویژگی‌ها استفاده کنیم.

۵- با خوشه‌بندی می‌توانیم یک دید و بینشی از از طبیعت و ساختار داده به دست آوریم که این می‌تواند برای ما با ارزش باشد. کشف زیررده‌های[۶] مجزا یا شباهت‌های بین الگوها ممکن است به طور چشمگیری در روش طراحی رده‌بندی‌کننده به ما پیشنهاد ارایه کند.

ما در جهانی پر از داده زندگی می‌کنیم. هر روز انسان‌ها با حجم وسیعی از اطلاعات مواجه هستند که باید آنها را ذخیره‌سازی یا نمایش دهند. یکی از روش‌های حیاتی کنترل و مدیریت این داده‌ها رده‌بندی یا گروه‌بندی داده‌های با خواص مشابه درون مجموعه‌ای از دسته‌ها یا خوشه‌ها است. همواره بشر هنگام یادگیری مطالب جدید یا برای فهم یک پدیده جدید سعی می‌کند، مشخصه و ویژگی‌هایی پیدا نماید تا بتواند آن را به کمک مقایسه‌ای با پدیدهای شناخته توصیف نماید. این مقایسه می‌تواند بر اساس شباهت یا تفاوت، معیارهای کلی مثل نزدیکی و یا مطابق با استانداردها و قانون‌های مشخص صورت گیرد.

به صورت کلی روش‌های داده‌کاوی به دو گروه با ناظر [۴] و بدون ناظر [۳] تقسیم‌بندی می‌شوند. در روش‌های باناظر یک متغیر هدف از قبل تعریف شده، وجود دارد. در این روش‌ها مثال‌های زیادی وجود دارند که مقدار متغیر هدف برای آنها مشخص است، بنابراین الگوریتم می‌تواند به کمک آن‌ها آموزش ببیند که کدام متغیر هدف با کدام مقادیر متغیرهای پیش‌بینی کننده متناظر است. یکی از متعارف‌ترین روش‌های یادگیری باناظر در داده‌کاوی رده‌بندی است. در رده‌بندی یک صفت مشخص به نام صفت رده وجود دارد که می‌تواند مقادیر معلومی را اخذ کند و قصد داریم که مدلی ایجاد شود تا داده‌ها و نمونه‌های جدید را در رده‌های از پیش تعیین شده قرار دهد. هدف رده‌بندی این است که ابتدا داده‌های آموزشی را تحلیل کرده و یک نسخه دقیق یا یک مدل برای هر رده با استفاده از صفات موجود در داده‌ها تعیین شود. سپس با استفاده از مدل بدست آمده، مجموعه آزمایشی رده‌بندی می‌شود تا توصیف بهتری برای هر رده ایجاد کند.

در روش‌های بدون ناظر متغیر هدفی تعریف نمی‌شود و الگوریتم داده کاوی همبستگی‌ها و ساختارهای بین تمام متغییرها را جستجو می‌کند. از مهم‌ترین روش‌های داده‌کاوی بدون ناظر، خوشه‌بندی را می‌توان نام برد.

مسأله خوشه‌بندی می‌تواند به صورت یک مسأله بهینه‌سازی فرمول‌بندی شود. نکته کلیدی این است که چطور متغیرهای تصمیم‌گیری و توابع هدف را تعریف نماییم. در خوشه‌بندی کلی‌ترین هدف، کمینه‌کردن تفاوت نمونه‌های هر خوشه (فشردگی) و بیشینه نمودن تفاوت نمونه‌های یک خوشه با خوشه‌های دیگر (تمایز) یا ترکیبی از این دو است. البته کیفیت نتایج خوشه‌بندی به روش اندازه‌گیری شباهت نیز وابستگی دارد. در بخش بعد به تشریح بعضی از معیارهای شباهت متداول می‌پردازیم. از دیگر معیارهای الگوریتم‌های خوشه‌بندی توانایی و قدرت آنها در کشف الگوهای مخفی میان داده‌هاست.

از طرف دیگر، خوشه‌بندی منجر به کاهش حجم اطلاعات نیز می‌شود؛ چون به جای نگهداری اطلاعات تعداد زیادی از اشیاء، می‌توان اطلاعات چند گروه همگن را نگهداری نمود. خوشه‌بندی در زمینه‌های بسیاری از قبیل مهندسی (یادگیری ماشین، هوش مصنوعی، تشخیص الگو، مهندسی مکانیک و الکترونیک)، علوم کامپیوتر (کاوش وب، تحلیل پایگاه داده فضایی، جمع‌آوری مستندات متنی، تقسیم‌بندی تصویر)، علوم پزشکی (ژنتیک، زیست‌شناسی، میکروب‌شناسی، فسیل‌شناسی، روان‌شناسی، بالین، آسیب شناسی)، علوم زمین شناسی (جغرافیا،

Journal of Iranian Association of Electrical and Electronics Engineers - Vol.13- No.2 Summer 2016

[DOR: 20.1001.1.26765810.1395.13.2.13.9]

زمین شناسی، نقشه برداری از زمین)، علوم اجتماعی (جامعه‌شناسی، روان‌شناسی، تاریخ، آموزش و پرورش) و اقتصاد (بازاریابی، تجارت) کاربرد دارد. خوشه‌بندی ممکن است با نام-های دیگری از قبیل علم رده‌بندی عددی، یادگیری بدون معلم (یا یادگیری بدون نظارت)، تحلیل گونه‌شناسی و افرازبندی بکار برده شود.

بخش بعد به پیشینه مورد نیاز می‌پردازد. بخش سوم بررسی اجمالی کارهای صورت گرفته در زمینه خوشه‌بندی ترکیبی مبتنی بر زیرمجموعه‌ای از نتایج اولیه می‌پردازد. روش پیشنهادی برای خوشه‌بندی ترکیبی مبتنی بر زیرمجموعه‌ای از نتایج اولیه در بخش چهارم تشریح شده است. بخش پنجم به بررسی و تفسیر نتایج آزمایش‌ها می‌پردازد. در نهایت جمع‌بندی حاصل از این پایان‌نامه در بخش ششم آمده است. همچنین، در این بخش کارهای آینده معرفی می‌شوند.

۲- پیشینه

در این بخش به مطالب پیشنیاز مورد استفاده در این مقاله پرداخته خواهد شد.

۲-۱- الگوریتم ژنتیک

الگوریتم ژنتیک[۷] رهیافتی است که تکامل طبیعی موجودات را الگو قرار می‌دهد [۸]. این روش تقلیدی از فرایند تکامل با استفاده از الگوریتم‌های کامپیوتری است. اساسی‌ترین اصل تکامل، وراثت است. هر نسل، خصوصیات نسل قبلی را به ارث می‌برد و به نسل بعد انتقال می‌دهد. این انتقال خصوصیات از نسلی به نسل بعد توسط ژن‌ها صورت می‌گیرد. جهانی که در آن زندگی می‌کنیم دائمًا در حال تغییر است. برای بقا در این سیستم پویا ، افراد باید توانایی داشته باشند که خود را با محیط، سازگار کنند.

سازگاری[۸] تعیین می‌کند که آن موجود چه مقدار زنده خواهد ماند و چقدر شانس دارد تا ژن‌های خود را به نسل بعد انتقال دهد. در تکامل بیولوژیکی، فقط برنده‌ها هستند که می-توانند در فرایند تکامل شرکت کنند. خصوصیات هر موجود زنده در ژن‌هایش، کدگذاری شده است و طی فرایند وراثت ، این ژن-ها به فرزندان[۹] منتقل می‌شوند.

مبتکر الگوریتم ژنتیک جان هلند در دهه هفتاد میلادی با الهام گرفتن از ویژگی‌های تئوری تکامل، الگوریتم جستجویی ابداع کرد که در این الگوریتم از همان اصولی که طبیعت فرایند تکامل را روی نمادهای ژنی انجام می‌دهد [۸]، برای تکامل جواب‌های مربوط به حل های یک مساله بهینه‌سازی استفاده می‌کند. فرایند با یک جمعیت اولیه تصادفی از جواب‌های ممکن شروع می‌شود.

هریک از جواب‌ها توسط یک ساختار رشته‌ای از بیت‌ها که مقدار کدگذاری شده متغیرهای تصمیم را در بردارند، نشان داده می‌شوند. سپس با تشکیل خانواده اولیه و ارزیابی هر یک از رشته‌ها، افراد مناسب برای تشکیل خانواده بعدی انتخاب می-شوند. جواب‌های جدید از خانواده جواب‌های اولیه با تغییر دادن ساختار رشته‌ها توسط عملگرهای الگوریتم ژنتیک تولید می-شوند.

رشته‌های جدید توسط روند طراحی الهام گرفته از مکانیزم ژنتیک طبیعی تولید می‌شوند. سپس مقدار برازندگی رشته‌های جدید با توجه به تابع هدف مسأله مورد نظر ارزیابی می‌شود. این روند موجب بهبود مداوم برازندگی خانواده حل‌ها شده و تا زمانی که حل‌ها همگرا شوند، تکرار می‌شود. دو جنبه مهم در الگوریتم ژنتیک وجود دارند که دائما جواب‌ها را آشفته کرده و مجال خروج از بهینه‌های موضعی را فراهم می‌آورند. یکی از این جنبه-ها آمیزش است که الگوریتم ژنتیک از آن برای تولید جواب استفاده می‌کند. جنبه دیگر که عملگر جهش نام دارد، قادر است مقادیر جدیدی به بیت‌ها بدهد که در گروه والدین وجود نداشته است. عملگر جهش کمک می‌کند که تنوع ژنتیک باقی بماند و جستجو به نواحی جدیدی برسد.

۲-۲- نورد شبیه‌سازی شده

نورد شبیه‌سازی شده یک روش بهینه‌سازی است که به جهت شباهت آن به فرایند حرارت فلزات و سرد کردن آرام آنها به این اسم نامیده می‌شود [۹]. در انتهای فرآیند بازپخت فیزیکی، جسم جامد به حالت کریستالی می‌رسد. در روش بهینه‌سازی، تابع هدف، انرژی فرآیند ترمودینامیکی را نشان می‌دهد و جواب بهینه به منزله حالت کریستالی جسم جامد می‌باشد. تفاوت این روش با فرآیند حرارت فلزات در این است که متغیر روش‌های پایه جستجوی موضعی، فرآیند جستجو، بهترین جواب را از میان نقاطی که در همسایگی جواب اولیه قرار دارند، انتخاب می‌کند. اگر بهترین جواب، بهتر از جواب اولیه نباشد فرآیند جستجو متوقف می‌شود و فرض می‌گردد که جواب بهینه پیدا شده است. این روش برای توابع هدفی مؤثر است که ساده باشند و فقط دارای یک نقطه حدی[۱۰] موضعی باشند (برای مسائل حداقل یا حداکثرسازی). برای توابع پیچیده، بطور مثال برای مسائل

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol.13 No.2 Summer 2016

حداقل‌سازی، این نقطه بهینه موضعی ممکن است با بهینه عمومی[11] کاملا متفاوت باشد و مدل بهینه‌سازی قادر به ارائه جواب‌های بهینه مورد نظر نباشد.

۳-۲- ترکیب خوشه‌بندها

از آنجایی که اکثر روش‌های خوشه‌بندی پایه روی جنبه‌های خاصی از داده‌ها تاکید می‌کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد می‌باشند. به همین دلیل، نیازمند روش‌هایی هستیم که بتواند با استفاده از ترکیب این الگوریتم‌ها و گرفتن نقاط قوت هر یک، نتایج بهینه‌تری را تولید کند. در واقع هدف اصلی خوشه‌بندی ترکیبی جستجوی نتایج بهتر و مستحکم‌تر، با استفاده از ترکیب اطلاعات و نتایج حاصل از چندین خوشه‌بندی اولیه است [۷] و [۱۰]. تاکنون مطالعات زیادی بر روی ترکیب رده‌بندها انجام شده است [۱۳-۱۱]. خوشه‌بندی ترکیبی نیز در این عرصه همپای رده‌بندی ترکیبی مورد توجه بسیار قرار گرفته است. تحقیقات اخیر در این زمینه نشان داده‌اند که خوشه‌بندی داده‌ها می‌تواند به طور چشمگیری از ترکیب چندین افراز داده سود ببرد. به علاوه، قدرت موازی‌سازی آنها یک انطباق طبیعی با نیاز داده‌کاوی توزیع‌شده دارد. خوشه‌بندی ترکیبی می‌تواند جواب‌های بهتری از نظر استحکام[12]، نو بودن[13]، پایداری[14] و انعطاف‌پذیری[15] نسبت به روش‌های پایه ارایه دهد [۷] و [۱۴-۱۶].

به طور خلاصه خوشه‌بندی ترکیبی شامل دو مرحله اصلی زیر می‌باشد [۵] و [۷]:

۱- تولید نتایج متفاوت از خوشه‌بندی‌ها، به عنوان نتایج خوشه‌بندی اولیه بر اساس اعمال روش‌های مختلف که این مرحله را، مرحله ایجاد تنوع یا پراکندگی[16] می‌نامند.

۲- ترکیب نتایج به دست آمده از خوشه‌بندی‌های متفاوت اولیه برای تولید خوشه نهایی؛ که این کار توسط تابع توافقی[17] (الگوریتم ترکیب‌کننده) انجام می‌شود.

مراحل ۱ و ۲ در زیربخش‌های ۳-۱ و ۳-۳ به ترتیب تشریح شده‌اند.

۳- کارهای مرتبط

روش‌های خوشه‌بندی ترکیبی سعی می‌کنند تا با ترکیب افرازهای[18] مختلف تولید شده از روش‌های خوشه‌بندی پایه، یک افراز مستحکم[19] از داده‌ها را تولید کنند [۷]، [۱۷] و [۱۸]. در اکثر مطالعات اخیر، همه افرازها با وزن برابر در ترکیب نهایی حاضر می‌شوند و همه خوشه‌های موجود در همه افرازها نیز با وزن برابر در ترکیب نهایی شرکت می‌کنند [۱۹]. استرل و گاش [۷] یک معیار برای انتخاب از میان ترکیبات ممکن ارایه کرده‌اند که مبتنی بر کیفیت کلی یک خوشه‌بندی است. برای این کار، آنها میزان ثبات بین افراز ترکیبی و افرازهای پایه را در نظر گرفته‌اند و با استفاده از یک قاعده ترکیبی ثابت، یک معیار شباهت دو به دو[20] را روی فضای ویژگی‌های d-بعدی به کار برده‌اند.

عظیمی [۱] از مفهوم پراکندگی[21] برای هوشمند نمودن خوشه‌بندی ترکیبی استفاده می‌کند. این روش که به صورت پویا اقدام به انتخاب زیرمجموعه بهینه‌ای از نتایج اولیه در ترکیب نهایی می‌کند، ابتدا یک خوشه‌بندی ترکیبی ساده انجام می‌شود. سپس این روش میزان شباهت تمام نتایج خوشه‌بندی‌های اولیه را نسبت به جواب به دست آمده ارزیابی می‌کند و سعی در طبقه‌بندی[22] مجموعه داده‌ها به سه مجموعه داده راحت[23]، معمولی[24] و سخت[25] می‌کند. در این طبقه‌بندی، مجموعه داده راحت به مجموعه داده‌ای اطلاق می‌شود که خوشه‌بندی‌های اولیه تفاوت چندانی با خوشه‌بندی ترکیبی به دست آمده نداشته باشند. به این معنی که هر خوشه‌بندی ساده بتواند تقریبا مانند خوشه‌بندی ترکیبی نتایج مشابه‌ای ارایه کند. مجموعه داده معمولی به مجموعه داده ای اطلاق می‌شود که خوشه‌بندی‌های اولیه نه تفاوت چندانی و نه تشابه چندانی با نتایج خوشه‌بندی ترکیبی به دست آمده دارند. مجموعه داده سخت به مجموعه داده‌ای اطلاق می‌شود که خوشه‌بندی‌های اولیه تشابه چندانی با خوشه‌بندی ترکیبی به دست آمده نداشته باشند. این رویداد نشان می‌دهد که داده‌های مجموعه مورد نظر کاملا دارای مرزهای مشترک هستند و روش‌های ساده و معمولی خوشه‌بندی همانند روش‌های پیچیده و قدرتمند خوشه‌بندی ترکیبی قادر به جداسازی نمونه‌ها نمی‌باشند. سپس کل نتایج خوشه‌بندی‌های اولیه به چهار زیرمجموعه متفاوت بر اساس میزان تطبیق دقت‌شان با نتایج خوشه‌بندی ترکیبی ساده تقسیم می‌شوند و بر اساس رده[26] هر مجموعه داده (راحت، معمولی و سخت) اقدام به انتخاب یکی از این زیرمجموعه‌ها برای ترکیب و به دست آوردن نتیجه نهایی می‌کنیم. نتایج تجربی[27] صورت گرفته در [۱] نیز نشان داده‌اند 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

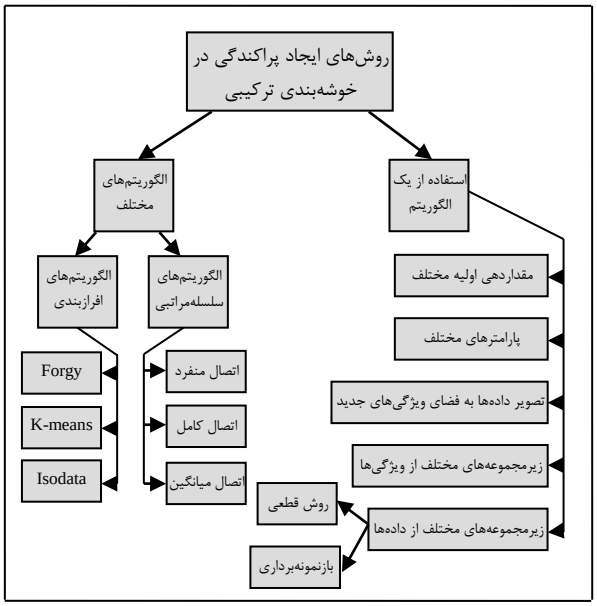
مـی‌شـود را از ترکیـب نهـایی خـارج کنـد و بـه ایـن ترتیـب خوشه‌بندی‌های ترکیبی اولیه‌ای را که دارای دقت نسبتا مناسبی هستند، وارد ترکیب نهایی کند. روش دیگری که بسـیار بـه ایـن روش نزدیک است، روش علیزاده [۲] می‌باشد که در آنجـا بـرای همه‌ی مجموعه داده‌ها ابتدا خوشه‌ها بـر اسـاس پایـداری مرتـب شده و سپس ۳۳ درصد پایدارتر انتخاب می‌شوند.

در اکثر الگـوریتم‌هـای پایـه بـرای خوشـه‌بنـدی ترکیبـی از نمونه‌برداری داده‌ها استفاده می‌شود. مسئله اصلی در این روش‌ها چگونگی ارزیابی خوشه و خوشه‌بندی (افـراز) اسـت. بامگـارتنر و همکاران [۲۰] یک روش مبتنی بر بازنمونه‌برداری را برای بررسی اعتبارسنجی نتایج خوشه‌بندی فازی ارایه کرده‌اند. در چند سـال اخیر، پایداری خوشه به عنوان یک معیـار ارزیـابی خوشـه مـورد توجه زیادی قرار گرفته است [۱۹] و [۲۳-۲۱]. ایـده‌هـای اولیـه برای اعتبارسنجی خوشه با اسـتفاده از بازنمونـه‌بـرداری در [۲۴] ارایه شده و بعدها در [۲۵] و [۲۶] کامـل‌تـر شـده اسـت. راس و همکاران [۲۷] و [۲۸]، نیز یک روش مبتنی بر بازنمونـه‌بـرداری برای اعتبارسنجی خوشه ارایه کرده‌اند. عنصر اصلی در این روش، که در واقع کامل‌شـده‌ی روش‌هـای پیشـین مـی‌باشـد، پایـداری خوشه است. معیار پایداری، میزان همبستگی افرازهای به دسـت آمده از دو نمونه‌برداری مستقل از مجموعـه داده را انـدازه‌گیـری می‌کند. هر چه میزان پایداری برای یک خوشه‌بندی بیشتر باشد، به این معنی است که اگر الگوریتم خوشـه‌بنـدی چنـدین مرتبـه دیگر روی آن نمونه‌ها به کار رود، نتایج مشابهی حاصل مـی‌شـود [۲۹] و [۳۰]. همچنین، راس و لانژ [۳۱] یـک الگـوریتم جدیـد برای خوشه‌بندی ارایه کرده‌انـد کـه مبتنـی بـر انتخـاب ویژگـی می‌باشد. در این روش از معیار پایداری مبتنی بر بازنمونه‌بـرداری داده‌ها، برای انتخاب پارامترهای الگوریتم خوشـه‌بنـدی اسـتفاده شده است. چندین روش اعتبارسنجی خوشـه[۲۸] مبتنـی بـر ایـده استفاده از پایداری پیشنهاد شده است [۳۲]. بن هور و همکـاران [۳۳] نیز روشی برای محاسبه پایداری ارایه کرده‌اند که بر مبنای شباهت بین نمونه‌ها در خوشه‌بندی‌های مختلف عمل می‌کند. در ایـن روش، ابتـدا مـاتریس همبسـتگی بـا اسـتفاده از روش بازنمونه‌برداری به دست می‌آید. سپس ضریب جاکارد[۲۹] به عنوان معیار پایداری بر اساس این ماتریس محاسبه می‌شود. همچنـین، کاسـترو و یانـگ [۳۴] روشـی بـرای ارزیـابی افرازهـای نهـایی خوشه‌بندی ارایه کرده‌اند که از ماشین بردار پشتیبان[۳۰] اسـتفاده می‌کند. این روش با شناسایی نویزها و داده‌هـای دورافتـاده[۳۱] بـه نتایج خوشه‌بندی دارای استحکام دست یافته است. مولر و رادک [۳۵] از بازنمونه‌برداری نزدیک‌ترین همسایه[۳۲] برای اعتبارسنجی نتایج خوشه‌بندی استفاده کـرده‌انـد. ایـن روش بازنمونـه‌بـرداری اولین بار در تحلیل سری‌های زمـانی بـه کـار رفتـه اسـت [۳۶]. اینوکوچی و همکاران [۳۷] معیار اعتبارسـنجی هسـته-محـور[۳۳] پیشنهاد کرده‌اند. هسـته در ایـن روش بـه معنـی تـابع مرکـزی استفاده شده در ماشین بردار پشتیبان می‌باشد. در این روش، دو شاخص مـورد توجـه قـرار گرفتـه اسـت: اولـی مجمـوع مقـادیر کواریانس فازی داخل خوشه‌هاست و دومـی شـاخص مبتنـی بـر هسته ژی-بن مـی‌باشـد [۳۸]. در ایـن روش از ایـن دو شـاخص برای ارزیابی نتایج خوشه‌بندی و همچنین، تعیین تعداد خوشه‌ها با مرزهای غیـر خطـی اسـتفاده شـده اسـت. داس و سـیل [۳۹] روشـی بـرای تعیـین تعـداد خوشـه‌هـا ارایـه کـرده‌انـد کـه از اعتبارسنجی خوشه‌ها برای تقسیم و ادغام آنها بهره می‌برد. فـرن و لین [۴۰] روشی برای خوشه‌بندی ترکیبی پیشنهاد کرده‌اند که از زیرمجموعه‌ی مـوثرتری از افرازهـای اولیـه در ترکیـب نهـایی استفاده می‌کند. در این روش اگر چه تعداد اعضای شرکت کننده در ترکیب نهایی کمتر از یک خوشه‌بندی ترکیبی کامل[۳۴] اسـت، به دلیل انتخاب افرازهای بـا کـارایی بـالاتر، نتـایج نهـایی بهبـود می‌یابند. پارامترهایی که در این روش مورد توجه قرار گرفته‌انـد، عبارتنـد از: کیفیـت و پراکنـدگی. ایـن روش سـعی مـی‌کنـد تـا زیرمجموعه‌ای از افرازهایی از نتـایج اولیـه را وارد ترکیـب نهـایی کند که از بالاترین میزان کیفیت برخوردار بوده و در عـین حـال نسبت به هم بیشترین پراکنـدگی را دارا باشـند. در ایـن روش از معیار مجموع اطلاعات متقابل نرمال‌شده[۳۵] (بـرای یـک افـراز در مقایسه با افرازهای دیگر ترکیب) برای اندازه‌گیـری کیفیـت یـک افـراز اسـتفاده شـده اسـت. همچنـین، معیـار اطلاعـات متقابـل نرمـال‌شـده[۳۶] (بـین تمـام افرازهـای موجـود در ترکیـب) بـرای اندازه‌گیری پراکندگی لازم برای ترکیب به کار رفته است. فـرن و لین [۴۰] نشان می‌دهند که روش آنها نسـبت بـه خوشـه‌بنـدی ترکیبـی کامـل و یـا روش انتخـاب تصـادفی از کـارایی بهتـری برخـوردار اسـت. لاو و بقیـه [۲۱] یـک روش خوشـه‌بنـدی چندهدفی[۳۷] ارایه کرده‌اند که مبتنی بر انتخاب خوشه‌های اولیـه تولید شده توسط الگوریتم‌های مختلف خوشه‌بندی، در طی یـک روال بهینه‌سازی می‌باشد. در این روش، بهترین مجموعه از توابع هدف برای بخش‌های مختلف از فضای ویژگی انتخاب شده است. فرد و جین [۱۹] یک روش خوشه‌بندی ترکیبی ارایه کرده‌اند که در آن بـا اسـتفاده از معیـار پایـداری خوشـه، شـباهت دو بـه دو آموزش داده می‌شود. در این روش، به جای استفاده از معیارهـای ارزیابی مبتنی بر افراز نهایی، افرازهـای حاصـل از الگـوریتم‌هـای

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol.13 No.2 Summer 2016

پایه در نواحی مختلف از فضای ویژگی d-بعدی مورد ارزیابی قرار گرفته است.

در اکثر روش‌های فوق از ماتریس همبستگی برای تجمیع اطلاعات خوشه‌بندی‌های اولیه استفاده شده است. اگر چند ماتریس وجود داشته باشد، روش‌های مختلفی وجود دارند که ماتریس همبستگی نهایی را از ادغام آنها تولید می‌کنند. بروزیر [۴۱] یک روش برای ادغام چند ماتریس فاصله فرامتری به یک ماتریس فاصله فرامتری ارایه نمود. شرط لازم و کافی برای یکتایی عملیات ادغام نیز مورد بررسی قرار گرفته است. همچنین، لاپوینته و لجندره [۴۲] دو روش برای تولید دندوگرام‌های تصادفی ارایه کرده‌اند. آنها این کار را با تصادفی کردن ماتریس کفنتیک وابسته[۳۸] انجام داده‌اند. هر وارده $d_C(i,j)$ در ماتریس کفنتیک بیانگر سطحی است که در آن نمونه‌های x_i و x_j برای اولین بار همدیگر را در یک خوشه ملاقات کرده‌اند [۴۳]. بنفیلد [۴۴] الگوریتمی برای محاسبه فاصله‌های فرامتری برای یک دندوگرام اتصال منفرد ارایه می‌کند. این الگوریتم یک ماتریس فاصله از ورودی می‌گیرد و از درخت پوشای کمینه‌ی[۳۹] حاصل از آن استفاده می‌کند. فاصله‌های فرامتری در این الگوریتم، به حداکثر اتصالات زنجیری گفته می‌شود که هر کدام از جفت نقاط درخت را به هم متصل می‌کند. در واقع فاصله‌های فرامتری سطوحی از دندوگرام هستند که جفت نقاط در آن سطح به هم متصل شده‌اند.



شکل (۱): طبقه‌بندی روش‌های ایجاد پراکندگی در خوشه‌بندی ترکیبی

۱-۳- پراکندگی در خوشه‌بندها

معمولا در مرحله اول از خوشه‌بندی ترکیبی تعدادی خوشه‌بندی‌های اولیه که هر کدام بر ویژگی خاصی از داده‌ها تاکید دارند، ایجاد می‌شود. اولین و ساده‌ترین روش برای ایجاد نتایج مختلف و پراکنده از یک مجموعه داده، استفاده از الگوریتم‌های مختلف خوشه‌بندی است. هر الگوریتم خوشه‌بندی از یک جنبه خاصی به مسئله نگاه می‌کند. بنابراین خطاهای موجود در روش‌های مختلف، می‌تواند با هم متفاوت باشد. این امر می‌تواند موجب ایجاد پراکندگی در نتایج الگوریتم‌های پایه خوشه‌بندی گردد. مهم‌ترین الگوریتم‌های خوشه‌بندی پایه که معمولا در خوشه‌بندی ترکیبی استفاده می‌شوند، شامل الگوریتم‌های خوشه‌بندی سلسله‌مراتبی[۴۰] [۵]، [۱۴] و [۴۵]، و الگوریتم‌های خوشه‌بندی افرازبندی[۴۱] [۴۳] و [۴۷-۴۵] می‌باشند. الگوریتم K-means، به دلیل سادگی و توانایی مناسب در خوشه‌بندی، معمولا در بیشتر روش‌های خوشه‌بندی ترکیبی به عنوان الگوریتم خوشه‌بندی پایه، استفاده می‌شود [۱۳]، [۴۳] و [۵۰-۴۸].

رویکرد دیگر برای ایجاد پراکندگی، به دست آوردن نتایج متنوع از یک الگوریتم خوشه‌بندی پایه با استفاده از یکی از روش‌های زیر می‌باشد.

- تغییر مقادیر اولیه[۴۲] الگوریتم خوشه‌بندی انتخاب شده [۱۰]
- تغییر پارامترهای الگوریتم خوشه‌بندی انتخاب شده [۵۱]
- استفاده از زیرمجموعه‌های مختلف از ویژگی‌ها[۴۳] [۷] و [۱۵]
- نگاشت داده‌ها به فضاهای ویژگی دیگر [۱۵] و [۵۲]
- تقسیم‌بندی داده‌های اصلی به زیر مجموعه‌هایی متفاوت و مجزا (بازنمونه‌برداری[۴۴]) [۷]، [۴۸] و [۵۳]
- طبقه‌بندی مهم‌ترین روش‌های ایجاد پراکندگی در نتایج اولیه در شکل ۱ ارایه شده است..

۲-۳- مشکلات پیش روی ترکیب خوشه‌بندها

ترکیب خوشه‌بندی‌ها کار مشکل‌تری از ترکیب رده‌بندی‌های باناظر است. به عبارت بهتر، برخلاف مسئله رده‌بندی که دارای ناظر[۴۵] و یک مجموعه یادگیری[۴۶] می‌باشد، در خوشه‌بندی هیچ‌گونه شناختی نسبت به مجموعه داده وجود ندارد. عدم وجود ناظر و مجموعه یادگیری، ارایه روش‌های مدرن و هوشمند

خوشه‌بندی داده‌ها که دارای کارایی بالا باشند را بسیار مشکل نموده است. همچنین، در غیاب داده آموزشی برچسب‌دار، ما با مشکل تناظر بین برچسب‌های خوشه در افرازهای مختلف از یک ترکیب مواجه هستیم.

بسیاری از مطالعات در سال‌های اخیر در این زمینه استوار بوده است که خوشه‌بندی‌های اولیه متنوع‌تری را برای ایجاد نتایج اولیه به کارگیرند [۱۴]، [۱۵]، [۵۴] و [۵۵]. مفهوم پراکندگی به طور گسترده‌ای در تحقیقات سال‌های اخیر مورد استفاده قرار گرفته است [۷]، [۱۵]، [۱۷] و [۱۹]. هدف اصلی در اکثر روش‌های اخیر خوشه‌بندی ترکیبی، تنها بررسی مجموعه داده از زوایای مختلف است و این سوال که "آیا پراکندگی بوجود آمده مفید می‌باشد یا نه؟" چندان مورد توجه قرار نگرفته است. در حقیقت به خاطر ماهیت بدون ناظر بودن مسئله خوشه‌بندی مطالعه این امر با دشواری‌های زیادی روبروست. اگرچه نتایج تجربی نشان داده‌اند که ایجاد پراکندگی در خوشه‌بندی‌های اولیه معمولا موجب بهبود خوشه‌بندی در اکثر مواقع می‌شود [۵۶]، عظیمی [۱] نشان داده است که در بعضی مجموعه داده‌ها، پراکندگی بیشتر لزوما کمکی به افزایش دقت در نتایج نهایی نمی‌کند.

عامل دیگری که معمولا برای بهبود عملکرد خوشه‌بندی ترکیبی از آن استفاده شده است، کیفیت نتایج اولیه می‌باشد. نشان داده شده است که هر چه نتایج اولیه علاوه بر داشتن پراکندگی لازم، از کیفیت بالاتری برخوردار باشند، کیفیت خوشه‌های نهایی نیز بهتر خواهد بود [۱۸]. اگرچه فرن و لین [۴۰] نشان داده‌اند که بهینه‌سازی همزمان دو عامل پراکندگی و کیفیت در نتایج اولیه خوشه‌بندی ترکیبی می‌تواند کارایی خوشه‌بندی ترکیبی را به طور چشمگیری بهبود بخشد، تنظیم و مصالحه بین این دو عامل مسئله‌ای است که حل دقیق آن هنوز با دشواری‌های فراوانی روبروست.

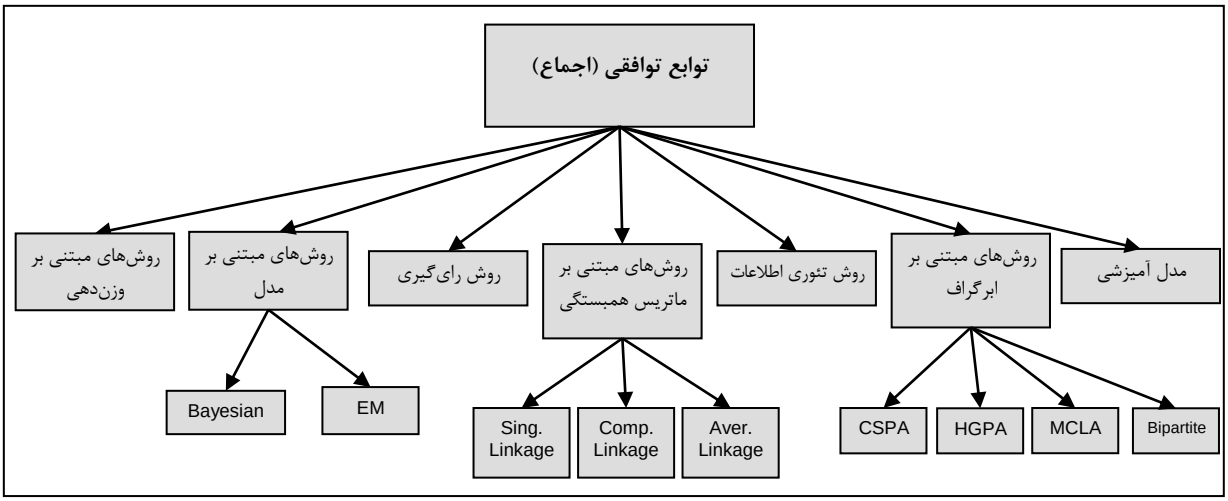
در این تحقیق سعی شده است تا زیرمجموعه‌ی نسبتا بهینه‌ای از نتایج خوشه‌بندی‌های اولیه برای شرکت در ترکیب نهایی یافت شود که بتواند کارایی خوشه‌بندی ترکیبی را بهبود بخشد. برای این منظور یک چارچوب کلی پیشنهاد شده است.

۳-۳- تابع جمع‌کننده

پس از اینکه نتایج اولیه (تا حد ممکن پراکنده) تولید شد، معمولا با استفاده از یک تابع ترکیب‌کننده این نتایج ترکیب می‌شوند. یکی از متداول‌ترین روش‌های ترکیب نتایج استفاده از ماتریس همبستگی است که در بخش۳-۵ به طور مفصل تشریح خواهد شد. روش خوشه‌بندی ترکیبی انباشت مدارک۴۷ که مبتنی بر ماتریس همبستگی است، اولین بار توسط فرد و جین [۱۰] مطرح شد و خیلی زود به صورت یک روش متداول درآمد. امروزه روش‌های دیگری نیز مبتنی بر ماتریس همبستگی ارایه شده‌اند [۱۶]. شکل ۲ یک طبقه‌بندی کلی از توابع توافقی گوناگون را نشان می‌دهد.

بسیاری از مطالعات در سال‌های اخیر در این زمینه استوار بوده است که خوشه‌بندی‌های اولیه متنوع‌تری را برای ایجاد نتایج اولیه به کارگیرند [۱۴]، [۱۵]، [۵۴] و [۵۵]. مفهوم پراکندگی به طور گسترده‌ای در تحقیقات سال‌های اخیر مورد استفاده قرار گرفته است [۷]، [۱۵]، [۱۷] و [۱۹]. هدف اصلی در اکثر روش‌های اخیر خوشه‌بندی ترکیبی، تنها بررسی مجموعه داده از زوایای مختلف است و این سوال که "آیا پراکندگی بوجود آمده مفید می‌باشد یا نه؟" چندان مورد توجه قرار نگرفته است. در حقیقت به خاطر ماهیت بدون ناظر بودن مسئله خوشه‌بندی مطالعه این امر با دشواری‌های زیادی روبروست. اگرچه نتایج تجربی نشان داده‌اند که ایجاد پراکندگی در خوشه‌بندی‌های اولیه معمولا موجب بهبود خوشه‌بندی در اکثر مواقع می‌شود [۵۶]، عظیمی [۱] نشان داده است که در بعضی مجموعه داده‌ها، پراکندگی بیشتر لزوما کمکی به افزایش دقت در نتایج نهایی نمی‌کند.

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol.13 No.2 Summer 2016



شکل (۲): طبقه‌بندی توابع توافقی در خوشه‌بندی ترکیبی

روش‌های مبتنی بر ابرگراف

استرل و گاش [۷] مفهوم مجمع را با دیدگاه ابرگراف[۴۸] بیان کرده‌اند. آنها سه روش مبتنی بر ابرگراف برای روش‌های اجماع پیشنهاد داده‌اند. آنها نواحی داده را به صورت یک ابرگراف در اولین مرحله از هر سه روش اجماع تبدیل کرده‌اند. در آن ابرگراف‌ها، خوشه‌ها به عنوان ابریال‌ها[۴۹] تلقی می‌شوند، در صورتی که راس‌ها، داده‌ها هستند. الگوریتم‌های حداقل برش[۵۰] ابرگراف می‌توانند جهت جداسازی راس‌ها یا نقاط داده‌ای استفاده شوند. با استفاده از الگوریتم برش حداقل k بر روی ابرگراف، از این ابرگراف، k ناحیه یا افراز استخراج می‌شود. مکاشفه‌های[۵۱] کارآمد جهت حل مساله برش حداقل k بر روی یک ابرگراف وجود دارد که بعضی از آنها دارای پیچیدگی محاسباتی از درجه $o(|\varepsilon|)$ که ε تعداد ابریال‌ها می‌باشد. سه الگوریتم ابرگراف CSPA[۵۲]، HGPA[۵۳] و MCLA[۵۴] در ادامه بیشتر توضیح داده شده‌اند.

CSPA

در CSPA فضای ویژگی نقاط داده‌ای در ابتدا به فضای ویژگی همبستگی ابرگراف نگاشت می‌شود. سپس یک الگوریتم حداقل برش ابرگراف شبیه متیس[۵۵] بر نقاط داده‌ای جدیدا فاصله‌دار شده به کار برده می‌شود. همانند قبل این روش فرض می‌کند که هر چه نقاط داده‌ای بیشتری در یک خوشه در افراز اولیه باشد، احتمال بیشتری دارد که آن نقاط داده‌ای ذاتا متعلق به یک خوشه باشند. CSPA ساده‌ترین مکاشفه در بین روش‌های مبتنی بر ابرگراف می‌باشد. پیچیدگی محاسباتی آن $O(kN^2M)$ است که k تعداد خوشه‌ها، N تعداد نقاط داده‌ای و M تعداد نواحی می‌باشد. دو روش بعدی پیچیدگی محاسباتی کمتری دارند.

HGPA

الگوریتم HGPA فرض می‌کند که راس‌ها نقاط داده‌ای و خوشه‌هایی که از افراز اولیه بیرون آمده‌اند، ابریال‌های آن هستند. حال دوباره یک الگوریتم ابرگراف حداقل برش شبیه متیس بر روی آن ابرگراف جهت جداسازی راس‌ها یعنی نقاط داده‌ای ابرگراف به k مولفه متفاوت به کار برده می‌شود. پیچیدگی محاسباتی آن $O(kNM)$ است که دوباره k تعداد خوشه‌ها، N تعداد نقاط داده‌ای و M تعداد نواحی می‌باشد.

MCLA

الگوریتم MCLA ابتدا خوشه‌ی بدست آمده از افراز اولیه را افراز می‌کند و سپس از یک مکانیسم مبتنی بر رای‌گیری جهت تولید افرازهای مجمع استفاده می‌کند. خوشه بندی خوشه با استفاده از متیس انجام شده است. پیچیدگی محاسباتی آن $O(k^2NM^2)$ است که k، N و M مانند قبل هستند. جهت جزئیات بیشتر در مورد روش‌های مبتنی بر ابرگراف به [۷] مراجعه نمایید.

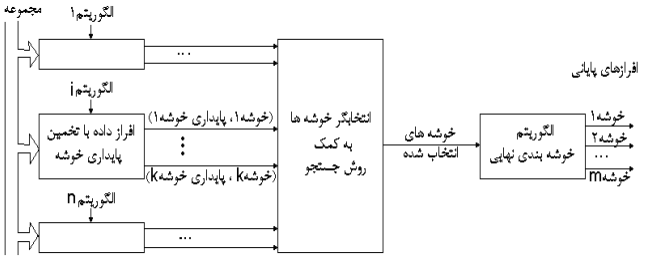
۴- چارچوب روش پیشنهادی

ایده اصلی در این روش استفاده از زیرمجموعه‌ای از خوشه‌های اولیه به جای کل خوشه‌ها در خوشه‌بندی ترکیبی است. نمای کلی از چارچوب کلی برای روال پیشنهادی در شکل ۳ نشان داده شده است.

در این روش ابتدا با استفاده از روش‌های ایجاد پراکندگی تعداد B خوشه‌بندی اولیه ایجاد می‌شود. این کار می‌تواند با

[DOR: 20.1001.1.26765810.1395.13.2.13.9]

استفاده از نمونه‌برداری از داده‌ها، استفاده از الگوریتم‌های مختلف خوشه‌بندی، استفاده از زیرمجموعه‌ای از ویژگی‌ها و یا انتخاب پارامترهای مختلف برای یک الگوریتم خوشه‌بندی انجام شود. در اینجا از الگوریتم K-means برای تولید نتایج اولیه استفاده شده است. پراکندگی لازم در نتایج اولیه برای الگوریتم K-means نیز، با انتخاب تصادفی نقاط اولیه مراکز خوشه‌ها و همچنین با نمونه‌برداری فراهم شده است. به علاوه یک روش عمده برای تولید پراکندگی لازم را از روش K-Means با K-های گوناگون بدست آورده شده است. در مرحله بعد اطلاعات متقابل نرمال شده هر کدام از خوشه‌های به دست آمده را با همه‌ی خوشه‌های به دست آمده محاسبه کرده و در یک ماتریس مربعی قرار می‌دهیم.



شکل (۳): روال الگوریتم پیشنهادی برای خوشه‌بندی ترکیبی

پس از اینکه پایداری هر خوشه نسبت به سایر خوشه‌ها محاسبه شد، در گام بعد، عمل انتخاب خوشه‌ها در دو زیربخش انجام می‌شود. در زیربخش اول با کمک الگوریتم ژنتیک به گونه‌ای یک زیرمجموعه از خوشه‌ها انتخاب می‌شود که خوشه‌های انتخاب شده با همدیگر بیشترین پایداری را داشته باشند. در زیربخش دوم برای تولید تنوع در خوشه‌های انتخاب شده به کمک الگوریتم ژنتیک خوشه‌هایی را انتخاب می‌کنیم که با هم کم‌ترین پایداری را داشته باشند. روش پیشنهاد شده برای انتخاب خوشه‌ها در بخش ۴-۲ تشریح خواهد شد.

در گام بعدی خوشه‌های انتخاب شده با هم ترکیب شده و خوشه‌های نهایی از آنها به دست می‌آید. روش‌های مختلفی برای ترکیب خوشه‌بندی‌های اولیه و به دست آوردن خوشه‌های نهایی وجود دارد. تفاوتی که در اینجا وجود دارد این است که در این روش ممکن است که از هر خوشه‌بندی اولیه، تنها تعدادی از خوشه‌ها در مجمع نهایی موجود باشند. در این مقاله دو روش پیشنهاد شده توسط علیزاده [۲] برای ترکیب نتایج خوشه‌های پایه و چگونگی بدست آوردن خوشه‌های نهایی از خوشه‌های پایه مورد استفاده قرار می‌گیرد. این روش‌ها که توانایی ساخت ماتریس همبستگی برای نمونه‌ها -در شرایطی که تنها تعدادی از خوشه‌ها موجود هستند- را دارند، در بخش ۴-۳ تشریح خواهد شد. پس از ساخت ماتریس همبستگی، می‌توان با استفاده از یکی از الگوریتم‌های سلسله‌مراتبی و یا روش‌های مبتنی بر ابرگراف خوشه‌های نهایی را به دست آورد.

از آنجایی که خوشه‌بندی ترکیبی فرایندی پیچیده شامل چندین مرحله می‌باشد، پیچیدگی محاسباتی بالاتی داشته و در مجموع فرایندی زمان‌بر و برون‌خط[۵۶] است. از همین رو، در تحقیقاتی که در این زمینه صورت می‌گیرد معمولا روی پیچیدگی محاسباتی روش‌های پیشنهادی بحث خاصی صورت نمی‌گیرد [۱۵] و [۴۱].

۱-۴- ارزیابی خوشه

در این مرحله خوشه‌های به دست آمده مورد ارزیابی قرار می‌گیرند تا کیفیت هر خوشه مشخص شود. از آن جایی که میزان برازندگی[۵۷] یک خوشه در میان کل نقاط داده معنی‌دار است، تابع برازندگی خوشه یعنی $g_j(C_i,D)$ علاوه بر پارامتر اول خود یعنی خوشه C_i به مجموعه داده D نیز وابسته است. یک تابع برازندگی باید خصوصیات زیر را داشته باشد [۲۲].

- باید با تابع f_j که توسط الگوریتم خوشه‌بندی خاص A_j بهینه می‌شود، ارتباط منطقی داشته باشد. به عبارت دیگر، مقدار بیش‌تر برای $g_j(C_i,D)$ به این معنی باشد که خوشه C_i نسبت به تابع f_j و متناظرا نسبت به الگوریتم خوشه‌بندی خاص A_j، برازنده‌تر (بهینه‌تر) است.
- باید نسبت به توابع خوشه‌بندی مختلف قابل مقایسه باشد. به عبارت دیگر، اگر $g_j(C_i,D) > g_l(C_i,D)$، آنگاه باید نتیجه گرفت که کیفیت خوشه C_i با توجه به تابع f_j از تابع f_l بهتر است. یعنی کیفیت خوشه C_i در خوشه‌بندی j-ام از کیفیت خوشه C_i در خوشه‌بندی l-ام بهتر می‌باشد.
- در نهایت مقدار تابع برازندگی خوشه باید نسبت به خوشه‌های مختلف قابل مقایسه باشد. به عبارت دیگر، $g_j(C_i,D) = g_j(C_l,D)$، باید نتیجه بدهد که خوشه‌های C_i و C_l نسبت به تابع f_j میزان برازندگی برابری دارند. یعنی این دو خوشه در خوشه‌بندی j-ام از کیفیت برابری برخوردارند.

یکی از معیارهایی که می‌تواند به عنوان تابع برازندگی خوشه در نظر گرفته شود، معیار پایداری خوشه [۲۴] است. پایداری

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol13 No.2 Summer 2016

خوشه، اثر آشفتگی[۵۸] در نتایج خوشه‌بندی‌های مختلف را انعکاس می‌دهد. یکی از مهم‌ترین روش‌های ایجاد آشفتگی در خوشه‌بندی استفاده از بازنمونه‌برداری است که می‌تواند به دو شکل رایج باجایگذاری و یا بدون‌جایگذاری انجام شود.

For l:=1 to M do
  Resample D to obtain the perturbed data set D';
  Run K-means over D' to obtain $P(D')$;
  Re-labeling $P(D')$ to $P(D)$;
  Compute $score[l] = sim(C_i,P(D))$;
End
$g_i(C_i,D)$:= *average* of $score[l]$;

شکل (۴): الگوریتم محاسبه پایداری خوشه C_i به عنوان تابع برازندگی در روش علیزاده

برای ارزیابی خوشه از معیارهای پایداری مختلفی استفاده شده است. اگر چه همه روش‌های پیشنهادی برای ارزیابی خوشه مبتنی بر اطلاعات متقابل نرمال شده هستند، هر کدام از این روش‌ها یکی از معایب آن را جبران می‌کند. یک روش ارزیابی خوشه به طور کامل در این بخش شرح داده شده است.

در این روش خوشه‌ها بر اساس معیار پایداری مبتنی بر اطلاعات متقابل نرمال شده (NMI) ارزیابی می‌شوند.

برای شناسایی خوشه‌های پایدارتر نیاز به مکانیزمی است تا بتواند پایداری را برای هر خوشه از یک خوشه‌بندی، مستقل از خوشه‌های دیگر به دست آمده از آن خوشه‌بندی، حساب کند. برای این کار، فرض کنید که می‌خواهیم پایداری خوشه C_i را محاسبه کنیم. در این روش ابتدا با نمونه‌برداری مجموعه داده‌های جدیدی درست می‌شود و خوشه‌بندی‌های مختلفی روی آن صورت می‌گیرد. سپس، سعی می‌شود تا به این سوال که "آیا این خوشه، در این خوشه‌بندی‌ها هم ظاهر شده است یا نه؟" پاسخ داده شود. برای این کار یک معیار شباهت بین آن خوشه (C_i) و خوشه‌بندی اولیه ($P(D)$) پیشنهاد می‌شود که با $sim(C_i,P(D))$ نشان داده می‌شود. با استفاده از این معیار، شباهت آن خوشه را با خوشه‌بندی‌های مختلف حاصل از نمونه‌برداری محاسبه می‌شود. سپس میانگین این معیارهای شباهت، به عنوان میزان پایداری این خوشه $g_i(C_i,D)$ برگردانده می‌شود. در واقع $sim(C_i,P(D))$ میزان اعتبار خوشه C_i را درخوشه‌بندی P روی مجموعه داده D مشخص می‌کند. شبه کد مربوط به این روال در شکل ۴ نشان داده شده است [۲].

برای محاسبه $sim(C_i,P(D))$ که میزان شباهت بین خوشه C_i و نتیجه خوشه‌بندی $P(D)$ است، به صورت زیر عمل می‌شود. ابتدا تمام نمونه‌های دیگر متعلق به مجموعه داده D که در خوشه C_i قرار ندارند، به صورت یک خوشه مستقل D/C_i نمایش داده می‌شود. حال یک خوشه‌بندی شامل دو خوشه C_i و D/C_i ایجاد شده است که آن را P_1 می‌نامیم $P_1=\{C_i, D/C_i\}$. اکنون خوشه‌بندی $P(D)$ که روی داده‌های نمونه‌برداری شده اعمال شده است، نیز باید به صورت دو خوشه‌ای ارایه شود تا در نهایت نتایج حاصل از این دو خوشه‌بندی طی فرآیندی با هم منطبق شوند. برای این منظور همه خوشه‌ها در $P(D)$ به دو خوشه C^* و D/C^* تقسیم می‌شوند. خوشه C^* از اجتماع همه خوشه‌هایی که بیش از ۵۰٪ از نمونه‌هایشان در خوشه C_i وجود دارند، تشکیل می‌شود و مابقی خوشه‌ها نیز در خوشه D/C^* قرار می‌گیرند. این خوشه‌بندی را P_2 می‌نامیم $P_2=\{C^*, D/C^*\}$. حال از اطلاعات متقابل[۵۹] نرمال شده (NMI) که معیار متداول برای ارزیابی شباهت بین دو افراز (نتیجه خوشه‌بندی) می‌باشد [۷]، [۱۰] و [۵۱]، برای اندازه‌گیری شباهت بین دو خوشه‌بندی P_1 و P_2 استفاده می‌شود. از آن جایی که معیار اطلاعات متقابل نرمال نشده (MI)، وابسته به اندازه خوشه‌هاست، معمولا از معیار NMI استفاده می‌شود. رابطه NMI بین دو خوشه‌بندی P_1 و P_2 به صورت زیر محاسبه می‌شود.

$$NMI(P_1,P_2)=\frac{MI(P_1,P_2)}{\frac{-1}{2m}\left(\sum_{i=0}^{1} p_{i.}\log\frac{p_{i.}}{m}+\sum_{j=0}^{1} p_{.j}\log\frac{p_{.j}}{m}\right)} \qquad (۱)$$

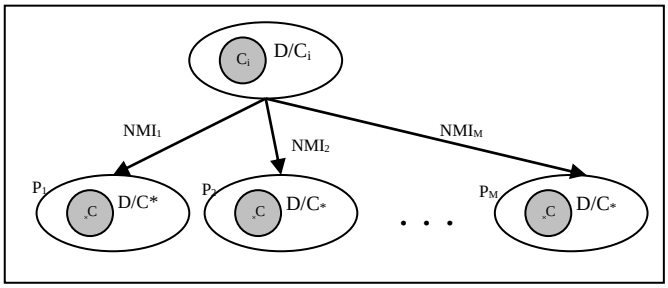
که اطلاعات متقابل، $MI(P_1,P_2)$ از رابطه ۲ به دست می‌آید.

$$MI(P_1,P_2)=\sum_{i=0}^{1}\sum_{j=0}^{1}\frac{r_{ij}}{m^2}\log\frac{mp_{ij}}{r_{ij}}$$
$$r_{ij}=p_{i.}p_{.j}\,,\quad p_{i.}=p_{i0}+p_{i1}\quad,\quad p_{.j}=p_{0j}+p_{1j} \qquad (۲)$$

که در این رابطه، p_{11} نشان‌دهنده تعداد نمونه‌های مشترک موجود در C^* و C_i است. p_{10} نشان‌دهنده تعداد نمونه‌های مشترک موجود در D/C^* و C_i است. p_{01} نشان‌دهنده تعداد نمونه‌های مشترک موجود در C^* و D/C_i است. p_{00} نشان‌دهنده تعداد نمونه‌های مشترک موجود در D/C^* و D/C_i است. همچنین m تعداد کل نمونه‌هاست. در واقع $p_{i.}$ و $p_{.j}$ به ترتیب بیانگر کل نمونه‌های موجود در C_i و C^* هستند.

شکل ۵ نمای کلی از این روش محاسبه پایداری خوشه را نشان می‌دهد. با توجه به شکل ۵، NMI_i نشان‌دهنده میزان شباهت خوشه‌بندی P_1 و P_2 می‌باشد. همچنین بیانگر میزان پایداری خوشه C_i در خوشه‌بندی i-ام نیز می‌باشد. این مقدار با

توجه به الگوریتم شکل ۴ در $sim(C_i,P(D))$ و سپس در $score[i]$ ذخیره می‌گردد.



شکل (۵): محاسبه پایداری خوشه C_i با روش مبتنی بر *NMI*

پایداری کل از میانگین کل این پایداری‌ها تشکیل می‌شود.

$$Stability(C_i)=\frac{1}{M}\sum_{i=1}^{M}NMI_i \qquad (۳)$$

که M تعداد خوشه‌بندی‌ها در مجموعه مرجع می‌باشد. این روش، در واقع خوشه‌هایی را که بیشترین تکرار را در خوشه‌بندی‌های مختلف دارند، به عنوان خوشه‌های پایدارتر معرفی می‌کند.

۲-۴- انتخاب زیرمجموعه‌ای از خوشه‌های اولیه

پس از اینکه پایداری هر خوشه محاسبه شد، در گام بعد، عمل انتخاب خوشه‌ها با توجه به مقدار پایداری خوشه انجام می‌شود. یک روش مبتنی بر الگوریتم‌های تکاملی برای انتخاب زیرمجموعه‌ای از خوشه‌های اولیه ارائه شده است که در این بخش به تشریح این الگوریتم می‌پردازیم.

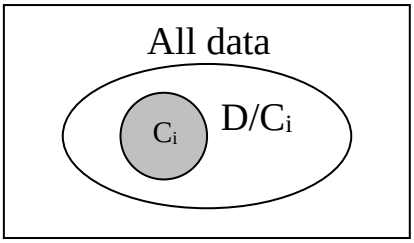
در اینجا عمل انتخاب خوشه‌ها در دو فاز انجام می‌پذیرد. ابتدا در فاز ۱، یک الگوریتم تکاملی سعی در یافتن زیرمجموعه‌ای از خوشه‌ها دارد که بیشترین پایداری را داشته باشند. این الگوریتم تکاملی دارای یک کروموزوم بیتی به طول تعداد کل خوشه‌های تولید شده در بخش تولید خوشه‌بندی‌های گوناگون می‌باشد. هر کدام از ژن‌های این کروموزوم می‌تواند عدد یک یا صفر را به خود بگیرد. عدد یک نشان‌دهنده آن می‌باشد که خوشه‌ی به شماره آن ژن در بین خوشه‌های انتخاب شده باشد و عدد صفر در یک ژن شماره m یعنی خوشه‌ی m-ام در بین خوشه‌های انتخاب شده نباشد. برای محاسبه تابع برازندگی این الگوریتم تکاملی، اختلاف میزان پایداری میانگین خوشه‌های انتخاب شده از عدد یک (حداکثر میزان پایداری میانگین خوشه‌های انتخاب شده می‌باشد)، محاسبه می‌شود.

$$FitnessFunction=1-\sum_i\sum_j\frac{Similarity(C_i,C_j)}{Card.(SelectedClusters)^2} \qquad (۴)$$
$$,i,j\in SelectedClusters$$

که $Similarity(C_i,C_j)$ از رابطه زیر محاسبه می‌شود.

$$Similarity(C_i,C_j)=NMI(P(C_i,D/C_i),P(C_j,D/C_j)) \qquad (۵)$$

در رابطه ۵، $P(C_i,D/C_i)$ یک خوشه بندی است که حاوی دو خوشه‌ی C_i و D/C_i (خوشه‌ای که همه‌ی داده‌ها را به جز داده‌های موجود در C_i در برمی‌گیرد) می‌باشد. این خوشه‌بندی در شکل ۶ نمایش داده شده است.

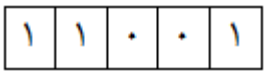


شکل (۶): افراز یک خوشه و مکمل آن

سپس در فاز ۲، برای تنوع بیشتر خوشه‌های انتخاب شده‌ی نهایی، زیرمجموعه‌ای از خوشه‌ها به کمک الگوریتم تکاملی انتخاب می‌شوند که کمترین پایداری را داشته باشند. در اینجا نیز الگوریتم تکاملی دارای یک کروموزوم بیتی به طول تعداد کل خوشه‌ها است. تابع برازندگی این الگوریتم تکاملی میزان پایداری میانگین خوشه‌های انتخاب شده می باشد.

$$FitnessFunction=\sum_i\sum_j\frac{Similarity(C_i,C_j)}{Card.(SelectedClusters)^2} \qquad (۶)$$
$$,i,j\in SelectedClusters$$

لازم به ذکر است که در هر مورد، هدف ما کمینه کردن تابع برازندگی می‌باشد، بنابراین هر چه مقدار تابع کوچکتر باشد، مجموعه‌ی خوشه‌های مشخص شده توسط کروموزوم، برازنده‌تر است. ما در اینجا به مجموعه‌ی انتخاب شده توسط الگوریتم تکاملی اول S_1 و به مجموعه‌ی انتخاب شده توسط الگوریتم تکاملی دوم S_2 می‌گوییم (شکل ۹).



شکل (۷): نمایش یک راه حل کاندید (کروموزوم)

برای روشن‌تر شدن روش انتخاب خوشه‌ها، مثال زیر را در نظر بگیرید. فرض کنید پنج خوشه‌ی اولیه C_1 تا C_5 داریم و می‌خواهیم تعدادی از این خوشه‌ها را به کمک الگوریتم تکاملی انتخاب نماییم. از آنجایی که طول کروموزوم بیتی به اندازه‌ی تعداد کل خوشه‌های اولیه تولید شده می‌باشد، بنابراین یک کروموزوم بیتی به طول ۵ خواهیم داشت. شکل ۷ یک راه حل

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol13 No.2 Summer 2016

کاندید را توسط کروموزوم بیتی نمایش می دهد که مشخص کننده‌ی انتخاب خوشه‌های C_1، C_2 و C_5 می‌باشد.

خوشه‌ها	C_1	C_2	C_3	C_4	C_5
C_1	۱	۰/۱۶	۰/۱۵	۰/۱۴	۰/۶۵
C_2	۰/۱۶	۱	۰/۹۷	۰/۴۲	۰/۰۳
C_3	۰/۱۵	۰/۹۷	۱	۰/۹۱	۰/۸۴
C_4	۰/۱۴	۰/۴۲	۰/۹۱	۱	۰/۹۳
C_5	۰/۶۵	۰/۰۳	۰/۸۴	۰/۹۳	۱

شکل (۸): ماتریس *Similarity* بین خوشه‌های اولیه

سپس الگوریتم تکاملی با استفاده از ماتریس $Similarity$ بین خوشه‌ها که از رابطه ۵ به دست می‌آید، تابع برازندگی راه حل کاندید را محاسبه می‌نماید. شکل ۸ یک ماتریس $Similarity$ نمونه بین خوشه‌های اولیه و همچنین چگونگی محاسبه تابع برازندگی راه حل کاندید شکل ۷ را نشان می‌دهد.

با توجه به ماتریس $Similarity$ و با استفاده از رابطه ۴، تابع برازندگی برای انتخاب پایدارترین خوشه‌ها محاسبه می‌شود.

$$FitnessFunction = 1 - \frac{1+0.16+0.65+0.16+1+0.03+0.65+0.03+1}{9} = 0.48$$

با توجه به رابطه ۶، بدیهی است که مقدار تابع برازندگی برای انتخاب ناپایدارترین خوشه‌ها برابر با ۰/۵۲ می‌باشد.

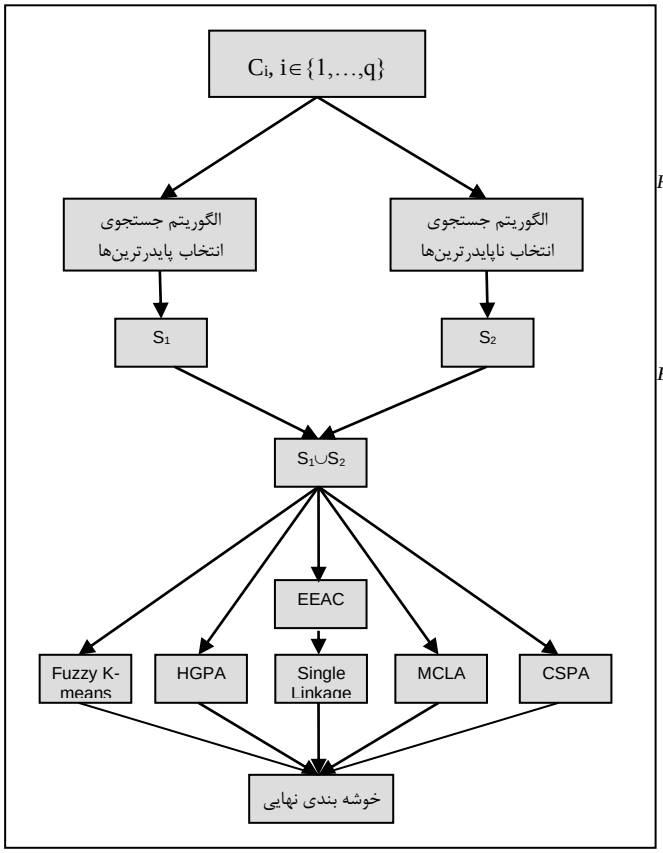
$$FitnessFunction = \frac{1+0.16+0.65+0.16+1+0.03+0.65+0.03+1}{9} = 0.52$$

۳-۴- ساخت نتایج افراز نهایی

در این مرحله خوشه‌های انتخاب شده با هم ترکیب شده و خوشه‌های نهایی از آنها به دست می‌آید. روش‌های مختلفی برای ترکیب خوشه‌بندی‌های اولیه و به دست آوردن خوشه‌های نهایی وجود دارد. برای این کار در اینجا از روش‌های HGPA، MCLA، CSPA و Fuzzy K-Means به عنوان جمع‌کننده نهایی استفاده شده است. در اینجا در حقیقت فرض شده است که برچسب خوشه‌ها برای داده‌ها، داده‌ها را به فضای جدیدی انتقال داده و در فضای جدید خوشه‌بندی را با یکی از روش‌های بالا انجام می‌دهیم.

راه دیگر استفاده از روش‌های خوشه‌بندی سلسله مراتبی می‌باشد که در اینجا از روش Single Linkage سلسله مراتبی استفاده شده است. تفاوتی که در اینجا وجود دارد این است که در این روش ممکن است که از هر خوشه‌بندی اولیه، تنها تعدادی از خوشه‌ها موجود باشند. از آن جایی که استفاده از روش انباشت مدارک (EAC) نمی‌تواند شباهت بین جفت نمونه‌ها را در حضور تنها تعدادی از خوشه‌ها به درستی تشخیص دهد، در این پایان‌نامه از دو روش علیزاده [۲] برای ترکیب نتایج استفاده شده است. این روش‌ها که توانایی استخراج ماتریس همبستگی برای نمونه‌ها -در شرایطی که تنها تعدادی از خوشه‌ها موجود هستند- را دارند، در بخش‌های بعدی تشریح شده است. پس از ساخت ماتریس همبستگی، می‌توان با استفاده از یکی از الگوریتم‌های سلسله‌مراتبی نظیر اتصال منفرد یا اتصال میانگین[۶۰]، خوشه‌های نهایی را استخراج کرد.

شکل ۹ چگونگی انتخاب زیرمجموعه‌ای از خوشه‌های اولیه به کمک روش‌های جستجو و همچنین نحوه‌ی بدست آوردن افراز نهایی از خوشه‌های انتخاب شده را نشان می‌دهد.



شکل (۹): روش پیشنهادی برای انتخاب زیرمجموعه‌ای از خوشه‌های اولیه و ساختن افراز نهایی

لازم به ذکر است که برای استفاده از الگوریتم‌های سلسله مراتبی ماتریس همبستگی استخراج می‌شود و خوشه‌های نهایی با اعمال این الگوریتم‌ها از این ماتریس به دست می‌آیند، ولی

برای استفاده از روش Fuzzy K-means و روش‌های مبتنی بر گراف مثل HGPA، MCLA و CSPA نیازی به ساخت ماتریس همبستگی نمی‌باشد و می‌توان خوشه‌بندی نهایی را با اعمال این روش‌ها بر خوشه‌های انتخاب شده به دست آورد.

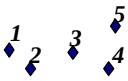
۱-۳-۴- ماتریس اطلاعات تجمعی

در روش انباشت مدارک (EAC) نتایج m خوشه‌بندی روی داده‌های نمونه‌برداری شده در ماتریس همبستگی $n\times n$ ذخیره می‌شوند. هر داده ورودی از این ماتریس در روش انباشت مدارک، به صورت رابطه ۷ محاسبه می‌شود.

$$C(i,j)=\frac{n_{i,j}}{m_{i,j}} \qquad (۷)$$

که $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j با هم در یک خوشه گروه‌بندی شده‌اند و $m_{i,j}$ تعداد نمونه‌برداری‌هایی است که هر دوی این جفت نمونه‌ها به طور همزمان در آن ظاهر شده‌اند.

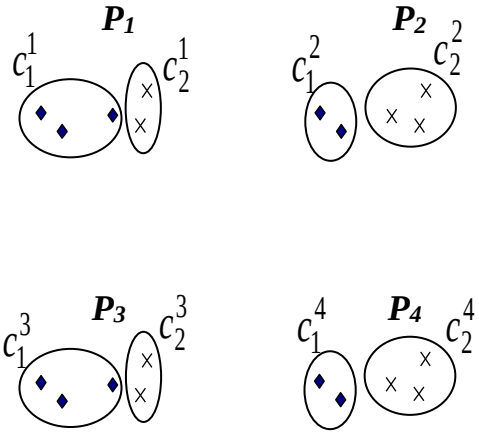
از آن جایی که پس از آستانه‌گیری در این روش ، تنها تعدادی از خوشه‌های اولیه در دسترس می‌باشند، روش EAC نمی‌تواند روابط بین جفت‌نمونه‌ها را تشخیص دهد. بنابراین برای ترکیب نتایج با استفاده از ماتریس همبستگی باید معیاری برای نشان‌دادن همبستگی نمونه‌ها تعریف شود که بتواند شباهت بین نمونه‌ها را با حضور تنها زیرمجموعه‌ای از خوشه‌های اولیه به درستی استخراج و محاسبه کند. این روش برای ساخت ماتریس همبستگی در شرایطی که تعدادی از خوشه‌ها حذف شده‌اند، انباشت مدارک توسعه یافته (EEAC) نامیده می‌شود. هر داده ورودی از ماتریس همبستگی در روش EEAC به صورت رابطه ۸ تعریف می‌شود.



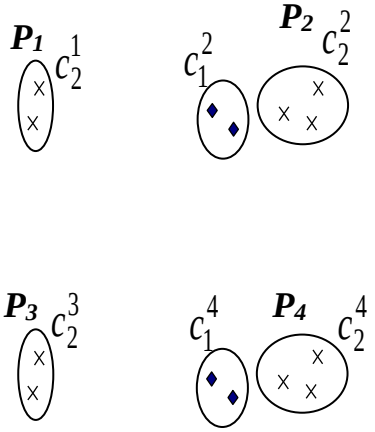
الف- یک مجموعه داده با ۵ داده

$$C(i,j)=\frac{n_{i,j}}{\max(n_i,n_j)} \qquad (۸)$$

n_i تعداد دفعاتی است که نمونه i در خوشه‌های انتخاب شده ظاهر شده است. به طور مشابه n_j نیز، تعداد دفعاتی است که نمونه j در خوشه‌های انتخاب شده ظاهر شده است. همچنین، $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j با هم در یک خوشه از خوشه‌های انتخاب شده ظاهر شده‌اند. بدیهی است که با



ب- نتایج چهار خوشه‌بندی دو-خوشه‌ای اولیه بر داده‌های شکل ۱۰-الف که در کل ۸ خوشه وجود دارد. فرض کنید که پایداری خوشه‌های c_1^1، c_2^1، c_1^2، c_2^2، c_1^3، c_2^3، c_1^4، و c_2^4 به ترتیب ۰،۷، ۰،۹، ۰،۹، ۰،۹، ۰،۶، ۰،۹، ۰،۹، ۰،۹، و ۰،۹ باشد.



ج- نتایج خوشه‌های اولیه انتخابی شکل ۱۰-ب با اعمال آستانه انتخاب ۰،۸

شکل (۱۰): محاسبه پایداری خوشه C_i با روش مبتنی بر NMI الف) مجموعه داده شامل ۵ نمونه ب) نتایج چهار خوشه‌بندی اولیه ج) خوشه‌های باقیمانده پس از آستانه‌گیری

در نظر گرفتن تعداد خوشه‌های ثابت در خوشه‌بندی‌های اولیه همواره n_i و n_j کمتر از تعداد کل افرازهای اولیه و همچنین، تعداد کل خوشه‌های ممکن می‌باشد. یعنی $n_i,n_j \le B \le k\times B$.

برای روشن‌تر شدن بحث، مثال زیر را در نظر بگیرید. فرض کنید یک مجموعه داده شامل ۵ نمونه مطابق شکل ۱۰-الف وجود دارند. همچنین فرض کنید چهار خوشه‌بندی اولیه P_1 تا P_4 به

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol13 No.2 Summer 2016

عنوان خوشه‌بندی‌های اولیه روی آن صورت گرفته است که در شکل ۱۰-ب نمایش داده شده است.

همچنین، فرض کنید که مقادیر پایداری خوشه‌های تولید شده به صورت زیر باشند (دقت شود این اعداد فرضی هستند و نه مقدار واقعی هستند؛ به دلیل آن که مجموعه مرجع را در اینجا نداریم مقدار آنها را نمی‌توان محاسبه کرد).

$$Stability(c_2^1) = Stability(c_2^3) = 1$$
$$Stability(c_1^2) = Stability(c_1^4) = 1$$
$$Stability(c_2^2) = Stability(c_2^4) = 0.82$$
$$Stability(c_1^1) = Stability(c_1^3) = 0.55$$

اگر مقدار آستانه برای انتخاب خوشه‌ها برابر با ۰٫۸ باشد، خوشه‌های اول از افرازهای اول و سوم حذف می‌شوند؛ چراکه پایداری آنها زیر آستانه ۰٫۸ است. اکنون در شکل ۱۰-ج ماتریس همبستگی باید با استفاده از بقیه خوشه‌ها ایجاد شود. اکنون خوشه‌های c_2^1، c_1^2، c_2^3، c_2^2، c_1^4، و c_2^4 انتخاب می‌شوند.

حال درایه C(2,3) در اجماع کل خوشه‌های موجود (یعنی خوشه‌های c_1^1، c_2^1، c_1^2، c_2^2، c_1^3، c_2^3، c_1^4، و c_2^4) درایه C(2,3)، برابر ۰٫۵ است؛ چرا که در ۴ خوشه موجود (یعنی خوشه‌های c_1^1، c_1^2، c_1^3، و c_1^4) داده دوم و در ۴ خوشه موجود (یعنی خوشه‌های c_1^1، c_2^2، c_1^3، و c_2^4) داده سوم حضور دارند؛ پس مخرج رابطه ۸، ۴ است. صورت آن رابطه نیز ۲ است؛ چرا که در دو خوشه (یعنی خوشه‌های c_1^1 و c_1^3) دو داده دوم و سوم به طور همزمان حضور دارند. پس قبل از انتخاب خوشه‌های پایدار، یعنی کل خوشه‌های شکل ۱۰-ب، درایه C(2,3) برابر ۰٫۵ است. اما بعد از انتخاب خوشه‌های پایدار، یعنی شکل ۱۰-ج، مقدار درایه C(2,3) به صفر تنزل پیدا می‌کند. برای نشان دادن این به خوشه‌های شکل ۱۰-ج توجه کنید، در ۲ خوشه (یعنی خوشه‌های c_1^2 و c_1^4) از ۶ خوشه موجود در شکل ۱۰-ج، داده دوم و در ۲ خوشه (یعنی خوشه‌های c_2^2 و c_2^4) از ۶ خوشه موجود در شکل ۱۰-ج، داده سوم حضور دارند؛ پس مخرج رابطه ۸، ۲ است. صورت آن رابطه نیز صفر است؛ چرا که در هیچ خوشه‌ای از ۶ خوشه موجود در شکل ۱۰-ج، دو داده دوم و سوم به طور همزمان حضور ندارند. به طور کامل مقادیر ماتریس C قبل از عمل انتخاب، در زیر آورده شده است.

$$c(1,2) = c(4,5) = \frac{4}{\max(4,4)} = 1$$
$$c(1,3) = c(2,3) = \frac{2}{\max(4,4)} = 0.5$$
$$c(1,4) = c(1,5) = c(2,4) = c(2,5) = c(3,4) = c(3,5) = \frac{0}{\max(4,4)} = 0$$

پس ماتریس همبستگی قبل از آستانه‌گیری و انتخاب، به صورت رابطه ۹ خواهد بود.

$$C_{before} = \begin{bmatrix} ۱ & ۱ & ۰.۵ & ۰ & ۰ \\ ۱ & ۱ & ۰.۵ & ۰ & ۰ \\ ۰.۵ & ۰.۵ & ۱ & ۰.۵ & ۰.۵ \\ ۰ & ۰ & ۰.۵ & ۱ & ۱ \\ ۰ & ۰ & ۰.۵ & ۱ & ۱ \end{bmatrix} \quad (۹)$$

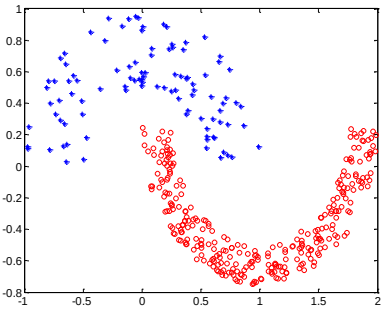
در این ماتریس نمونه سوم می‌تواند با احتمال ۵۰٪ به هر کدام از دو خوشه بچسبد. اگر بتوان اطلاعات اضافه‌تری به این ماتریس اضافه کرد، به گونه‌ای که خوشه پایدارتر دارای وزن بیشتری در مقادیر همبستگی بین نمونه‌هایش شود، می‌توان به عملکرد بهتر خوشه‌بندی ترکیبی امیدوار بود.

مقادیر ماتریس C بعد از انتخاب، در زیر آورده شده است.

$$c(1,2) = \frac{2}{\max(2,2)} = 1$$
$$c(1,3) = c(2,3) = \frac{2}{\max(4,4)} = 0.5$$
$$c(1,3) = c(1,4) = c(1,5) = c(2,3) = c(2,4) = c(2,5) = \frac{0}{\max(2,4)} = 0$$
$$c(3,4) = c(3,5) = \frac{2}{\max(2,4)} = 0.5$$
$$c(4,5) = \frac{4}{\max(4,4)} = 1$$

پس ماتریس همبستگی بعد از آستانه‌گیری و انتخاب، به صورت رابطه ۱۰ خواهد بود.

[DOR: 20.1001.1.26765810.1395.13.2.13.9]



شکل(۱۱): مجموعه داده "نیم‌حلقه‌ها" با ۴۰۰ الگو (۱۰۰-۳۰۰ در هر رده)

$$C_{after}=\begin{bmatrix}1&1&0&0&0\\1&1&0&0&0\\0&0&1&0.5&0.5\\0&0&0.5&1&1\\0&0&0.5&1&1\end{bmatrix} \qquad (۱۰)$$

با مشاهده این دو ماتریس می‌توان دریافت که چگونه حذف خوشه‌های ناپایدار می‌تواند باعث بهبود ماتریس همبستگی شود. از آن جایی که خوشه مربوط به نمونه‌های $\{۱،۲،۳\}$ دارای مقدار پایداری پایینی می‌باشند، حذف آنها موجب روشن‌تر شدن ماتریس همبستگی می‌شود. اکنون یک الگوریتم سلسله‌مراتبی ساده نیز می‌تواند خوشه‌های موجود در ماتریس همبستگی (بعد از انتخاب) را استخراج کند.

پس از اینکه ماتریس همبستگی با روش EEAC ساخته شد، در مرحله بعد از یک تابع توافقی برای استخراج خوشه‌های نهایی از این ماتریس استفاده می‌شود. معمولا از یکی از الگوریتم‌های سلسله‌مراتبی برای استخراج خوشه‌های نهایی از ماتریس همبستگی استفاده می‌شود. در این مقاله از الگوریتم سلسله‌مراتبی اتصال منفرد[۶۱] استفاده شده است.

۲-۳-۴- اشتراک به اجتماع

این روش نیز مشابه روش انباشت مدارک توسعه یافته عمل می‌کند. ایده اصلی پشت این روش شمارش تمام حالت‌های ممکنی است که نمونه‌های i و j نسبت به هم در خوشه‌های انتخاب شده داشته‌اند. در واقع تفاوت اصلی این دو روش در مخرج کسر می‌باشد. هر داده ورودی از ماتریس همبستگی در روش اشتراک به اجتماع[۶۲] (ItoU) به صورت رابطه ۱۱ تعریف می‌شود.

$$C(i,j)=\frac{\bigcap(n_i,n_j)}{\bigcup(n_i,n_j)}=\frac{n_{i,j}}{n_i+n_j-n_{i,j}} \qquad (۱۱)$$

n_i تعداد دفعاتی است که نمونه i در خوشه‌های انتخاب شده ظاهر شده است. به طور مشابه n_j نیز، تعداد دفعاتی است که نمونه j در خوشه‌های انتخاب شده ظاهر شده است. همچنین، $n_{i,j}$ تعداد دفعاتی است که جفت نمونه‌های i و j با هم در یک خوشه از خوشه‌های انتخاب شده ظاهر شده‌اند.

۵- نتایج و تفسیر

در این بخش نتایج تجربی برای ارزیابی روش پیشنهادی با استفاده از چندین مجموعه داده گزارش شده است. مجموعه داده‌های استفاده شده مجموعه داده‌های استاندارد UCI می‌باشد که تقریبا نتایج تمام مطالعات اخیر دنیا در زمینه خوشه‌بندی با استفاده از این مجموعه داده‌ها گزارش می‌شوند. همچنین، نتایج آزمایش‌ها که نشان از کارایی نسبتا بالای روش‌های پیشنهادی در مواجهه با مجموعه داده‌های مختلف می‌باشد، در این بخش ارایه شده است.

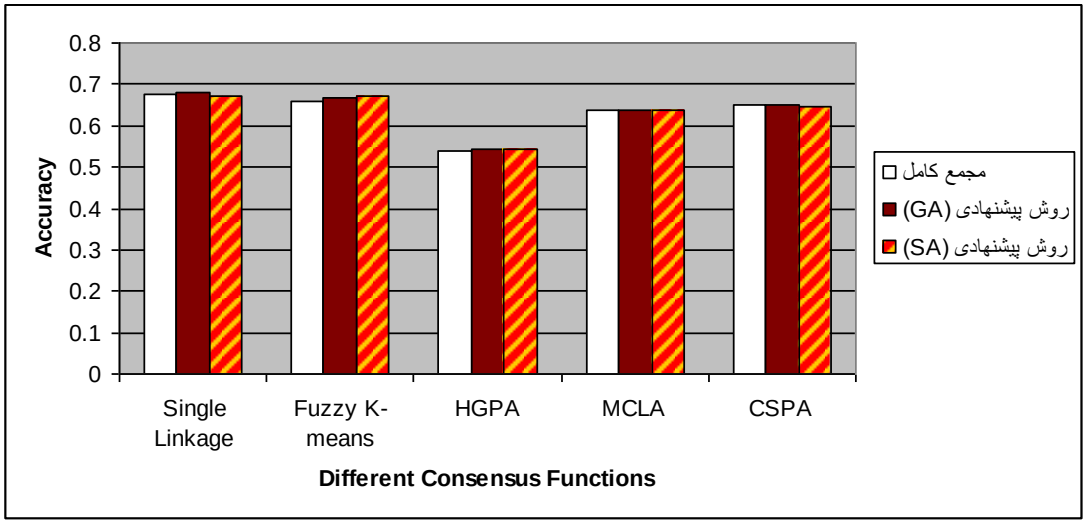
۱-۵- مجموعه داده‌ها

روش پیشنهادی بر روی ۱۰ مجموعه داده استاندارد نرمال شده مورد آزمایش قرار گرفته است. برای انجام آزمایش‌ها سعی شده است تا مجموعه داده‌ها از لحاظ تعداد کلاس‌ها، تعداد ویژگی‌ها و همچنین تعداد نمونه‌ها از حداکثر تنوع برخوردار باشند تا نتایج آزمایش‌ها تا حد ممکن دارای استحکام و قابل تعمیم باشد. جدول ۱ اطلاعات مختصری از مجموعه داده‌های استاندارد مورد استفاده را در اختیار می‌گذارد.

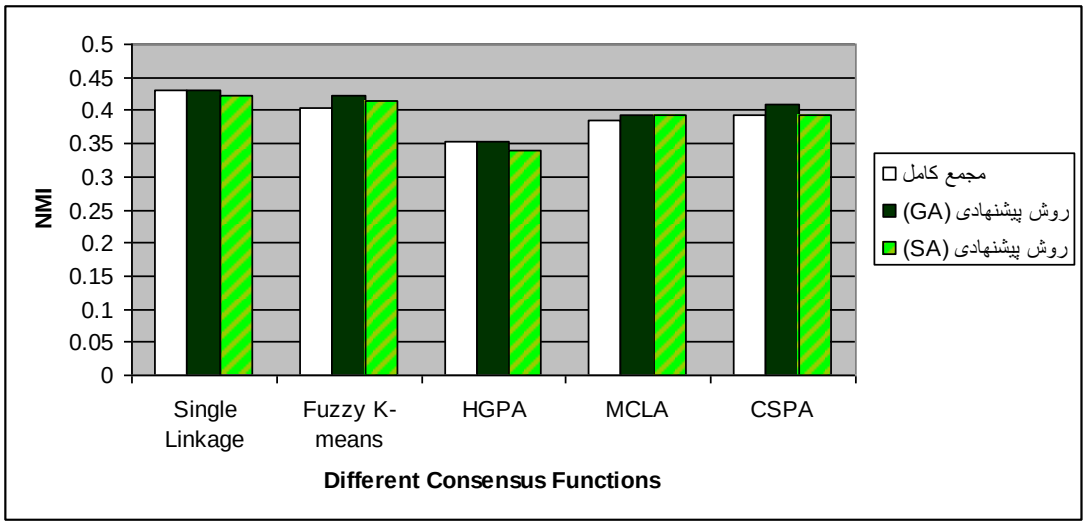
به خاطر مصنوعی‌بودن مجموعه داده نیم‌حلقه‌ها، این مجموعه داده نیز توصیف شده است. برای اطلاعات بیشتر در مورد هر کدام از مجموعه داده‌های جدول ۱ (غیر از نیم حلقه‌ها) می‌توان به [۵۷] رجوع کرد.

مجموعه داده نیم‌حلقه‌ها که یک مجموعه داده مصنوعی می‌باشد -همانطور که در شکل ۱۱ نشان داده شده است- شامل دو خوشه می‌باشد که خوشه‌ها با ۱۰۰ نقطه و ۳۰۰ نقطه، نامتوازن هستند. الگوریتم K-means به تنهایی قادر به تشخیص دو خوشه طبیعی نیست، به خاطر اینکه صورت پیش‌فرض خوشه‌ها را به شکل ابرکره تصور می‌کند.

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol13 No.2 Summer 2016



شکل (۱۲): نمودار میانگین دقت بر روی تمام مجموعه داده‌ها در مقابل روش‌های گوناگون



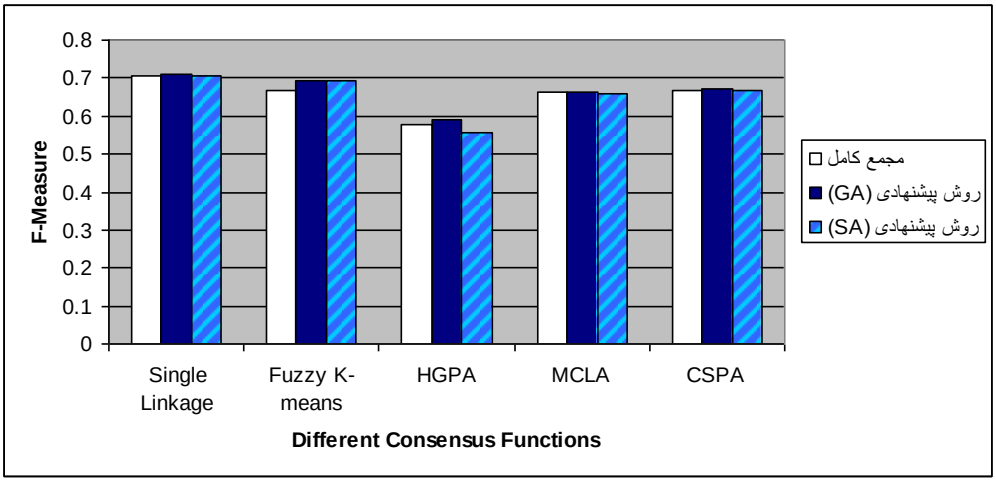
شکل (۱۳): نمودار میانگین اطلاعات متقابل نرمال شده بر روی تمام مجموعه داده‌ها در مقابل روش‌های گوناگون

تعدادی از نتایج آزمایش‌ها بر روی ویژگی‌های نرمال شده از این مجموعه داده‌ها گزارش شده است. برای عملیات نرمال‌سازی، هر کدام از ویژگی‌های این مجموعه داده‌ها با میانگین صفر و واریانس یک، $N(0,1)$، نرمال شده‌اند. برای همه این مجموعه داده‌ها، تعداد خوشه‌ها و برچسب واقعی نمونه‌ها از قبل معلوم هستند. بنابراین، درصد نمونه‌هایی که درست تشخیص داده شده‌اند، به عنوان معیار کارایی روش خوشه‌بندی مورد استفاده قرار گرفته است. در واقع بعد از حل مسئله تناظر بین برچسب‌های به دست آمده و خوشه‌های واقعی، می‌توان نرخ خطا را تعیین کرد. همچنین NMI و F-Measure نیز گزارش خواهد شد.

۲-۵- مجموعه داده‌ها

در در این بخش نتایج به کارگیری روش پیشنهادی روی مجموعه داده‌های مختلف و پارامترهای مورد استفاده گزارش شده است. روش پیشنهادی در محیط MATLAB[7.1] پیاده‌سازی و مورد آزمایش قرار گرفته است. نتایج آزمایش‌ها روی میانگین ۳۰ بار اجرای مستقل برنامه گزارش شده است. عملکرد روش‌های مختلف خوشه‌بندی با سه معیار دقت، NMI و F-Measure محاسبه شده است.

[DOR: 20.1001.1.26765810.1395.13.2.13.9]



شکل (۱۴): نمودار میانگین معیار فیشر بر روی تمام مجموعه داده‌ها در مقابل روش‌های گوناگون

در تمامی روش‌های مورد استفاده از الگوریتم K-means به عنوان الگوریتم پایه استفاده شده است. تعداد نتایج اولیه تولید شده نیز در تمام روش‌ها ثابت و برابر با ۱۲۰ می‌باشد. در واقع این تعداد با دستکاری پارامتر k از الگوریتم K-means به دست آمده است. به این صورت که چهار گروه ۳۰ تایی از نتایج اولیه، با در نظر گرفتن تعداد خوشه‌های مورد استفاده توسط این الگوریتم با اندازه‌های k، k+1، k+2 و k+3 حاصل شده است. همچنین، برای ایجاد پراکندگی بیشتر در نتایج اولیه از نمونه‌برداری بدون جاگذاری با نرخ ۵۰٪ استفاده شده است. همچنین، برای ساختن افراز نهایی از روش اتصال منفرد[۶۳] بر روی ماتریس همبستگی، روش Fuzzy K-means و روش‌های مبتنی بر گراف HGPA، MCLA و CSPA استفاده شده است.

جدول (۱): خلاصه‌ای از مشخصه‌های مجموعه داده‌های استاندارد مورد استفاده

مجموعه داده	تعداد کلاس	تعداد ویژگی	تعداد نمونه
Wine	3	13	178
breast-cancer	2	9	683
Bupa	2	6	345
Galaxy	7	4	323
Glass	6	9	214
Halfring	2	2	400
Iris	3	4	150
Ionosphere	2	34	351
Saheart	2	9	462
Yeast	10	8	1484

۳-۵- اعمال الگوریتم‌های تکاملی برای انتخاب خوشه‌ها

الگوریتم‌های تکاملی به کار رفته در این مقاله، یکی الگوریتم ژنتیک و دیگری نورد شبیه سازی شده بوده است. پارامترهای الگوریتم ژنتیک شامل اندازه‌ی جمعیت ۱۰۰۰، تعداد نسل‌های ۵۰۰ و طول کروموزوم برابر با ۱۲۰ برابر تعداد خوشه‌های واقعی به علاوه ۱۸۰ می‌باشد. همچنین از احتمال جهش ۰/۰۱ عملگر تقطیع یکنواخت و عملگر انتخاب مرتب سازی استفاده شده است.

الگوریتم نورد شبیه‌سازی شده نیز با T=۰/۹ انجام پذیرفته است. میزان خطای دو جواب پی‌درپی الگوریتم نورد شبیه‌سازی شده نباید کمتر از ۰/۰۰۱ باشد (ε=۰/۰۰۱). این الگوریتم هم از همان نمایش کروموزوم الگوریتم ژنتیک و تابع برازندگی آن استفاده می‌کند.

جدول (۲): مقایسه دقت روش پیشنهادی با روش مجمع کامل و روش‌های پیشین

مجموعه داده	مجمع کامل	GA	SA	علیزاده	عظیمی
Wine	**96.74**	96.63	96.63	96.63	96.63
Breast-Cancer	**97.03**	95.29	95.17	95.73	95.91
Bupa	55.01	**55.10**	55.07	54.33	54.75
Galaxy	30.03	**32.82**	30.65	31.27	29.97
Glass	56.81	**57.86**	45.79	57.76	55.05
Halfring	**76.38**	74.50	74.50	74.48	67.70
Iris	**89.60**	89.33	89.33	89.33	89.33
Ionosphere	70.71	**70.74**	70.65	70.60	**70.74**
Saheart	**63.44**	63.29	63.27	63.36	56.06
Yeast	39.42	42.93	43.05	42.75	**43.40**
ALL	67.517	**67.831**	66.411	67.642	65.954

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol13 No.2 Summer 2016

در عمل اعمال الگوریتم تکاملی برای انتخاب خوشه‌ها، دو مجموعه‌ی خوشه‌های پایدار و خوشه‌های غیر پایدار به دست می‌آید. نتایج مجمع کامل و مجمعی که خوشه‌های را با استفاده از الگوریتم تکاملی انتخاب می‌کند بر حسب NMI و F-Measure و دقت بر روی مجموعه داده‌های گوناگون محاسبه و برای سهولت در تحلیل، میانگین نتایج بر روی ۱۰ مجموعه داده در شکل‌های ۱۲، ۱۳ و ۱۴ آورده شده است. چنان که مشاهده می‌گردد نتایج نه تنها کاهش نداشته، بلکه در اغلب موارد بهبود نیز یافته است.

چنان که در اشکال میانگین‌ها مشاهده می‌شود، در اغلب موارد بهبود کارایی حاصل شده است، لذا نتیجه می‌گیریم که نه تنها کاهش خوشه‌های انتخاب شده باعث کاهش کارایی نشده است بلکه افزایش کارایی را نیز در اکثر موارد موجب شده است.

همچنین از آن جهت که این کار در ادامه‌ی کاری است که قبلا توسط علیزاده و مینایی انجام شده است، مقایسه‌ای بین این کار با کار آنها و همچنین کار پیشتر از آن که توسط آقای عظیمی انجام شده است، صورت پذیرفته است.

برای آن که تطبیقی بین نتایج روش این مقاله با روش‌های مرتبط پیشینی که در دانشگاه علم و صنعت انجام شده است وجود داشته باشد، نتایج جدول ۲ همگی با تابع Single Linkage به عنوان تابع جمع‌کننده ارائه شده‌اند.

همان‌طور که از نتایج بر می‌آید، این روش نه تنها از روش‌های پیشین بدتر عمل نمی‌کند، بلکه در مواردی بهتر نیز عمل می‌کند. علت این کار می‌تواند در این باشد که مجموعه‌ی انتخاب شده توسط الگوریتم ژنتیک اول از پایداری بالایی برخوردار است و در حقیقت خوشه‌های ساده را به درستی در خود نشان می‌دهد. همچنین الگوریتم ژنتیک دوم خوشه‌هایی را انتخاب می‌کند که بیشترین تنوع را داشته باشند و این کمک می‌کند به اصل تنوع نتایج و بالا رفتن دقت نتایج نهایی که این پیش از انجام آزمایشات انتظار می‌رفت. در حقیقت این نتایج دور از انتظار نبودند.

اگر چه روش پیشنهادی از نظر دقت در جدول ۲، بهتر از سایر روش‌ها بوده است، لیکن هنوز نمی‌توان ادعایی دال بر این که بهترین روش روش پیشنهادی است نمی‌توان داشت. باید دید که آیا این نتایج تصادفی نبوده باشد که با تغییر دوباره پارامترها و مقداردهی اولیه متفاوت، نتایج به گونه‌ای دیگر رقم نخواهد خورد. برای بررسی دقیق‌تر و پی بردن به این نکته که آیا این برتری بامعنی است، باید به یکی از روش‌های راستی‌سنجی آماری پناه برد. در این جا از روش راستی‌سنجی آماری فریدمن استفاده می‌کنیم. این روش را به این دلیل بر می‌گزینیم که مناسب مقایسه چندین روش به طور همزمان است. این روش هر یک از روش‌ها را بر اساس کارایی آنها در یک مجموعه داده مرتب می‌کند و رتبه روشی با بیشترین کارایی را ۱ در نظر می‌گیرد و رتبه روشی با کمترین کارایی را M (M تعداد روش‌ها است) در نظر می‌گیرد. در مواردی که چند روش با رتبه‌های یکسان باشد، میانگین رتبه برای آنها در نظر گرفته می‌شود. برای مثال اگر روش A و B دارای دومین و سومین روش کارا باشند، یعنی رتبه‌های آنها ۲ و ۳ باشد، ولی کارایی آنها برابر باشد، رتبه‌های آنها به ترتیب برابر ۲,۵ و ۲,۵ خواهند بود. در ادامه این روش به تفصیل امده است.

فرضیه صفر روش فریدمن بیان می‌دارد که روش‌ها تفاوت با معنی ندارند. برای رد این فرضیه نشان دادن اینکه روش‌ها تفاوت با معنی دارند، باید به شکل زیر اقدام کنیم:

ابتدا فرض کنید r_i^j نشان دهنده رتبه روش i-ام در مجموعه داده j-ام باشد. میانگین رتبه روش j-ام از رابطه ۱۲ محاسبه می‌شود.

$$R_j = \frac{1}{N}\sum_{i=1}^{N} r_i^j \qquad (۱۲)$$

میانگین رتبه‌های همه روش‌ها را محاسبه می‌کنیم. درجه آزادی مسئله $k-1$ است که k نشان‌گر تعداد روش‌ها است. چون ۵ روش داریم، درجه آزادی مسئله ۴ می‌باشد. حال با رابطه ذیل مقدار χ_F^2 را از رابطه ۱۳ محاسبه می‌کنیم.

$$\chi_F^2 = \frac{12N}{k(k+1)}\left(\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right) \qquad (۱۳)$$

با محاسبه رابطه ۱۳، مقدار χ_F^2 برابر ۱۰/۱۶ خواهد شد که این مقدار از مقدار مورد انتظار ما در جدول توزیع چای با درجه آزادی ۴ که برابر ۹/۴۸۹ است بیشتر است. پس فرض صفر رد شده و به این نتیجه می‌رسیم که اختلاف‌های بین روش‌ها با معنی می‌باشد. از آنجا که رتبه متوسط هر یک از روش‌های (الف) مجمع کامل، (ب) GA، (ج) SA، (د) علیزاده و (ی) عظیمی به ترتیب ۲,۳، ۲,۰، ۳,۴، ۳,۵ و ۳,۸ است، پس روش GA به شکل بامعنایی از سایر روش‌ها بهتر است.

۶- جمع‌بندی و کارهای آینده

در این مقاله یک روش برای بهبود کارایی خوشه‌بندی ترکیبی پیشنهاد شد. این روش همانند کار پیشین علیزاده [۲] و عظیمی [۱] مبتنی بر به‌کارگیری زیرمجموعه‌ای از نتایج اولیه می‌باشند. برای انتخاب زیرمجموعه‌ای از نتایج اولیه، نیاز به تعریف معیار

[DOR: 20.1001.1.26765810.1395.13.2.13.9]

Journal of Iranian Association of Electrical and Electronics Engineers - Vol.13- No.2 Summer 2016

ارزیابی می‌باشد که در این تحقیق همچون کارهای پیشین از معیار NMI ویرایش شده برای ارزیابی خوشه استفاده شده است. نتایج تجربی انجام شده بر روی ۱۰ مجموعه داده استاندارد نشان از کارایی بالای روش‌های پیشنهادی در مقایسه با سایر روش‌های خوشه‌بندی ترکیبی دارد. همچنین، نتایج نشان می‌دهند که اگرچه در روش‌های پیشنهادی تنها مقدار کمی از نتایج اولیه در ترکیب نهایی استفاده می‌شوند، کارایی این روش‌ها حتی از روش ترکیب کامل[۶۴] هم بالاتر است. به طور دقیق‌تر روش‌های اطلاعات نرمال اصلاح شده (ENMI) در ترکیب با روش اشتراک به اجتماع برای ساخت ماتریس همبستگی، بیشترین بهبود در نتایج را ایجاد می‌کند. این بهبود روی کلیه ۱۰ مجموعه آزمایش‌شده نسبت به روش‌های ترکیب کامل، روش علیزاده و روش عظیمی قابل مشاهده می‌باشد.

از جمله کارهایی که می‌توان در ادامه این تحقیق به آن پرداخت، جستجوی بیشتر برای ارایه معیارهای مناسب برای ارزیابی یک خوشه و یا یک افراز از داده‌ها می‌باشد. اگر بتوان یک معیار یا یک مجموعه از معیارهای دقیق برای ارزیابی برازندگی خوشه و افراز یافت که مستقل از نوع خوشه‌ها و دادگان عمل کند، می‌توان با استفاده از آنها، از روش‌های جستجوی هوشمند نیز بهره گرفت. همچنین می‌توان از راهکارهای تولید تنوع شبیه تقویت[۶۵] و کیسه[۶۶] برای مرحله‌ی تولید استفاده کرد. راهکارهایی در روش‌های ترکیب رده‌بندها برای تولید تنوع وجود دارد که می‌توان با توجه به آنها از بین چندین رده‌بند متنوع‌ترین رده-بندها را انتخاب کرد. می‌توان از این روش‌ها در خوشه‌بندی ترکیبی نیز استفاده کرد.

سپاسگزاری

این مقاله مستخرج از یک طرح پژوهشی به نام دانشگاه آزاد اسلامی، واحد استهبان بوده است و تحت حمایت مالی این واحد بوده است.

مراجع

[۱] عظیمی ج.، بررسی پراکندگی در خوشه‌بندی ترکیبی، پایان نامه کارشناسی ارشد، دانشگاه علم و صنعت ایران، خرداد، ۱۳۸۶.

[۲] علیزاده ح.، خوشه بندی ترکیبی مبتنی بر زیرمجموعه ای از نتایج اولیه، پایان نامه کارشناسی ارشد، دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران، اسفند، ۱۳۸۷.

[۳] علیزاده ح.، حسین زاده ر.، ناظمی ا.، تشخیص اجتماعات ترکیبی در شبکه های اجتماعی، نشریه مهندسی برق و الکترونیک ایران، ۱۱(۲): ۶۰-۴۹، ۱۳۹۳.

[۴] مقیمی م.، اکبری پور ح.، امین‌ناصری م.ر.، طراحی سیستم خبره به منظور تشخیص حمله‌های فیشینگ در بانکداری الکترونیکی، نشریه مهندسی برق و الکترونیک ایران، ۱۲(۲): چاپ آنلاین، ۱۳۹۴.

[5] Jain A., Murty M. N., and Flynn P., Data clustering: A review. ACM Computing Surveys, 31(3):264–323, 1999.

[6] Faceli K., Marcilio C.P. Souto d., Multi-objective Clustering Ensemble, Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06),2006

[7] Strehl A. and Ghosh J., Cluster ensembles - a knowledge reuse framework for combining multiple partitions. Journal of Machine Learning Research, 3(Dec):583–617, 2002.

[8] Melanie M., An Introduction to Genetic Algorithms, A Bradford Book The MIT Press, Cambridge, Massachusetts. London, England, Fifth printing, 1999.

[9] Aarts E. H. L. and Korst J., Simulated Annealing and Boltzmann Machines, John Wiley & Sons, Essex, U.K, 1989.

[10] Fred, A. and Jain, A. K., "Data Clustering Using Evidence Accumulation", Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR02, Quebec City, pp. 276 – 280, 2002.

[11] Parvin H., Alizadeh H. and Minaei-Bidgoli B., A New Method for Constructing Classifier Ensembles, International Journal of Digital Content: Technology and its Application, JDCTA, ISSN: 1975-9339, 2009.

[12] Parvin H., Alizadeh H. and Minaei-Bidgoli B., Using Clustering for Generating Diversity in Classifier Ensemble, International Journal of Digital Content: Technology and its Application, JDCTA, ISSN: 1975-9339, Vol. 3, No.1, pp. 51-57, 2009.

[13] Alizadeh H., Minaei-Bidgoli B. and Amirgholipour S.K., A New Method for Improving the Performance of K Nearest Neighbor using Clustering Technique, International Journal of Convergence Information Technology, JCIT, ISSN: 1975-9320,2009.

[14] Topchy, A., Jain, A.K. and Punch, W.F., Combining Multiple Weak Clusterings, Proc. 3d IEEE Intl. Conf. on Data Mining, pp. 331-338,2003.

[15] Fred A. and Lourenco A., Cluster Ensemble Methods: from Single Clusterings to Combined Solutions, Studies in Computational Intelligence (SCI), 126, 3–30, 2008.

[16] Ayad H.G. and Kamel M.S., Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters, IEEE Trans. on Pattern

[DOR: 20.1001.1.26765810.1395.13.2.13.9]

Journal of Iranian Association of Electrical and Electronics Engineers Vol13 No.2 Summer 2016

[32] Lange T., Roth V., Braun M.L., and Buhmann J.M., Stability-based validation of clustering solutions. Neural Computation, 16(6):1299–1323, 2004.
, 2004.Ben-Hur A., Elisseeff A. and Guyon I. (2002), A stability based method for discovering structure in clustered data, in Pasific Symposium on Biocomputing, vol. 7, pp. 6-17.
[33] Estivill-Castro V. and Yang J., Cluster Validity Using Support Vector Machines, DaWaK 2003, LNCS 2737, pp. 244–256, 2003.
[34] Moller U., Radke D., A Cluster Validity Approach based on Nearest-Neighbor Resampling, In Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06), 2006.
[35] Brandsma T. and Buishand T.A., "Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling", Hydrology and Earth System Sciences 2, pp. 195-209, 1998.
[36] Inokuchi R., Nakamura T. and Miyamoto S. , Kernelized Cluster Validity Measures and Application to Evaluation of Different Clustering Algorithms, in proc. of the IEEE Int. Conf. on Fuzzy Systems, Canada, July 16-21, 2006.
[37] Xie X.L., Beni G., A Validity measure for Fuzzy Clustering, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.13, No.4, pp. 841–846, 1991.
[38] Das A.K. and Sil J., Cluster Validation using Splitting and Merging Technique, in proc. of Int. Conf. on Computational Intelligence and Multimedia Applications, ICCIMA, 2007.
[39] Fern X. and Lin W., Cluster Ensemble Selection, SIAM International Conference on Data Mining (SDM08),2008.
[40] Brossier G., Piecewise hierarchical clustering, Journal of Classification, Springer New York, Vol. 7, No. 2, pp. 197-216, 1990.
[41] Lapointe F.J. and Legendre P., The generation of random ultrametric matrices representing dendrograms. Journal of Classification, Springer New York, Vol. 8, No. 2, pp 177-200, 1991.
[42] Jain A.K. and Dubes R.C., Algorithms for Clustering Data. Prentice Hall, 1988.
[43] Banfield C.F., "Ultrametric Distances for a Single Linkage Dendrogram", JSTOR: Applied Statistics, Statistical Algorithms, Vol. 25, No. 3, pp. 313-315, 1976.
[44] Duda R.O., Hart P.E., and Stork D.G., Pattern Classification, second ed. Wiley, 2001.
[45] Kaufman L. and Rosseeuw P.J., Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc, 1990.
[46] Man Y. and Gath I., Detection and Separation of Ring-Shaped Clusters Using Fuzzy Clusters, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 8, pp. 855-861, 1994.
[47] Minaei-Bidgoli B., Topchy A. and Punch W.F., Ensembles of Partitions via Data Resampling, in Proc. Intl. Conf. on Information Technology, ITCC 04, Las Vegas, 2004.
Analysis and Machine Intelligence, VOL. 30, NO. 1, 160-173, 2008.
[17] Fred A.L. and Jain A.K., Combining Multiple Clusterings Using Evidence Accumulation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 27(6):835–850, 2005.
[18] Kuncheva L.I. and Hadjitodorov S., Using diversity in cluster ensembles. In Proc. of IEEE Intl. Conference on Systems, Man and Cybernetics, pages 1214–1219, 2004.
[19] Fred A. and Jain A.K. , Learning Pairwise Similarity for Data Clustering, In Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06), 2006.
[20] Baumgartner R., Somorjai R., Summers R., Richter W., Ryner L., and Jarmasz M., Resampling as a Cluster Validation Technique in fMRI, JOURNAL OF MAGNETIC RESONANCE IMAGING 11: pp. 228–231, 2000.
[21] Law M.H.C., Topchy A.P., and Jain A.K., Multiobjective data clustering. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 424–430, Washington D.C, 2004.
[22] Shamiry O., Tishby N., Cluster Stability for Finite Samples, 21st Annual Conference on Neural Information Processing Systems (NIPS07), 2007.
[23] Lange T., Braun M.L., Roth V., and Buhmann J.M., Stability-based model selection. In Advances in Neural Information Processing Systems 15. MIT Press, 2003.
[24] Breckenridge J., Replicating cluster analysis: Method, consistency and validity, Multivariate Behavioral research, 1989.
[25] Fridlyand J. and Dudoit S., Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method. Stat. Berkeley Tech Report. No. 600, 2001.
[26] Levine E., Domany E., Resampling Method for Unsupervised Estimation of Cluster Validity. Neural Computation 13: 2573-2593, 2001.
[27] Roth V., Lange T., Braun M., and Buhmann J., A Resampling Approach to Cluster Validation, Intl. Conf. on Computational Statistics, COMPSTAT, 2002.
[28] Roth V., Braun M.L., Lange T., and Buhmann J.M., Stability-Based Model Order Selection in Clustering with Applications to Gene Expression Data, ICANN 2002, LNCS 2415, pp. 607–612, 2002.
[29] Rakhlin A. and Caponnetto A., Stability of k-means clustering, In Advances in Neural Information Processing Systems 19, MIT Press, Cambridge, MA, 2007.
[30] Luxburg U.V. and Ben-David S., Towards a statistical theory of clustering, Technical report, PASCAL workshop on clustering, London, 2005.
[31] Roth V. and Lange T., Feature Selection in Clustering Problems, In Advances in Neural Information Processing Systems, NIPS04, 2004.

[48] Alizadeh H., Amirgholipour S.K., Seyedaghaee N.R. and Minaei-Bidgoli B., Nearest Cluster Ensemble (NCE): Clustering Ensemble Based Approach for Improving the performance of K-Nearest Neighbor Algorithm, 11th Conf. of the International Federation of Classification Societies, IFCS09, March 13–18, 2009.
[49] Mohammadi M., Alizadeh H. and Minaei-Bidgoli B., Neural Network Ensembles using Clustering Ensemble and Genetic Algorithm, Intl. Conf. on Convergence and hybrid Information Technology, ICCIT08, Nov. 11-13, IEEE CS, 2008.
[50] Barthelemy J.P. and Leclerc B., The median procedure for partition, In Partitioning Data Sets, AMS DIMACS Series in Discrete Mathematics, Cox, I. J. et al eds., 19, pp. 3-34, 1995.
[51] Fern, X. and Brodley, C. E., Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach, In Proc. 20th Int. conf. on Machine Learning, ICML 2003, 2003.
[52] Dudoit S. and Fridlyand, J., Bagging to improve the accuracy of a clustering procedure, Bioinformatics, 19 (9), pp. 1090-1099, 2003.
[53] Fischer B. and Buhmann J.M., Bagging for path-based clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, pp.1411–1415, 2003.
[54] Fred A. and Jain A.K., Robust data clustering, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR ,USA, vol. II, pp. 128–136, 2003.
[55] Kuncheva L.I. and Whitaker C.J., Measures of diversity in classifier ensembles, Machine Learning, 2003.
[56] Newman C.B.D.J., Hettich S. and Merz C., UCI repository of machine learning databases, http://www.ics.uci.edu/˜mlearn/MLSummary.html, 1998.

زیرنویس‌ها

[1] Data Mining
[2] Classification
[3] Classifier
[4] Unsupervised
[5] Tracking
[6] SubClass
[7] Genetic Algorithm (GA)
[8] Fitness
[9] Offspring
[10] Extreme
[11] Global Optimum
[12] Robustness
[13] Novelty
[14] Stability
[15] Flexibility
[16] Diversity
[17] Consensus Function
[18] Partitions
[19] Robust
[20] Pairwise
[21] Diversity
[22] Classification
[23] Easy
[24] Intermediate
[25] Hard
[26] Class
[27] Experimental Results
[28] Cluster Validity
[29] Jaccard Coefficient
[30] Support Vector Machine
[31] Outliers
[32] Nearest Neighbor Resampling
[33] Kernelized Validity Measure
[34] Full Ensemble
[35] Sum of Normalized Mutual Information
[36] Normalized Mutual Information
[37] Multiobjective
[38] Associated Cophenetic Matrix
[39] Minimum Spanning Tree
[40] Hierarchical
[41] Partitional
[42] Initialization
[43] Features
[44] Resampling
[45] Supervisor
[46] Train
[47] Evidence Accumulation Clustering
[48] Hypergraph
[49] Hyperedges
[50] Minimum-cut
[51] Heuristic
[52] Cluster-based Similarity Partitioning Algorithm
[53] Hyper-Graph Partitioning Algorithm
[54] Meta-CLustering Algorithm
[55] METIS
[56] Offline
[57] Goodness
[58] Perturbation
[59] Mutual Information (MI)
[60] Single Linkage(SL) or Average Linkage(AL)
[61] Single Linkage
[62] Intersection to Union
[63] Single Linkage
[64] Full Ensemble
[65] Boosting
[66] Bagging

[DOR: 20.1001.1.26765810.1395.13.2.13.9]
Journal of Iranian Association of Electrical and Electronics Engineers Vol13 No.2 Summer 2016