

خوشه‌بندی ترکیبی مبتنی بر زیرمجموعه‌ای از

خوشه‌های اولیه

حسین علیزاده، محسن مشکی، حمید پروین و بهروز مینایی بیدگلی
دانشکده مهندسی کامپیوتر، دانشگاه علم و صنعت ایران

چکیده:

اکثر مطالعات اخیر در حوزه خوشه‌بندی ترکیبی سعی می‌کنند ابتدا خوشه‌بندی‌های اولیه‌ای تولید کنند که تا حد ممکن دارای پراکندگی باشند، سپس با اعمال یک تابع توافقی همه این نتایج را با هم ترکیب می‌کنند. در این مقاله یک روش جدید خوشه‌بندی ترکیبی ارائه شده است که در آن به جای استفاده از تمام نتایج اولیه، تنها از زیرمجموعه‌ای از خوشه‌های اولیه استفاده می‌شود. ایده اصلی در این روش استفاده از خوشه‌های پایدار در ترکیب نهایی است. برای ترکیب خوشه‌های انتخابی، از تابع توافقی مبتنی بر ماتریس همبستگی استفاده شده است. از آن جایی که ساخت ماتریس همبستگی با در دسترس بودن تنها تعدادی از خوشه‌ها، با روش‌های موجود امکان‌پذیر نمی‌باشد، در این مقاله یک روش جدید به نام خوشه‌بندی انباشت مدارک توسعه یافته، برای ساخت ماتریس همبستگی از زیرمجموعه‌ای از خوشه‌ها پیشنهاد شده است. برای ارزیابی خوشه‌ها، از پایداری مبتنی بر اطلاعات متقابل استفاده شده است. نتایج تجربی روی چندین مجموعه داده استاندارد نشان می‌دهد که روش پیشنهادی به‌طور مؤثری نتایج خوشه‌بندی‌های اولیه را بهبود می‌دهد. همچنین، مقایسه نتایج در مقایسه با سایر روش‌های خوشه‌بندی ترکیبی، نشان از کارایی بالای روش پیشنهادی دارد.

واژگان کلیدی: خوشه‌بندی ترکیبی، پایداری خوشه، اطلاعات متقابل، ماتریس همبستگی.

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 گیرند و با نمونه‌های گروه‌های دیگر حداکثر تفاوت را داشته باشند (Faceli, et al., 2006, Jain, et al., 1999). از آن جا که اکثر روش‌های خوشه‌بندی پایه روی جنبه‌های خاصی از داده‌ها تأکید می‌کنند، در نتیجه روی مجموعه داده‌های خاصی کارآمد می‌باشند. به همین دلیل، نیاز به روش‌هایی است 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 (Strehl and Ghosh, 2002, Fred and Lourenco, 2008, Fred and Jain, 2002).

تحقیقات اخیر در این زمینه نشان داده‌اند که خوشه‌بندی داده‌ها می‌تواند به‌طور چشم‌گیری از ترکیب چندین افراز داده سود ببرد. خوشه‌بندی ترکیبی می‌تواند جواب‌های بهتری از نظر استحکام^۱، نو بودن^۲، پایداری^۳ و انعطاف‌پذیری^۴ نسبت به روش‌های پایه ارائه دهد (Fred and Lourenco, Strehl and Ghosh, 2002, Topchy, et al., 2003, Ayad and Kamel, 2008).

۱-۱- ایجاد پراکندگی در خوشه‌بندی ترکیبی

به‌طور خلاصه خوشه‌بندی ترکیبی شامل دو مرحله اصلی (الف) تولید نتایج اولیه و (ب) ترکیب نتایج برای استخراج خوشه‌های نهایی می‌باشد (Strehl and Ghosh, 2002). در مرحله اول تعدادی خوشه‌بندی اولیه ایجاد می‌شود که هر

¹ Robustness

² Novelty

³ Stability

⁴ Flexibility

۲- کارهای مرتبط

روش‌های خوشه‌بندی ترکیبی سعی می‌کنند تا با ترکیب افزایش‌های^۱ مختلف تولید شده از روش‌های خوشه‌بندی پایه، یک افزایش مستحکم^۲ از داده‌ها را تولید کنند (Strehl and Fred and Jain, Ayad and Kamel, 2008, Ghosh, 2002, 2005, Kuncheva and Hadjitodorov, 2004). در اکثر مطالعات اخیر، همه افزایش‌ها با وزن برابر در ترکیب نهایی حاضر می‌شوند و همه خوشه‌های موجود در همه افزایش‌ها نیز با وزن برابر در ترکیب نهایی شرکت می‌کنند (Fred and Jain, 2006). استرل و گاش در (Strehl and Ghosh, 2002) یک معیار برای انتخاب از میان ترکیبات ممکن از پایه کرده‌اند که مبتنی بر کیفیت کلی یک خوشه‌بندی است. برای این کار، آنها میزان ثبات بین افزایش ترکیبی و افزایش پایه را در نظر گرفته‌اند و با استفاده از یک قاعده ترکیبی ثابت، یک معیار شباهت دو به دو^۳ را روی فضای ویژگی‌های d -بعدی به کار برده‌اند.

در اکثر الگوریتم‌های پایه برای خوشه‌بندی ترکیبی از نمونه‌برداری داده‌ها استفاده می‌شود. مسئله اصلی در این روش‌ها چگونگی ارزیابی خوشه و خوشه‌بندی (افراز) است. بامگارتنر و همکاران در (Baumgartner, et al., 2000) یک روش مبتنی بر بازنمونه‌برداری را برای بررسی اعتبارسنجی نتایج خوشه‌بندی فازی ارائه کرده‌اند. در چند سال اخیر، پایداری خوشه به‌عنوان یک معیار ارزیابی خوشه مورد توجه زیادی قرار گرفته است (Fred and Jain, 2006, Law, et al., 2004, Lange, et al., 2003). ایده‌های اولیه برای اعتبارسنجی خوشه با استفاده از بازنمونه‌برداری در (Breckenridge, 1989) ارائه شده و بعدها در (Fridlyand and Dudoit, 2001, Levine, and Domany, 2001) کامل‌تر شده است. راس و همکاران نیز در (b2002, a2002) یک روش مبتنی بر بازنمونه‌برداری برای اعتبارسنجی خوشه ارائه کرده‌اند. عنصر اصلی در این روش، که در واقع کامل‌شده روش‌های پیشین می‌باشد، پایداری خوشه است. معیار پایداری، میزان همبستگی افزایش‌های به‌دست آمده از دو نمونه‌برداری مستقل از مجموعه داده را اندازه‌گیری می‌کند. هر چه میزان پایداری برای یک خوشه‌بندی بیشتر باشد، به این معنی است که اگر الگوریتم خوشه‌بندی چندین مرتبه دیگر روی آن نمونه‌ها به کار رود، نتایج مشابهی حاصل می‌شود (Rakhlin and Luxburg and Ben-David, 2005, Caponnetto, 2007). هم‌چنین، (Roth and Lange, 2003) یک الگوریتم جدید

برای خوشه‌بندی ارائه کرده‌اند که مبتنی بر انتخاب ویژگی می‌باشد. در این روش از معیار پایداری مبتنی بر بازنمونه‌برداری داده‌ها، برای انتخاب پارامترهای الگوریتم خوشه‌بندی استفاده شده است. چندین روش اعتبارسنجی خوشه^۴ مبتنی بر ایده استفاده از پایداری پیشنهاد شده است (Lange, et al., 2004). (Ben-Hur, et al., 2002) نیز روشی برای محاسبه پایداری ارائه کرده‌اند که بر مبنای شباهت بین نمونه‌ها در خوشه‌بندی‌های متفاوت عمل می‌کند. در این روش، ابتدا ماتریس همبستگی با استفاده از روش بازنمونه‌برداری به‌دست می‌آید. سپس ضریب جاکارد^۵ به‌عنوان معیار پایداری بر اساس این ماتریس محاسبه می‌شود. هم‌چنین، (Estivill-Castro and Yang, 2003) روشی برای ارزیابی افزایش‌های نهایی خوشه‌بندی ارائه کرده‌اند که از ماشین بردار پشتیبان^۶ استفاده می‌کند. این روش با شناسایی نویزها و داده‌های دورافتاده^۷ به نتایج خوشه‌بندی دارای استحکام دست یافته است. (Moller and Radke, 2006) از بازنمونه‌برداری نزدیک‌ترین همسایه (NNR^۸) برای اعتبارسنجی نتایج خوشه‌بندی استفاده کرده‌اند. این روش بازنمونه‌برداری اولین بار در تحلیل سری‌های زمانی به کار رفته است (Brandsma and Buishand, 1998). (Inokuchi, et al., 2006) معیار اعتبارسنجی هسته - محور^۹ پیشنهاد کرده‌اند. هسته در این روش به معنی تابع مرکزی استفاده شده در ماشین بردار پشتیبان می‌باشد. در این روش، دو شاخص مورد توجه قرار گرفته است: اولی مجموع مقادیر کواریانس فازی داخل خوشه‌هاست و دومی شاخص مبتنی بر هسته ژی - بن (ژی و بنی ۱۹۹۱). در این روش از این دو شاخص برای ارزیابی نتایج خوشه‌بندی و هم‌چنین، تعیین تعداد خوشه‌ها با مرزهای غیر خطی استفاده شده است. (Das and Sil, 2007) روشی برای تعیین تعداد خوشه‌ها ارائه کرده‌اند که از اعتبارسنجی خوشه‌ها برای تقسیم و ادغام آن‌ها بهره می‌برد. (Fern and Lin, 2008) روشی برای خوشه‌بندی ترکیبی پیشنهاد کرده‌اند که از زیرمجموعه‌ی مؤثرتری از افزایش‌های اولیه در ترکیب نهایی استفاده می‌کند. در این روش اگر چه تعداد اعضای شرکت کننده در ترکیب نهایی کمتر از یک خوشه‌بندی ترکیبی کامل^{۱۰} است، به دلیل انتخاب افزایش‌های با کارایی بالاتر، نتایج نهایی بهبود می‌یابند. پارامترهایی که در این روش مورد توجه قرار گرفته‌اند، عبارتند از: کیفیت و

⁴ Cluster Validity

⁵ Jaccard Coefficient

⁶ Support Vector Machine

⁷ Outliers

⁸ Nearest Neighbor Resampling

⁹ Kernelized Validity Measure

¹⁰ Full Ensemble

¹ Partitions

² Robust

³ Pairwise

۲- کارهای مرتبط

روش‌های خوشه‌بندی ترکیبی سعی می‌کنند تا با ترکیب افزای‌های^۱ مختلف تولید شده از روش‌های خوشه‌بندی پایه، یک افزای مستحکم^۲ از داده‌ها را تولید کنند (Strehl and Fred and Jain, Ayad and Kamel, 2008, Ghosh, 2002, 2005, Kuncheva and Hadjitodorov, 2004). در اکثر مطالعات اخیر، همه افزای‌ها با وزن برابر در ترکیب نهایی حاضر می‌شوند و همه خوشه‌های موجود در همه افزای‌ها نیز با وزن برابر در ترکیب نهایی شرکت می‌کنند (Fred and Jain, 2006). استرل و گاش در (Strehl and Ghosh, 2002) یک معیار برای انتخاب از میان ترکیبات ممکن از پایه کرده‌اند که مبتنی بر کیفیت کلی یک خوشه‌بندی است. برای این کار، آنها میزان ثبات بین افزای ترکیبی و افزای پایه را در نظر گرفته‌اند و با استفاده از یک قاعده ترکیبی ثابت، یک معیار شباهت دو به دو^۳ را روی فضای ویژگی‌های d -بعدی به کار برده‌اند.

در اکثر الگوریتم‌های پایه برای خوشه‌بندی ترکیبی از نمونه‌برداری داده‌ها استفاده می‌شود. مسئله اصلی در این روش‌ها چگونگی ارزیابی خوشه و خوشه‌بندی (افزای) است. بامگارتنر و همکاران در (Baumgartner, et al., 2000) یک روش مبتنی بر بازنمونه‌برداری را برای بررسی اعتبارسنجی نتایج خوشه‌بندی فازی ارائه کرده‌اند. در چند سال اخیر، پایداری خوشه به‌عنوان یک معیار ارزیابی خوشه مورد توجه زیادی قرار گرفته است (Fred and Jain, 2006, Law, et al., 2004, Lange, et al., 2003). ایده‌های اولیه برای اعتبارسنجی خوشه با استفاده از بازنمونه‌برداری در (Breckenridge, 1989) ارائه شده و بعدها در (Fridlyand and Dudoit, 2001, Levine, and Domany, 2001) کامل‌تر شده است. راس و همکاران نیز در (b2002, a2002) یک روش مبتنی بر بازنمونه‌برداری برای اعتبارسنجی خوشه ارائه کرده‌اند. عنصر اصلی در این روش، که در واقع کامل‌شده روش‌های پیشین می‌باشد، پایداری خوشه است. معیار پایداری، میزان همبستگی افزای‌های به‌دست آمده از دو نمونه‌برداری مستقل از مجموعه داده را اندازه‌گیری می‌کند. هر چه میزان پایداری برای یک خوشه‌بندی بیشتر باشد، به این معنی است که اگر الگوریتم خوشه‌بندی چندین مرتبه دیگر روی آن نمونه‌ها به کار رود، نتایج مشابهی حاصل می‌شود (Rakhlin and Luxburg and Ben-David, 2005, Caponnetto, 2007). همچنین، (Roth and Lange, 2003) یک الگوریتم جدید

برای خوشه‌بندی ارائه کرده‌اند که مبتنی بر انتخاب ویژگی می‌باشد. در این روش از معیار پایداری مبتنی بر بازنمونه‌برداری داده‌ها، برای انتخاب پارامترهای الگوریتم خوشه‌بندی استفاده شده است. چندین روش اعتبارسنجی خوشه^۴ مبتنی بر ایده استفاده از پایداری پیشنهاد شده است (Lange, et al., 2004). (Ben-Hur, et al., 2002) نیز روشی برای محاسبه پایداری ارائه کرده‌اند که بر مبنای شباهت بین نمونه‌ها در خوشه‌بندی‌های متفاوت عمل می‌کند. در این روش، ابتدا ماتریس همبستگی با استفاده از روش بازنمونه‌برداری به‌دست می‌آید. سپس ضریب جاکارد^۵ به‌عنوان معیار پایداری بر اساس این ماتریس محاسبه می‌شود. هم‌چنین، (Estivill-Castro and Yang, 2003) روشی برای ارزیابی افزای‌های نهایی خوشه‌بندی ارائه کرده‌اند که از ماشین بردار پشتیبان^۶ استفاده می‌کند. این روش با شناسایی نویزها و داده‌های دورافتاده^۷ به نتایج خوشه‌بندی دارای استحکام دست یافته است. (Moller and Radke, 2006) از بازنمونه‌برداری نزدیک‌ترین همسایه (NNR^۸) برای اعتبارسنجی نتایج خوشه‌بندی استفاده کرده‌اند. این روش بازنمونه‌برداری اولین بار در تحلیل سری‌های زمانی به کار رفته است (Brandsma and Buishand, 1998). (Inokuchi, et al., 2006) معیار اعتبارسنجی هسته - محور^۹ پیشنهاد کرده‌اند. هسته در این روش به معنی تابع مرکزی استفاده شده در ماشین بردار پشتیبان می‌باشد. در این روش، دو شاخص مورد توجه قرار گرفته است: اولی مجموع مقادیر کواریانس فازی داخل خوشه‌هاست و دومی شاخص مبتنی بر هسته ژی - بن (ژی و بنی (۱۹۹۱)). در این روش از این دو شاخص برای ارزیابی نتایج خوشه‌بندی و هم‌چنین، تعیین تعداد خوشه‌ها با مرزهای غیر خطی استفاده شده است. (Das and Sil, 2007) روشی برای تعیین تعداد خوشه‌ها ارائه کرده‌اند که از اعتبارسنجی خوشه‌ها برای تقسیم و ادغام آن‌ها بهره می‌برد. (Fern and Lin, 2008) روشی برای خوشه‌بندی ترکیبی پیشنهاد کرده‌اند که از زیرمجموعه‌ی مؤثرتری از افزای‌های اولیه در ترکیب نهایی استفاده می‌کند. در این روش اگر چه تعداد اعضای شرکت کننده در ترکیب نهایی کمتر از یک خوشه‌بندی ترکیبی کامل^{۱۰} است، به دلیل انتخاب افزای‌های با کارایی بالاتر، نتایج نهایی بهبود می‌یابد. پارامترهایی که در این روش مورد توجه قرار گرفته‌اند، عبارتند از: کیفیت و

⁴ Cluster Validity

⁵ Jaccard Coefficient

⁶ Support Vector Machine

⁷ Outliers

⁸ Nearest Neighbor Resampling

⁹ Kernelized Validity Measure

¹⁰ Full Ensemble

سال ۱۳۸۹ شماره ۱ پیاپی ۱۳

¹ Partitions

² Robust

³ Pairwise

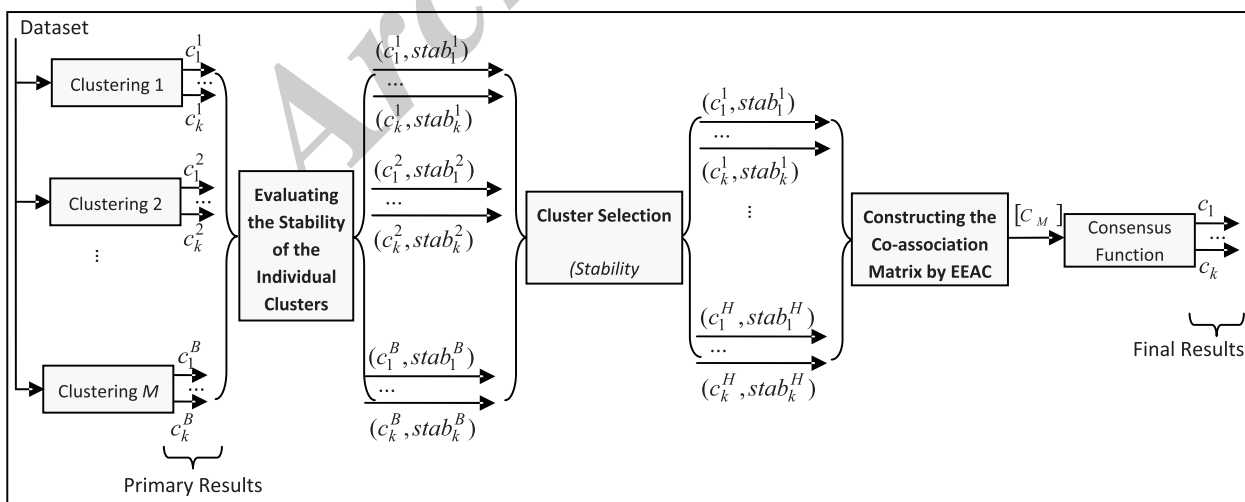
است. از آنجایی که به نظر می‌رسد تمام خوشه‌های حاصل از الگوریتم‌های خوشه‌بندی اولیه از پایداری بالایی برخوردار نباشند، در این روش تنها از خوشه‌های با مقدار پایداری بالا در ماتریس همبستگی و در نتیجه در ترکیب نهایی استفاده شده است. انتخاب خوشه‌ها بر اساس معیار پایداری خوشه مبتنی بر اطلاعات متقابل هنجارسازی شده (NMI) صورت می‌گیرد. نمای کلی از روش پیشنهادی در (شکل ۲) نشان داده شده است.

در این روش ابتدا با استفاده از روش‌های ایجاد پراکندگی تعداد B خوشه‌بندی اولیه انجام می‌شود. این کار می‌تواند با 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 و الگوریتم‌های سلسله‌مراتبی 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. در مرحله بعد خوشه‌های به دست آمده، مورد ارزیابی قرار می‌گیرند تا کیفیت هر خوشه مشخص شود. برای ارزیابی خوشه از معیار پایداری خوشه استفاده شده است. چگونگی ارزیابی خوشه‌ها در بخش بعدی به‌طور مفصل تشریح شده است. پس از این که پایداری هر خوشه محاسبه شد، در گام بعد، عمل انتخاب خوشه‌ها با توجه به مقدار پایداری خوشه انجام می‌شود. برای این کار خوشه‌هایی که از مقدار آستانه th فراتر باشند برای شرکت در خوشه‌بندی نهایی انتخاب

پراکندگی. این روش سعی می‌کند تا زیرمجموعه‌ای از افزایندهایی از نتایج اولیه را وارد ترکیب نهایی کند که بالاترین میزان کیفیت برخوردار بوده و در عین حال نسبت به هم بیشترین پراکندگی را دارا باشند. در این روش از معیار مجموع اطلاعات متقابل هنجارسازی شده (SNMI¹) (برای یک افزاز در مقایسه با افزازهای دیگر ترکیب) برای اندازه‌گیری کیفیت یک افزاز استفاده شده است. هم‌چنین، معیار اطلاعات متقابل هنجارسازی شده (NMI²) (بین تمام افزازهای موجود در ترکیب) برای اندازه‌گیری پراکندگی لازم برای ترکیب به کار رفته است. (Fern and Lin, 2008) نشان می‌دهند که روش آن‌ها نسبت به خوشه‌بندی ترکیبی کامل و یا روش انتخاب تصادفی از کارآیی بهتری برخوردار است. (Law, et al., 2004) یک روش خوشه‌بندی چندهدفی³ ارائه کرده‌اند که مبتنی بر انتخاب خوشه‌های اولیه تولید شده توسط الگوریتم‌های مختلف خوشه‌بندی، در طی یک روال بهینه‌سازی می‌باشد. در این روش، بهترین مجموعه از توابع هدف برای بخش‌های مختلف از فضای ویژگی انتخاب شده است. (Fred and Jain, 2006) یک روش خوشه‌بندی ترکیبی ارائه کرده‌اند که در آن با استفاده از معیار پایداری خوشه، شباهت دو به دو آموزش داده می‌شود. در این روش، به جای استفاده از معیارهای ارزیابی مبتنی بر افزاز نهایی، افزازهای حاصل از الگوریتم‌های پایه در نواحی مختلف از فضای ویژگی d -بعدی مورد ارزیابی قرار گرفته است.

۳- روش پیشنهادی

ایده اصلی در این روش استفاده از زیرمجموعه‌ای از خوشه‌های اولیه به جای کل خوشه‌ها در خوشه‌بندی ترکیبی



(شکل ۲) روال الگوریتم پیشنهادی برای خوشه‌بندی ترکیبی

می‌شوند. در گام بعدی خوشه‌های انتخاب شده با هم ترکیب شده و خوشه‌های نهایی از آن‌ها به دست می‌آید. روش‌های

¹ Sum of Normalized Mutual Information
² Normalized Mutual Information
³ Multiobjective

یعنی این دو خوشه در خوشه‌بندی Z -ام از کیفیت برابری برخوردارند.

یکی از معیارهایی که می‌تواند به‌عنوان تابع برازندگی خوشه در نظر گرفته شود، معیار پایداری خوشه (Lange, et al., 2003) است. پایداری خوشه، اثر آشفتگی^۴ 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 با جای‌گذاری و یا بدون جای‌گذاری انجام شود.

در این روش خوشه‌های پایدارتر از خوشه‌بندی‌های اولیه شناسایی می‌شوند و ماتریس همبستگی نهایی، تنها از این خوشه‌های پایدار تشکیل می‌شود. یک خوشه پایدار، خوشه‌ای است که اگر آن روش خوشه‌بندی را چند بار دیگر هم، روی آن مجموعه داده (یا روی مجموعه‌های مختلف حاصل از نمونه‌برداری از آن مجموعه داده) اجرا کنیم، با احتمال زیاد این خوشه باز هم دیده خواهد شد. به‌عبارت دیگر، خوشه‌های پایدار به خوشه‌هایی اطلاق می‌شود که در خوشه‌بندی‌های مختلف روی زیرمجموعه‌های به‌دست آمده از نمونه‌برداری‌های مختلف بیشترین تکرار را داشته باشند. یعنی با تغییرات جزئی در مجموعه داده، آن خوشه‌ها باز هم تکرار شوند. برای شناسایی خوشه‌های پایدارتر نیاز به سازوکارهایی است تا بتواند پایداری را برای هر خوشه از یک خوشه‌بندی، مستقل از خوشه‌های دیگر به‌دست 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 سؤال 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 $\text{sim}(C_i, P(D))$ 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، به‌عنوان میزان پایداری این خوشه $g_i(C_i, D)$ برگردانده می‌شود. در واقع $\text{sim}(C_i, P(D))$ میزان اعتبار خوشه C_i را در خوشه‌بندی P روی مجموعه داده D مشخص می‌کند. الگوریتم این روال در (شکل ۳) نشان داده شده است.

مختلفی برای ترکیب خوشه‌بندی‌های اولیه و به‌دست آوردن خوشه‌های نهایی وجود دارد. تفاوتی که در این جا وجود دارد این است که در این روش ممکن است که از هر خوشه‌بندی اولیه، تنها تعدادی از خوشه‌ها موجود باشند. از آنجایی که استفاده از روش انباشت مدارک (EAC) نمی‌تواند شباهت بین جفت نمونه‌ها را در حضور تنها تعدادی از خوشه‌ها به‌درستی تشخیص دهد، در این مقاله یک روش جدید برای ترکیب نتایج پیشنهاد شده است که آن را انباشت مدارک توسعه‌یافته (EEAC)^۱ نامیده‌ایم. این روش که توانایی استخراج ماتریس همبستگی برای نمونه‌ها - در شرایطی که تنها تعدادی از خوشه‌ها موجود هستند- را دارد، 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^۲، خوشه‌های نهایی را استخراج کرد.

۳-۱- ارزیابی خوشه

از آنجایی که میزان برازندگی^۳ یک خوشه در میان کل نقاط داده معنی‌دار است، تابع برازندگی خوشه یعنی $g_i(C_i, D)$ 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 (Law, et al., 2004):

- باید با تابع f_i که توسط الگوریتم خوشه‌بندی خاص A_i بهینه می‌شود، ارتباط منطقی داشته باشد. به عبارت دیگر، مقدار بیشتر برای $g_i(C_i, D)$ به این معنی باشد که خوشه C_i نسبت به تابع f_i و به تناظر نسبت به الگوریتم خوشه‌بندی خاص A_i ، برانزده‌تر (بهینه‌تر) است.
- باید نسبت به توابع خوشه‌بندی مختلف قابل مقایسه باشد. به عبارت دیگر، اگر $g_j(C_j, D) > g_i(C_i, D)$ ، آن‌گاه باید نتیجه گرفت که کیفیت خوشه C_j با توجه به تابع f_j از تابع f_i بهتر است. یعنی کیفیت خوشه C_i در خوشه‌بندی Z -ام از کیفیت خوشه C_j در خوشه‌بندی Z -ام بهتر می‌باشد.
- در نهایت مقدار تابع برازندگی خوشه باید نسبت به خوشه‌های مختلف قابل مقایسه باشد. به عبارت دیگر، $g_j(C_j, D) = g_i(C_i, D)$ باید نتیجه دهد که خوشه‌های C_i و C_j نسبت به تابع f_j میزان برازندگی برابری دارند.

¹ Extended EAC (EEAC)

² Single Linkage (SL) or Average Linkage (AL)

³ Goodness

⁴ Perturbation

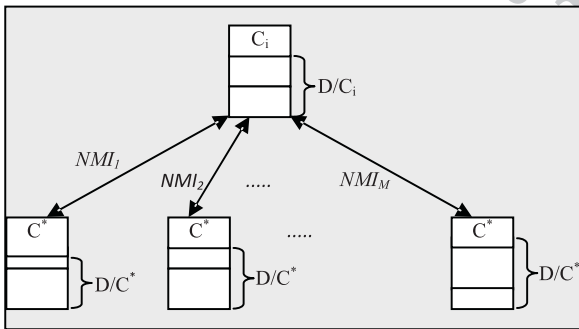
که اطلاعات متقابل، $MI(P_1, P_2)$ از (رابطه ۲) به دست می‌آید.

$$MI(P_1, P_2) = \sum_{i=0}^1 \sum_{j=0}^1 \frac{r_{ij}}{m^2} \log \frac{mp_{ij}}{r_{ij}} \quad (2)$$

$$r_{ij} = P_{i,j}, \quad P_{i.} = P_{i0} + P_{i1}, \quad P_{.j} = P_{0j} + P_{1j}$$

که در این رابطه، p_{11} نشان‌دهنده تعداد نمونه‌های مشترک موجود در C_i و C^* است. p_{10} نشان‌دهنده تعداد نمونه‌های مشترک موجود در D/C^* و D/C_i است. p_{01} نشان‌دهنده تعداد نمونه‌های مشترک موجود در D/C_i و C^* است. p_{00} نشان‌دهنده تعداد نمونه‌های مشترک موجود در D/C_i و D/C^* 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 ۴) نمای کلی از این روش محاسبه پایداری خوشه را نشان می‌دهد.

با توجه به (شکل ۴)، 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 ۳) در $\text{sim}(C_i, P(D))$ و سپس در $\text{score}[i]$ ذخیره می‌گردد. پایداری کل از میانگین کل این پایداری‌ها تشکیل می‌شود.



(شکل ۴) محاسبه پایداری خوشه C_i با روش مبتنی بر NMI

$$\text{Stability}(C_i) = \frac{1}{M} \sum_{i=1}^M NMI_i \quad (3)$$

که M تعداد خوشه‌بندی‌ها در مجموعه مرجع می‌باشد. این روش، در واقع خوشه‌هایی را که بیشترین تکرار را در خوشه‌بندی‌های مختلف دارند، به‌عنوان خوشه‌های پایدارتر معرفی می‌کند. پس از این مرحله خوشه‌های بالاترین مقدار پایداری انتخاب شده و وارد مرحله بعد می‌شوند تا خوشه‌بندی نهایی را شکل دهند. پارامتر th این مقدار آستانه را مشخص می‌کند.

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 $\text{score}[l] = \text{sim}(C_i, P(D))$;
End
 $g_i(C_i, D) := \text{average of score}[l]$;

(شکل ۳) الگوریتم محاسبه پایداری خوشه C_i به‌عنوان تابع برازندگی

برای محاسبه $\text{sim}(C_i, P(D))$ که میزان شباهت بین خوشه C_i و نتیجه خوشه‌بندی $P(D)$ است، به‌صورت زیر عمل می‌شود. ابتدا تمام نمونه‌های دیگر متعلق به مجموعه داده D که در خوشه C_i قرار ندارند، به‌صورت 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 $P_2 = \{C^*, D/C^*\}$ می‌نامیم. حال از اطلاعات متقابل^۱ هنجارسازی شده (NMI) (Fred and Jain, 2002 Strehl and Ghosh, 2002) (Dudoit and Fridlyand, 2003) که معیار متداول برای ارزیابی شباهت بین دو افراز (نتیجه خوشه‌بندی) است، برای اندازه‌گیری شباهت بین دو خوشه‌بندی P_1 و P_2 استفاده می‌شود. از آن جایی که معیار اطلاعات متقابل هنجارسازی شده (MI)، وابسته به اندازه خوشه‌هاست، به‌طور معمول از معیار NMI استفاده می‌شود. رابطه NMI بین دو خوشه‌بندی P_1 و P_2 به‌صورت زیر محاسبه می‌شود.

(۱)

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

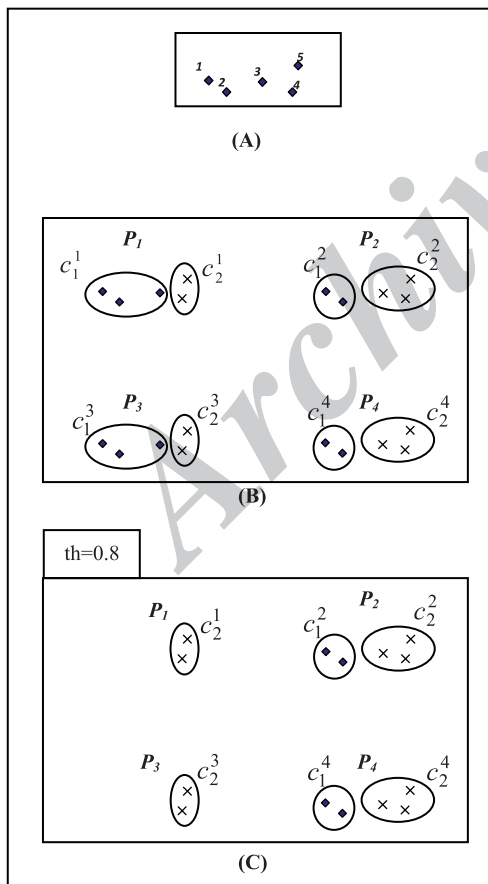
¹ Mutual Information (MI)

خوشه‌بندی‌های اولیه روی آن صورت گرفته است (شکل ۵ قسمت B). همچنین، فرض کنید که مقادیر پایداری خوشه‌های تولید شده به صورت زیر باشند:

$$\begin{aligned} \text{Stability}(c_2^1) &= \text{Stability}(c_2^3) = 1 \\ \text{Stability}(c_1^2) &= \text{Stability}(c_4^1) = 1 \\ \text{Stability}(c_2^2) &= \text{Stability}(c_2^4) = 0.82 \\ \text{Stability}(c_1^1) &= \text{Stability}(c_3^1) = 0.55 \end{aligned}$$

اگر مقدار آستانه برای انتخاب خوشه‌ها برابر با 0.8 باشد، خوشه‌های اول از افزایش اول و سوم حذف می‌شوند. اکنون ماتریس همبستگی باید با استفاده از بقیه خوشه‌ها درست شود. یعنی (شکل ۵، قسمت C).

$$\begin{aligned} C(1,2) &= \frac{2}{\max(2,2)} = \frac{2}{2} = 1 \\ C(1,3) &= C(2,3) = \frac{0}{\max(2,2)} = \frac{0}{2} = 0 \\ C(3,4) &= C(3,5) = \frac{2}{\max(2,4)} = \frac{2}{4} = 0.5 \\ C(4,5) &= \frac{4}{\max(4,4)} = \frac{4}{4} = 1 \end{aligned}$$



(شکل ۵) ساخت ماتریس همبستگی با روش EEAC (A) مجموعه داده شامل ۵ نمونه، (B) نتایج چهار خوشه‌بندی اولیه (C) خوشه‌های باقیمانده پس از آستانه‌گیری

۲-۳- ترکیب نتایج اولیه با روش انباشت مدارک توسعه یافته

در این مرحله، خوشه‌های انتخاب‌شده ماتریس همبستگی را تشکیل می‌دهند. 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 (EAC)، به صورت رابطه (۴) محاسبه می‌شود.

$$C(i, j) = \frac{n_{i,j}}{m_{i,j}} \quad (4)$$

که $n_{i,j}$ تعداد دفعاتی است که جفت‌نمونه‌های i و j با هم در یک خوشه گروه‌بندی شده‌اند و $m_{i,j}$ تعداد نمونه‌برداری‌هایی است که هر دوی این جفت‌نمونه‌ها به‌طور هم‌زمان در آن ظاهر شده‌اند.

از آن جایی که پس از آستانه‌گیری در روش پیشنهادی، 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 یک معیاری برای نشان دادن همبستگی نمونه‌ها تعریف شود که بتواند شباهت بین نمونه‌ها را با حضور تنها زیرمجموعه‌ای از خوشه‌های اولیه به درستی استخراج و محاسبه کند. ما روش خود را برای ساخت ماتریس همبستگی در شرایطی که تعدادی از خوشه‌ها حذف شده‌اند، انباشت مدارک توسعه یافته (EEAC) می‌نامیم. هر داده ورودی از ماتریس همبستگی در روش EEAC به صورت رابطه (۵) تعریف می‌شود.

$$C(i, j) = \frac{n_{i,j}}{\max(n_i, n_j)} \quad (5)$$

n_i تعداد دفعاتی است که نمونه i در خوشه‌های انتخاب شده ظاهر شده است. به‌طور 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 خوشه‌های انتخاب شده ظاهر شده‌اند. بدیهی است که با در نظر گرفتن تعداد خوشه‌های ثابت در خوشه‌بندی‌های اولیه همواره n_i و n_j کمتر از تعداد کل افزایش‌های اولیه و همچنین، تعداد کل خوشه‌های ممکن می‌باشد. یعنی $n_i, n_j \leq B \leq k \times B$.

برای روشن‌تر شدن بحث، مثال زیر را در نظر بگیرید. فرض کنید ۵ نمونه مطابق (شکل ۵)، (قسمت A) وجود دارند که چهار خوشه‌بندی اولیه P_1 تا P_4 به‌عنوان

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 باشند. (جدول ۱) اطلاعات مختصری از این مجموعه داده‌ها در اختیار می‌گذارد. برای اطلاعات بیشتر در مورد هر کدام از این مجموعه داده‌ها می‌توان به (Newman, et al., 1998) رجوع کرد.

جدول ۱) مجموعه داده‌ها

	Class	Features	Samples
Glass	6	9	214
Breast-C	2	9	683
Wine	3	13	178
Bupa	2	6	345
Yeast	10	8	1484

نتایج آزمایش‌ها بر روی ویژگی‌های هنجارسازی شده از این مجموعه داده‌ها گزارش شده است. به عبارت دیگر هر کدام از ویژگی‌های این مجموعه داده‌ها با میانگین صفر و واریانس یک، $N(0,1)$ ، هنجارسازی شده‌اند.

۴-۲- نتایج آزمایش‌ها

روش پیشنهادی در محیط MATLAB (ver 7.1) پیاده‌سازی و مورد آزمایش قرار گرفته است. نتایج آزمایش‌ها روی میانگین ده بار اجرای مستقل برنامه گزارش شده است. عملکرد روش‌های مختلف خوشه‌بندی با استفاده از فرآیند بازبرچسب‌گذاری^۱ بین خوشه‌های به‌دست آمده و کلاس‌های واقعی و مقایسه آن‌ها محاسبه شده است. (جدول ۲) عملکرد روش‌های مختلف را در مقایسه با روش پیشنهادی نشان می‌دهد. مقادیر موجود در این جدول درصد خوشه‌بندی درست نمونه‌ها را بیان می‌کند. چهار سطر اول از این جدول نتایج عملکرد الگوریتم‌های پایه روی مجموعه داده‌های مختلف آمده است. نتایج آزمایش‌ها نشان می‌دهند که اگرچه هر کدام از این الگوریتم‌ها می‌توانند روی مجموعه داده خاصی نتایج قابل قبولی ارائه کنند، ولی روی مجموعه داده‌های دیگر کارایی پایینی از خود نشان می‌دهند. به‌عنوان مثال، الگوریتم K-means روی مجموعه Wine نتیجه^۲ تا اندازه‌ای خوب نسبت به روش‌های سلسله‌مراتبی اتصالی (Linkage) به‌دست می‌آورد. اما روی مجموعه داده‌های Bupa از بقیه ضعیف‌تر عمل می‌کند. هم‌چنین، روش اتصالی

ماتریس همبستگی قبل و بعد از آستانه‌گیری به ترتیب، به صورت روابط (۶) و (۷) خواهد بود:

$$C_{before} = \begin{bmatrix} 1 & 1 & 0.5 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 \end{bmatrix} \quad (6)$$

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، به‌گونه‌ای که خوشه پایدارتر دارای وزن بیشتری در مقادیر همبستگی بین نمونه‌هایش شود، می‌توان به عملکرد بهتر خوشه‌بندی ترکیبی امیدوار بود.

$$C_{after} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 \end{bmatrix} \quad (7)$$

با مشاهده این دو ماتریس می‌توان دریافت که چگونه حذف خوشه‌های ناپایدار می‌تواند باعث بهبود ماتریس همبستگی شود. از آنجایی 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 این که ماتریس همبستگی با روش EEAC ساخته شد، در مرحله بعد از یک تابع توافقی برای استخراج خوشه‌های نهایی از این ماتریس استفاده می‌شود. به‌طور معمول از یکی از الگوریتم‌های سلسله‌مراتبی برای استخراج خوشه‌های نهایی از ماتریس همبستگی استفاده می‌شود. در این مقاله از الگوریتم سلسله‌مراتبی اتصال منفرد^۱ استفاده شده است.

۴- نتایج تجربی

در این بخش نتایج به‌کارگیری روش پیشنهادی روی مجموعه داده‌های مختلف و پارامترهای مورد استفاده گزارش شده است.

۴-۱- مجموعه داده‌ها

روش پیشنهادی بر روی پنج مجموعه داده استاندارد مورد آزمایش قرار گرفته است. برای انجام آزمایش‌ها سعی شده

² Relabing

¹ Single Linkage



الگوریتم‌های سلسله‌مراتبی تحت شرایط ثابت همواره به جواب‌های به‌طور کامل مشابهی منجر می‌شود، از این الگوریتم‌ها به تعداد محدودی می‌توان در تولید نتایج اولیه استفاده کرد. در این آزمایش‌ها از پارامتر تعداد خوشه‌ها برای ایجاد پراکندگی در نتایج استفاده شده است. برای این کار، تعداد خوشه‌های از پیش تعیین شده برای الگوریتم‌های سلسله‌مراتبی مورد استفاده برابر با $k \pm 2$ انتخاب شده است که k تعداد واقعی خوشه‌ها می‌باشد. الگوریتم‌های مورد استفاده برای خوشه‌بندی سلسله‌مراتبی اولیه و فاصله مورد استفاده توسط هر کدام از این الگوریتم‌ها در (جدول ۳) آمده است (Duda, et al., 2001).

سطر آخر از (جدول ۲) نتایج روش پیشنهادی را روی این مجموعه داده‌ها نشان می‌دهد. نتایج اولیه برای روش پیشنهادی به‌طور دقیق همان نتایج روش ترکیب کامل (سطر دوم از روش‌های ترکیبی) می‌باشد. تفاوت روش پیشنهادی با روش ترکیب کامل در تعداد خوشه‌های اولیه‌ای است که در ترکیب نهایی شرکت می‌کنند. در واقع در روش پیشنهادی تنها خوشه‌های با مقدار پایداری بالا در ماتریس همبستگی و در نتیجه در ترکیب نهایی ظاهر شده‌اند. مقدار آستانه برای حضور یا عدم حضور یک خوشه در ترکیب نهایی به‌صورت تطبیقی^۲ انتخاب شده است. به این معنی که ابتدا مقدار $0/95$ به‌عنوان مقدار اولیه انتخاب شده است. اگر خوشه‌های انتخاب شده با این مقدار آستانه، کمتر از 90% داده‌ها را شامل شوند، مقدار آستانه به اندازه $0/05$ کاهش می‌یابد. این کار تا جایی ادامه می‌یابد که بیش از 90% از داده‌ها توسط خوشه‌های انتخاب شده خوشه‌بندی شده باشند. نتایج آزمایش‌ها نشان‌دهنده کارایی روش پیشنهادی نسبت به سایر روش‌های پایه و همچنین ترکیبی است.

(جدول ۳) الگوریتم‌های سلسله‌مراتبی مورد استفاده

Linkage Methods	Used Distance Measure
Single Linkage	کم‌ترین فاصله بین نمونه‌ها
Average Linkage	بیش‌ترین فاصله بین نمونه‌ها
Complete Linkage	میانگین فاصله بین همه جفت نمونه‌ها در دو خوشه
Weighted Linkage	میانگین فاصله وزن‌دار
Ward Linkage	مجموع افزایشی مربعات به‌عنوان نتیجه اتصال دو خوشه

² Adaptive

کامل (Complete-Linkage) روی مجموعه داده Breast-Cancer نیز به همین شکل عمل می‌کند.

یکی از دلایل این امر شکل‌های متنوع مجموعه‌های داده است که با توجه به این که هر کدام از روش‌های خوشه‌بندی پایه برای شکل‌های خاصی از داده‌ها طراحی شده‌اند، کارایی آن‌ها نیز روی داده‌های مختلف متفاوت است. به‌عنوان مثال الگوریتم K-means روی مجموعه داده Breast-Cancer به خوبی عمل می‌کند، اما روی مجموعه داده Bupa کارایی ضعیف‌تری از خود نشان می‌دهد (سطر اول از جدول ۲) که این نتایج نشان می‌دهند که شکل داده‌ها Breast-Cancer به‌نسبه کروی است در حالی که در Bupa با داده‌های غیر کروی مواجه هستیم. نتایج الگوریتم‌های Linkage روی این مجموعه داده‌ها (سطرهای دوم و سوم از جدول ۲) نیز صحت این گفته را تأیید می‌کند. سه سطر آخر از (جدول ۲) عملکرد روش‌های ترکیبی را در برابر مجموعه داده‌های مختلف نشان می‌دهد. با یک نگاه کلی به نتایج این سه سطر در مقایسه با چهار سطر فوق، بر این ادعا که "روش‌های ترکیبی منجر به تولید نتایج مستحکم‌تری نسبت به روش‌های غیر ترکیبی می‌شوند" صحت می‌گذارد. در اولین سطر از روش‌های ترکیبی، نتیجه ترکیب یکصد بار K-means با استفاده از روش EAC گزارش شده است. برای ایجاد پراکندگی در نتایج اولیه این روش ترکیبی، از نمونه‌برداری 90% از داده‌ها به روش زیرنمونه‌برداری^۱ (بدون جای‌گذاری) و مقداردهی اولیه تصادفی استفاده شده است. همچنین، الگوریتم اتصال منفرد SL به‌عنوان تابع توافقی برای به‌دست آوردن خوشه‌های نهایی از ماتریس همبستگی به کار رفته است.

سطر دوم از روش‌های ترکیبی، خوشه‌بندی ترکیبی کامل است که در آن از الگوریتم‌های مختلفی به‌عنوان الگوریتم پایه استفاده شده است. در این روش تعداد اعضای شرکت‌کننده در ماتریس همبستگی و در نتیجه در ترکیب نهایی برابر با یکصد می‌باشد. از آن جایی که الگوریتم K-means در مورد خوشه‌های کروی خوب عمل می‌کند و در برابر خوشه‌های غیر کروی عملکرد پایینی دارد، در این روش تعداد هفتاد مورد از نتایج اولیه حاصل از اجراهای مستقل الگوریتم K-means (با همان پارامترهای ذکر شده) است. سی نتیجه باقیمانده از اجرای الگوریتم‌های سلسله‌مراتبی به‌دست آمده است. از آن جایی که نتایج حاصل از اجرای

¹ Subsampling

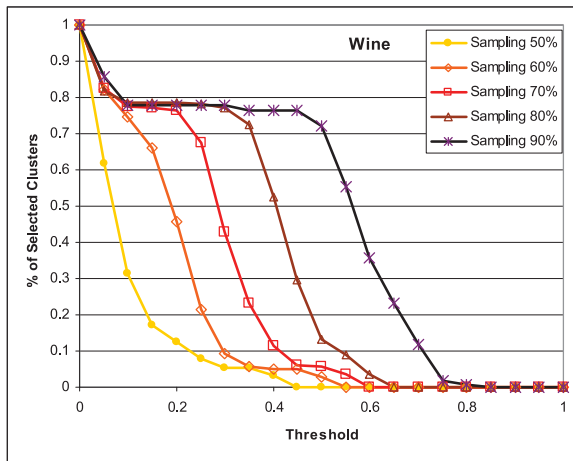
(جدول ۲) نتایج آزمایش‌ها روی میانگین ده بار اجرای مستقل

		Glass %	Breast-C %	Wine %	Bupa %	Yeast %	
Simple Methods (جین و همکاران ۱۹۹۹، تاپچی و همکاران ۲۰۰۳، دودا و همکاران ۲۰۰۱، جین و دوپس ۱۹۸۸، کافمن و روزیو ۱۹۹۰، من و گاس ۱۹۹۴)	K-means	45.65	95.13	96.63	54.55	40.20	
	Single Linkage	36.45	65.15	37.64	57.68	34.38	
	Average Linkage	37.85	70.13	38.76	57.10	35.11	
	Complete Linkage	40.65	94.73	83.71	55.94	38.91	
Ensemble Methods	EAC (فرد و جین ۲۰۰۲، ۲۰۰۵)	Kmeans Ensemble	47.76	95.46	96.63	54.49	45.46
		Full Ensemble	47.83	95.10	97.08	56.78	47.17
	EEAC	Proposed Ensemble	48.88	98.33	98.31	58.39	47.17

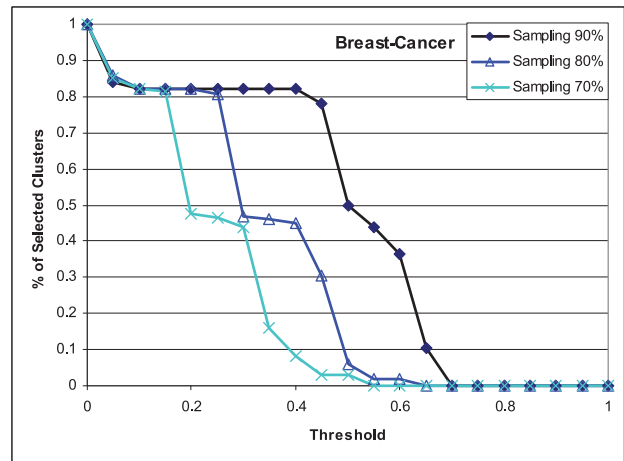
داده می‌شود. به همین دلیل در (شکل ۷)، وقتی هیچ خوشه‌ای انتخاب نشود، نرخ خوشه‌بندی از $1/k$ کمتر نمی‌شود. نتیجه دیگری که از (شکل ۷) برمی‌آید این است که با استفاده از روش پیشنهادی می‌توان تنها در حضور زیرمجموعه کوچکی از خوشه‌های اولیه به جواب بالاتری از روش ترکیب کامل دست یافت. با توجه به (شکل ۷) با روش نمونه‌برداری ۹۰٪، اگر مقدار آستانه برابر با $0.1/6$ در نظر گرفته شود، روش پیشنهادی در هر دو مجموعه داده به جواب تاحدودی بهینه‌ای دست می‌یابد. اگر این مقدار آستانه در (شکل ۶) مورد توجه قرار گیرد، مشاهده می‌شود که در هر دو مجموعه داده، تنها ۳۵٪ از کل خوشه‌های اولیه برای ترکیب نهایی مورد استفاده قرار می‌گیرند.

شایان ذکر است روش پیشنهادی می‌تواند روی انواع مجموعه داده‌های بزرگ یا کوچک مورد استفاده قرار گیرد. با توجه به این که روش پیشنهادی اندازه مجمع خوشه‌های اولیه را به‌طور چشم‌گیری کاهش می‌دهد، استفاده از آن برای مجموعه داده‌های بزرگ می‌تواند به بهبود پیچیدگی زمانی و سرعت بالاتر مسئله کمک کند.

(شکل ۶) اثر مقدار آستانه را روی درصد خوشه‌های انتخابی در دو مجموعه داده Wine و Breast-Cancer نشان می‌دهد. با توجه به این شکل هر چه نرخ نمونه‌برداری افزایش یابد، میانگین پایداری خوشه‌های به‌دست آمده نیز افزایش خواهد یافت. به‌عنوان مثال در مجموعه داده Breast-Cancer با مقدار آستانه ۰/۴، تنها حدود ۸٪ از خوشه‌های با نرخ نمونه‌برداری ۷۰٪ انتخاب می‌شوند؛ درحالی که حدود ۸۲٪ از خوشه‌های با نرخ نمونه‌برداری ۹۰٪، با همین مقدار آستانه انتخاب می‌شوند. (شکل ۷) اثر مقدار آستانه را روی کارایی روش پیشنهادی در دو مجموعه داده Wine و Breast-Cancer نشان می‌دهد. وقتی که مقدار آستانه از بیشینه مقدار پایداری خوشه‌ها بیشتر باشد، هیچ خوشه‌ای انتخاب نمی‌شود و بنابراین تمام درایه‌های ماتریس همبستگی برابر با صفر خواهند شد، در چنین حالتی الگوریتم SL که وظیفه استخراج خوشه‌های نهایی را از روی ماتریس همبستگی بر عهده دارد، به‌طور تصادفی یک نمونه را به هر یک از $k-1$ خوشه اختصاص می‌دهد و بقیه نمونه‌ها را به خوشه k -ام تخصیص می‌دهد. طی فرآیند بازبرچسب‌گذاری، خوشه k -ام به بزرگترین کلاس اختصاص

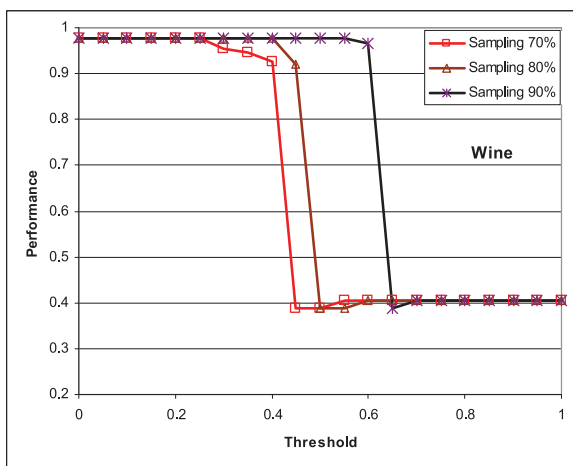


(A)

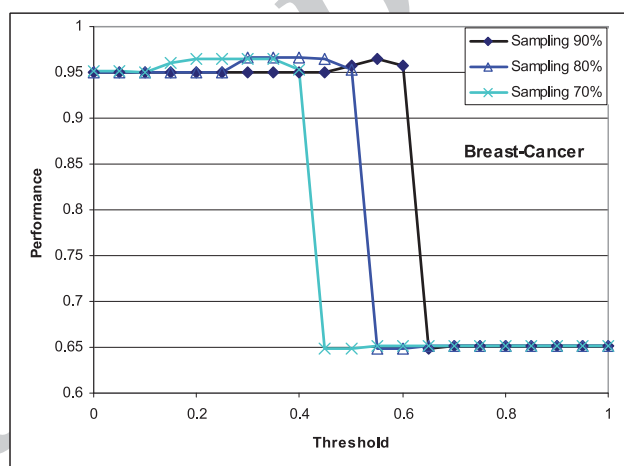


(B)

شکل ۶) اثر مقدار آستانه روی درصد خوشه‌های انتخابی در مجموعه داده‌های: A) Wine B) Breast-Cancer



(A)



(B)

شکل ۷) اثر مقدار آستانه روی کارایی روش پیشنهادی در مجموعه داده‌های: A) Wine B) Breast-Cancer

تجربی روش پیشنهادی خوشه‌بندی ترکیبی بر روی پنج مجموعه داده مختلف و متنوع نشان می‌دهد که این روش نسبت به روش‌های متداول و هم‌چنین سایر روش‌های ترکیبی برتری قابل ملاحظه‌ای دارد. نتایج آزمایش‌ها نشان می‌دهند که استفاده از به‌طور متوسط ۳۵٪ از خوشه‌های اولیه می‌تواند نتایج خوشه‌بندی ترکیبی را به‌طور مؤثری بهبود بخشد. هم‌چنین، بررسی‌ها نشان می‌دهند که اگرچه روش پیشنهادی از زیرمجموعه کوچکی از نتایج خوشه‌های اولیه استفاده می‌کند، به خاطر مؤثر بودن این زیرمجموعه، و هم‌چنین، حذف خوشه‌های با کیفیت پایین که تأثیر منفی روی میزان همبستگی واقعی نمونه‌ها می‌گذارند، نتایج نهایی حتی از ترکیب کامل هم بهتر می‌شود.

۵- نتیجه‌گیری

در این مقاله یک روش جدید برای خوشه‌بندی ترکیبی پیشنهاد شده است که مبتنی بر زیرمجموعه‌ای از خوشه‌های اولیه می‌باشد. از آنجایی که کیفیت خوشه‌های حاصل از الگوریتم‌های پایه برابر نیست و حتی حضور تعدادی از آن‌ها می‌تواند منجر به بدتر شدن نتیجه خوشه‌بندی ترکیبی شود؛ در این مقاله روشی برای انتخاب زیرمجموعه بهینه‌تر و مؤثرتر از خوشه‌های اولیه برای شرکت در ترکیب نهایی پیشنهاد شد. با توجه به این که روش‌های پیشین مبتنی بر ماتریس همبستگی برای ترکیب نتایج اولیه، توانایی لازم برای استخراج خوشه‌های نهایی از زیرمجموعه‌ای از خوشه‌های اولیه را ندارند، در این مقاله یک روش جدید برای انباشت نتایج در ماتریس همبستگی نیز پیشنهاد شده است که آن را روش انباشت مدارک توسعه‌یافته نامیده‌ایم. نتایج

- Duda, R.O., Hart, P.E., Stork, D.G., 2001. Pattern Classification, 2nd ed. Wiley, 2001.
- Dudoit, S., Fridlyand, J., 2003. Bagging to improve the accuracy of a clustering procedure, *Bioinformatics*, 19 (9), pp. 1090-1099.
- Faceli, K., Marcilio, C.P., Souto, D., 2006. Multi-objective Clustering Ensemble, *Proceedings of the Sixth International Conference on Hybrid Intelligent Systems (HIS'06)*.
- Fern, X., Brodley, C.E., 2003. Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach, In *Proc. 20th Int. conf. on Machine Learning, ICML 2003*.
- Fern, X., Lin, W., 2008. Cluster Ensemble Selection, *SIAM International Conference on Data Mining (SDM08)*.
- Fred, A., Jain, A.K., 2002. Data Clustering Using Evidence Accumulation, *Proc. of the 16th Intl. Conf. on Pattern Recognition, ICPR02, Quebec City*, pp. 276 – 280.
- Fred, A.L., Jain, A.K., 2005. Combining Multiple Clusterings Using Evidence Accumulation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(6):835–850.
- Fred, A.L., Jain, A.K., 2006. Learning Pairwise Similarity for Data Clustering, In *Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06)*.
- Fred, A., Lourenco, A., 2008. Cluster Ensemble Methods: from Single Clusterings to Combined Solutions, *Studies in Computational Intelligence (SCI)*, 126, 3–30.
- Fridlyand, J., Dudoit, S., 2001. 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*.
- Inokuchi, R., Nakamura, T., Miyamoto, S., 2006. 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*.
- Jain, A., Dubes, R.C., 1988. *Algorithms for Clustering Data*. Prentice Hall.
- Alizadeh, H., Amirgholipour, S.K., Seyedaghaee, N.R., Minaei-Bidgoli, B., 2009. 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*.
- Ayad, H.G., Kamel, M.S., 2008. Cumulative Voting Consensus Method for Partitions with a Variable Number of Clusters, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, VOL. 30, NO. 1, 160-173.
- Ayad, H., Kamel, M., 2005. Cluster-based cumulative ensembles. In N. Oza and R. Polikar, editors, *Proc. the 6th Intl. Workshop on Multiple Classifier Systems*, pages 236–245. *LCNS 3541*.
- Barthelemy, J.P., Leclerc, B., 1995. 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.
- Baumgartner, R., Somorjai, R., Summers, R., Richter, W., Ryner, L., Jarmasz, M., 2000. Resampling as a Cluster Validation Technique in fMRI, *JOURNAL OF MAGNETIC RESONANCE IMAGING 11*: pp. 228–231.
- Ben-Hur, A., Elisseeff, A., Guyon, I., 2002. A stability based method for discovering structure in clustered data, in *Pacific Symposium on Biocomputing*, vol. 7, pp. 6-17.
- Estivill-Castro, V., Yang, J., 2003. Cluster Validity Using Support Vector Machines, *DaWaK 2003, LNCS 2737*, pp. 244–256.
- Brandsma, T., Buishand, T.A., 1998. Simulation of extreme precipitation in the Rhine basin by nearest-neighbour resampling, *Hydrology and Earth System Sciences 2*, pp. 195-209.
- Breckenridge, J., 1989. Replicating cluster analysis: Method, consistency and validity, *Multivariate Behavioral research*.
- Das, A.K., Sil, J., 2007. Cluster Validation using Splitting and Merging Technique, in *proc. of Int. Conf. on Computational Intelligence and Multimedia Applications, ICCIMA*.

Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06).

Newman, C.B.D.J., Hettich, S., Merz, C., 1998. UCI repository of machine learning databases, <http://www.ics.uci.edu/~mllearn/MLSummary.html>

Rakhlin, A., Caponnetto, A., 2007. Stability of k-means clustering, In Advances in Neural Information Processing Systems 19, MIT Press, Cambridge, MA.

Roth, V., Braun, M.L., Lange, T., Buhmann, J.M., 2002b. Stability-Based Model Order Selection in Clustering with Applications to Gene Expression Data, ICANN 2002, LNCS 2415, pp. 607–612.

Roth, V., Lange, T., 2004. Feature Selection in Clustering Problems, In Advances in Neural Information Processing Systems, NIPS04.

Roth, V., Lange, T., Braun, M., Buhmann, J., 2002a. A Resampling Approach to Cluster Validation, Intl. Conf. on Computational Statistics, COMPSTAT.

Shamiry, O., Tishby, N., 2007. Cluster Stability for Finite Samples, 21st Annual Conference on Neural Information Processing Systems (NIPS07).

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

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



حسین علیزاده مدرک کارشناسی خود

را در رشته مهندسی کامپیوتر- نرم‌افزار از دانشگاه پیام نور بابل اخذ کرده است.

سیس، کارشناسی ارشد خود را در رشته

مهندسی کامپیوتر در گرایش هوش مصنوعی و رباتیک از دانشگاه علم و صنعت ایران در سال ۱۳۸۷ گرفت. او از سال ۱۳۸۸ دانشجوی دکترای مهندسی کامپیوتر، گرایش هوش مصنوعی و رباتیک در دانشگاه علم و صنعت ایران می‌باشد.

نشانی رایانامک ایشان عبارت است از:

halizadeh@iust.ac.ir

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

Kaufman, L., Rosseeuw, P.J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley & Sons, Inc.

Kuncheva, L.I., Hadjitodorov, S., 2004. Using diversity in cluster ensembles. In Proc. of IEEE Intl. Conference on Systems, Man and Cybernetics, pages 1214–1219.

Lange, T., Braun, M.L., Roth, V., Buhmann, J.M., 2003. Stability-based model selection. In Advances in Neural Information Processing Systems 15. MIT Press.

Lange, T., Roth, V., Braun, M.L., Buhmann, J.M., 2004. Stability-based validation of clustering solutions. Neural Computation, 16(6):1299–1323.

Law, M.H.C., Topchy, A.P., Jain, A.K., 2004. Multiobjective data clustering. In Proc. of IEEE Conference on Computer Vision and Pattern Recognition, volume 2, pages 424–430, Washington D.C.

Levine, E., Domany, E., 2001. Resampling Method for Unsupervised Estimation of Cluster Validity. Neural Computation 13: 2573-2593.

Luxburg, U.V., Ben-David, S., 2005. Towards a statistical theory of clustering, Technical report, PASCAL workshop on clustering, London.

Man, Y., Gath, I., 1994. Detection and Separation of Ring-Shaped Clusters Using Fuzzy Clusters, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 16, no. 8, pp. 855-861.

Minaei-Bidgoli, B., Topchy, A., Punch, W.F., 2004. Ensembles of Partitions via Data Resampling, in Proc. Intl. Conf. on Information Technology, ITCC 04, Las Vegas.

Mohammadi, M., Alizadeh, H., Minaei-Bidgoli, B., 2008. Neural Network Ensembles using Clustering Ensemble and Genetic Algorithm, 2008 Intl. Conf. on Convergence and hybrid Information Technology, ICCIT08, Nov. 11-13, IEEE CS, Korea.

Moller, U., Radke, D., 2006. A Cluster Validity Approach based on Nearest-Neighbor Resampling, In



محسن مشکی کارشناسی ارشد خود را در رشته مهندسی کامپیوتر در گرایش هوش مصنوعی و رباتیک از دانشگاه علم و صنعت ایران در سال ۱۳۸۷ گرفت. او از سال ۱۳۸۸ دانشجوی دکتری مهندسی کامپیوتر، گرایش هوش مصنوعی و رباتیک در دانشگاه علم و صنعت ایران می‌باشد.

نشانی رایانامک ایشان عبارت است از:

moshki@iust.ac.ir



حمید پروین کارشناسی ارشد خود را در رشته مهندسی کامپیوتر در گرایش هوش مصنوعی و رباتیک از دانشگاه علم و صنعت ایران در سال ۱۳۸۷ گرفت. او از سال ۱۳۸۸ دانشجوی دکتری مهندسی کامپیوتر، گرایش هوش مصنوعی و رباتیک در دانشگاه علم و صنعت ایران می‌باشد.

نشانی رایانامک ایشان عبارت است از:

parvin@iust.ac.ir



بهروز مینایی بیدگلی دکتری خود را در رشته علوم و مهندسی کامپیوتر از دانشگاه ایالتی میشیگان آمریکا در سال ۱۳۸۴ گرفت. تخصص او هوش مصنوعی و داده کاوی است و هم اکنون به‌عنوان عضو هیئت علمی دانشکده مهندسی کامپیوتر دانشگاه علم و صنعت به تدریس دروس هوش مصنوعی و نرم‌افزار مشغول می‌باشد. ایشان سرپرستی گروه متن کاوی برای متون عربی و فارسی را در پژوهشکده داده کاوی نور نیز به عهده دارد. از سال ۱۳۸۶ ریاست بنیاد ملی بازی‌های رایانه‌ای بر عهده ایشان می‌باشد.

نشانی رایانامک ایشان عبارت است از:

b_minaci@iust.ac.ir

Archives of Science