دانشكده مهندسي كامپيوتر

خوشه‌بندي تركيبي مبتني بر زيرمجموعه‌اي از نتايج اوليه

پايان‌نامه براي دريافت درجه كارشناسي ارشد

در رشته مهندسي كامپيوتر گرايش هوش مصنوعي و رباتيک

حسين عليزاده

استاد راهنما:

دكتر بهروز مينايي بيدگلي

اسفندماه ۱۳۸۷

دانشكده مهندسي كامپيوتر

خوشه‌بندي تركيبي مبتني بر زيرمجموعه‌اي از نتايج اوليه

پايان‌نامه براي دريافت درجه كارشناسي ارشد

در رشته مهندسي كامپيوتر گرايش هوش مصنوعي و رباتيک

حسين عليزاده

استاد راهنما:

دكتر بهروز مينايي بيدگلي

اسفندماه ۱۳۸۷


بسم الله الرحمن الرحيم

تأييديه‌ي هيأت داوران جلسه‌ي دفاع از پايان‌نامه

نام دانشكده: مهندسي كامپيوتر

نام دانشجو: حسين عليزاده

عنوان پايان‌نامه: خوشه‌بندي تركيبي مبتني بر زيرمجموعه‌اي از نتايج اوليه

تاريخ دفاع: ۱۳۸۷/۱۲/۱۰

رشته: مهندسي كامپيوتر

گرايش: هوش مصنوعي و رباتيک

رديف	سمت	نام و نام خانوادگي	مرتبه دانشگاهي	دانشگاه يا مؤسسه	امضا
۱	استاد راهنما	دكتر بهروز مينايي بيدگلي	استاديار	دانشگاه علم و صنعت ايران	
۲	استاد راهنما				
۳	استاد مشاور				
۴	استاد مشاور				
۵	استاد مدعو خارجي	دكتر حميد سلطانيان‌زاده		دانشگاه تهران	
۶	استاد مدعو خارجي				
۷	استاد مدعو داخلي	دكتر مرتضي آنالويي		دانشگاه علم و صنعت ايران	
۸	استاد مدعو داخلي				

تأييديه‌ي صحت و اصالت نتايج

باسمه تعالي

اينجانب حسين عليزاده به شماره دانشجويي ۸۵۷۲۲۰۲۸ دانشجوي رشته مهندسي كامپيوتر مقطع تحصيلي كارشناسي ارشد تأييد مي‌نمايم كه كليه‌ي نتايج اين پايان‌نامه حاصل كار اينجانب و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداري‌شده از آثار ديگران را با ذكر كامل مشخصات منبع ذكر كرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخيص دانشگاه مطابق با ضوابط و مقررات حاكم (قانون حمايت از حقوق مؤلفان و مصنفان و قانون ترجمه و تكثير كتب و نشريات و آثار صوتي، ضوابط و مقررات آموزشي، پژوهشي و انضباطي ...) با اينجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مكتسب و تشخيص و تعيين تخلف و مجازات را از خويش سلب مي‌نمايم. در ضمن، مسؤوليت هرگونه پاسخگويي به اشخاص اعم از حقيقي و حقوقي و مراجع ذي‌صلاح (اعم از اداري و قضايي) به عهده‌ي اينجانب خواهد بود و دانشگاه هيچ‌گونه مسؤوليتي در اين خصوص نخواهد داشت.

نام و نام خانوادگي: حسين عليزاده

امضا و تاريخ: ۱۳۸۷/۱۱/۲۵

مجوز بهره‌برداري از پايان‌نامه

بهره‌برداري از اين پايان‌نامه در چهارچوب مقررات كتابخانه و با توجه به محدوديتي كه توسط استاد راهنما به شرح زير تعيين مي‌شود، بلامانع است:

☐ بهره‌برداري از اين پايان‌نامه براي همگان بلامانع است.

☐ بهره‌برداري از اين پايان‌نامه با اخذ مجوز از استاد راهنما، بلامانع است.

☐ بهره‌برداري از اين پايان‌نامه تا تاريخ ..... ممنوع است.

نام استاد يا اساتيد راهنما: بهروز مينايي بيدگلي

تاريخ: ۱۳۸۷/۱۱/۲۵

امضا:

تقديم به:

روح پاک پدرم که نيکنامي را به من آموخت.

و مادر فداکارم که ايثار و فداکاري را در مکتب ايشان مي‌آموزم.

و برادر عزيزم که برايم سرمشق درستکاري و مسئوليت‌پذيري هستند.

با تشكر و قدرداني از:
زحمات استاد عزيز و گرانقدرم جناب آقاي دكتر مينايي كـه همـواره مـرا از راهنمـايي‌هـاي ارزشـمند خويش برخوردار كرده‌اند.

همچنين، بر خود لازم مي‌دانم تا از زحمات دوست عزيز و مهربانم آقاي حميد پروين كـه همـواره در تمام مراحل يار و ياور من بوده‌اند، كمال تشكر و قدرداني را بنمايم.
به علاوه از دوستان عزيزم آقايان محسن مشكي و جواد عظيمي كه با همفكري‌هاي خود مرا در به ثمر رساندن اين تحقيق ياري نموده‌اند، تشكر و سپاس‌گذاري مي‌كنم.

چكيده

خوشه‌بندي داده‌ها يكي از مراحل اصلي در داده‌كاوي است كه وظيفه كاوش الگوهاي پنهان در داده‌هاي بدون برچسب را بر عهده دارد. به خاطر پيچيدگي مسئله و ضعف روش‌هاي خوشه‌بندي پايه، امروزه اكثر مطالعات به سمت روي روش‌هاي خوشه‌بندي تركيبي هدايت شده است. پراكندگي در نتايج اوليه يكي از مهم‌ترين عواملي است كه مي‌تواند در كيفيت نتايج نهايي اثرگذار باشد. همچنين، كيفيت نتايج اوليه نيز عامل ديگري است كه در كيفيت نتايج حاصل از تركيب موثر است. هر دو عامل در تحقيقات اخير خوشه‌بندي تركيبي مورد توجه قرار گرفته‌اند. اما پاسخ به بعضي سوالات در اين زمينه همچنان با ابهامات زيادي روبروست. اين كه در حضور يک مجموعه از نتايج خوشه‌بندي‌هاي اوليه، آيا استفاده از زيرمجموعه‌اي از نتايج مي‌تواند بهتر از استفاده از كل نتايج باشد يا نه؟ كدام زيرمجموعه از نتايج اوليه مي‌تواند منجر به بهبود عملكرد خوشه‌بندي تركيبي شود؟ نتايج اوليه با چه معياري نسبت به هم ارزيابي شوند؟ در اين پايان‌نامه سعي شده است تا پاسخي درخور به اين سوالات داده شود. در اينجا يک چهارچوب جديد براي بهبود كارايي خوشه‌بندي تركيبي پيشنهاد شده است كه مبتني بر استفاده از زيرمجموعه‌اي از خوشه‌هاي اوليه مي‌باشند. به علاوه، براي هر مرحله از اين چهارچوب پيشهادي تعدادي روش جديد ارائه شده است. ايده‌هاي اصلي در روش‌هاي پيشنهادي براي انتخاب زيرمجموعه‌اي از خوشه‌ها، استفاده از خوشه‌هاي پايدار به علاوه استفاده از مفاهيم پراكندگي و كيفيت در نتايج اوليه مي‌باشند. براي ارزيابي خوشه‌ها، از معيار پايداري مبتني بر اطلاعات متقابل به علاوه چند معيار پيشنهادي ديگر استفاده شده است. همچنين براي ساخت ماتريس همبستگي با در دسترس بودن تنها تعدادي از خوشه‌ها، دو روش جديد پيشنهاد شده است. نتايج تجربي روي چندين مجموعه داده استاندارد نشان مي‌دهد كه روش‌هاي پيشنهادي مي‌توانند به طور موثري نتايج خوشه‌بندي‌هاي اوليه و همچنين روش تركيب كامل را بهبود دهند. مقدار متوسط بهبود روي ۱۱ مجموعه آزمايش شده نسبت به روش تركيب كامل ۲٫۳٪ مي‌باشد. به علاوه چندين مجموعه داده مصنوعي نيز توليد و در آزمايش‌ها مورد استفاده قرار گرفته‌اند.

واژه‌هاي كليدي: خوشه‌بندي تركيبي، ارزيابي خوشه، اطلاعات متقابل، زيرمجموعه‌اي از نتايج اوليه، خوشه‌بندي انباشت مدارک، ماتريس همبستگي.

فهرست مطالب

فهرست اشکال

فهرست جداول

فصل ۱:

مقدمه

۱-۱- مقدمه

ايده اصلي خوشه‌بندي اطلاعات، جداكردن نمونه‌ها از يكديگر و قرار دادن آنها در گروه‌هاي شبيه به هم مي‌باشد. به اين معني كه نمونه‌هاي شبيه به هم بايد در يك گروه قرار بگيرند و با نمونه‌هاي گروه‌هاي ديگر حداكثر تفاوت را دارا باشند (جين[1] و همكاران، ۱۹۹۹؛ فاصلي[2] و همكاران، ۲۰۰۶). در واقع خوشه‌بندي داده‌ها يک ابزار ضروري براي يافتن گروه‌ها در داده‌هاي بدون برچسب است (استرل و گاش[3]، ۲۰۰۲). حداقل ۵ دليل اصلي براي اهميت خوشه‌بندي وجود دارد:

۱-جمع‌آوري و برچسب‌گذاري يک مجموعه بزرگ از الگوهاي نمونه مي‌تواند بسيار با ارزش باشد.

۲-ممكن است ما به دنبال‌كردن در جهت معكوس علاقمند باشيم. يعني آموزش با مقدار زياد داده‌هاي بدون برچسب و سپس تنها استفاده از ناظر براي برچسب‌گذاري خوشه‌هاي پيدا شده. اين مي‌تواند براي كاربردهاي داده‌كاوي[4] بزرگ كه محتويات يک پايگاه داده از قبل شناخته شده نيست، مناسب باشد (داودي و همكاران، ۱۳۸۵).

۳-در خيلي از كاربردها مشخصه‌هاي الگوها مي‌توانند به آهستگي با زمان تغيير كنند، مثل رده‌بندي[5] خودكار مواد غذايي با تغيير فصل. اگر اين تغييرات بتواند با يک رده‌بندي‌كننده[6] به صورت بدون ناظر[7] رهگيري[8] شود، عملكرد بهتري مي‌تواند به دست آيد.

۴-مي‌توانيم از روش‌هاي بدون‌ناظر (خوشه‌بندي) براي پيداكردن و استخراج ويژگي‌ها استفاده كنيم.

۵-با خوشه‌بندي مي‌توانيم يک ديد و بينشي از از طبيعت و ساختار داده به دست آوريم كه اين

---

[1] Jain
[2] Faceli
[3] Strehl and Gosh
[4] Data Mining
[5] Classification
[6] Classifier
[7] Unsupervised
[8] Tracking

مي‌تواند براي ما با ارزش باشد. كشف زيرردەهاي[1] مجزا يا شباهت‌هاي بين الگوها ممكن است به طور چشمگيري در روش طراحي رده‌بندي‌كننده به ما پيشنهاد ارايه كند.

۱-۲- خوشه‌بندي تركيبي

از آنجايي كه اكثر روش‌هاي خوشه‌بندي پايه روي جنبه‌هاي خاصي از داده‌ها تاكيد مي‌كنند، در نتيجه روي مجموعه داده‌هاي خاصي كارآمد مي‌باشند. به همين دليل، نيازمند روش‌هايي هستيم كه بتواند با استفاده از تركيب اين الگوريتم‌ها و گرفتن نقاط قوت هر يک، نتايج بهينه‌تري را توليد كند. در واقع هدف اصلي خوشه‌بندي تركيبي جستجوي نتايج بهتر و مستحكم‌تر، با استفاده از تركيب اطلاعات و نتايج حاصل از چندين خوشه‌بندي اوليه است (استرل و گاش، ۲۰۰۲؛ فرد و جين[2]، ۲۰۰۲؛). تا كنون مطالعات زيادي بر روي تركيب رده‌بندها انجام شده است (پروين[3] و همكاران، ۲۰۰۹a؛ ۲۰۰۹b؛ عليزاده و همكاران، ۲۰۰۹b). خوشه‌بندي تركيبي نيز در اين عرصه همپاي رده‌بندي تركيبي مورد توجه بسيار قرار گرفته است. تحقيقات اخير در اين زمينه نشان داده‌اند كه خوشه‌بندي داده‌ها مي‌تواند به طور چشمگيري از تركيب چندين افراز داده سود ببرد. به علاوه، قدرت موازي‌سازي آنها يک انطباق طبيعي با نياز داده‌كاوي توزيع‌شده دارد. خوشه‌بندي تركيبي مي‌تواند جواب‌هاي بهتري از نظر استحكام[4]، نو بودن[5]، پايداري[6] و انعطاف‌پذيري[7] نسبت به روش‌هاي پايه ارايه دهد (استرل و گاش، ۲۰۰۲؛ تاپچي[8] و همكاران، ۲۰۰۳؛ فرد و لورنسو[9]، ۲۰۰۸؛ آيد و كمل[10]، ۲۰۰۸).

[1] SubClass
[2] Fred and Jain
[3] Parvin
[4] Robustness
[5] Novelty
[6] Stability
[7] Flexibility
[8] Topchy
[9] Fred and Lourenco
[10] Ayad and Kamel

به طور خلاصه خوشه‌بندي تركيبي شامل دو مرحله اصلي زير مي‌باشد (جين و همكاران، ۱۹۹۹؛ استرل و گاش، ۲۰۰۲):

۱- توليد نتايج متفاوت از خوشه‌بندي‌ها، به عنوان نتايج خوشه‌بندي اوليه بر اساس اعمال روش‌هاي مختلف كه اين مرحله را، مرحله ايجاد تنوع يا پراكندگي[1] مي‌نامند.

۲- تركيب نتايج به دست آمده از خوشه‌بندي‌هاي متفاوت اوليه براي توليد خوشه نهايي؛ كه اين كار توسط تابع توافقي[2] (الگوريتم تركيب‌كننده) انجام مي‌شود.

هر كدام از اين مراحل در زيربخش‌هاي زير تشريح شده‌اند.

۱-۲-۱- ايجاد پراكندگي در خوشه‌بندي تركيبي

معمولا در مرحله اول از خوشه‌بندي تركيبي تعدادي خوشه‌بندي‌هاي اوليه كه هر كدام بر ويژگي خاصي از داده‌ها تاكيد دارند، ايجاد مي‌شود. اولين و ساده‌ترين روش براي ايجاد نتايج مختلف و پراكنده از يک مجموعه داده، استفاده از الگوريتم‌هاي مختلف خوشه‌بندي است. هر الگوريتم خوشه‌بندي از يک جنبه خاصي به مسئله نگاه مي‌كند. بنابراين خطاهاي موجود در روش‌هاي مختلف، مي‌تواند با هم متفاوت باشد. اين امر مي‌تواند موجب ايجاد پراكندگي در نتايج الگوريتم‌هاي پايه خوشه‌بندي گردد. مهم‌ترين الگوريتم‌هاي خوشه‌بندي پايه كه معمولا در خوشه‌بندي تركيبي استفاده مي‌شوند، شامل الگوريتم‌هاي خوشه‌بندي سلسله‌مراتبي[3] (جين و همكاران، ۱۹۹۹؛ تاپچي و همكاران، ۲۰۰۳؛ دودا[4] و همكاران، ۲۰۰۱) و الگوريتم‌هاي خوشه‌بندي افرازبندي[5] (دودا و همكاران، ۲۰۰۱؛ جين و دوبس[6]، ۱۹۸۸؛ كافمن و روسيو[7]، ۱۹۹۰؛ من و گاس[8]، ۱۹۹۴) مي‌باشند. الگوريتم K-means،

[1] diversity
[2] Consensus Function
[3] Hierarchical
[4] Duda
[5] Partitional
[6] Jain and Dubes
[7] Kaufman and Rosseeuw
[8] Man and Goth

به دليل سادگي و توانايي مناسب در خوشه‌بندي، معمولا در بيشتر روش‌هاي خوشه‌بندي تركيبي به عنوان الگوريتم خوشه‌بندي پايه، استفاده مي‌شود (جين و دوبس، ۱۹۸۸؛ مينايي و همكاران، ۲۰۰۴؛ عليزاده و مينايي، ۱۳۸۷الف؛ عليزاده و همكاران، ۲۰۰۹a؛ ۲۰۰۹b؛ محمدي و همكاران، ۲۰۰۸).

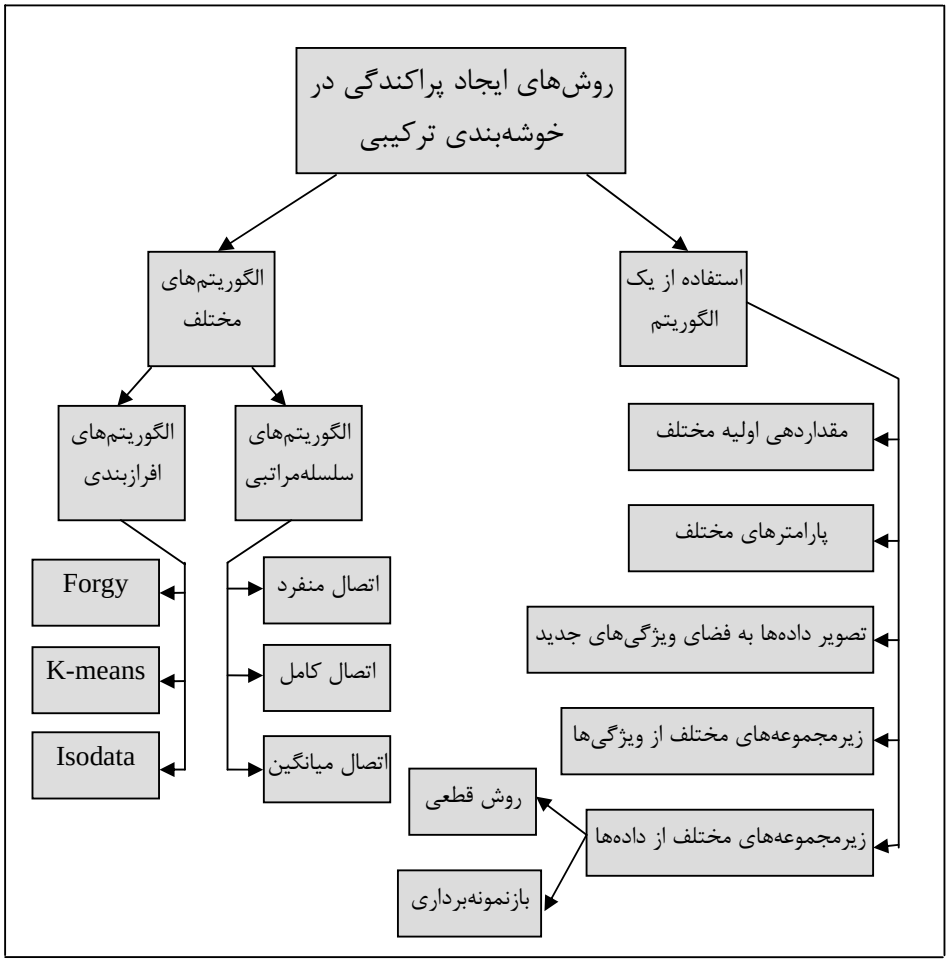
رويكرد ديگر براي ايجاد پراكندگي، به دست آوردن نتايج متنوع از يک الگوريتم خوشه‌بندي پايه با استفاده از يكي از روش‌هاي زير مي‌باشد.

- تغيير مقادير اوليه[1] الگوريتم خوشه‌بندي انتخاب شده (فرد و جين، ۲۰۰۲)
- تغيير پارامترهاي الگوريتم خوشه‌بندي انتخاب شده (بارتلمي و ركلرک[2]، ۱۹۹۵)
- استفاده از زيرمجموعه‌هاي مختلف از ويژگي‌ها[3] (استرل و گاش، ۲۰۰۲؛ فرد و لورنسو، ۲۰۰۸)
- نگاشت داده‌ها به فضاهاي ويژگي ديگر (فرن و برادلي[4]، ۲۰۰۳؛ فرد و لورنسو، ۲۰۰۸)
- تقسيم‌بندي داده‌هاي اصلي به زير مجموعه‌هايي متفاوت و مجزا (بازنمونه‌برداري[5]) (استرل و گاش، ۲۰۰۲؛ مينايي و همكاران، ۲۰۰۴؛ دودوايت و فريدليانت[6]، ۲۰۰۳)

طبقه‌بندي مهم‌ترين روش‌هاي ايجاد پراكندگي در نتايج اوليه در شكل ۱-۱ ارايه شده است.

در اين پايان‌نامه با استفاده از چندين روش سعي در ايجاد تنوع در خوشه‌بندي‌هاي اوليه شده است. بازنمونه‌برداري و ايجاد زيرمجموعه‌هاي متفاوت از داده‌ها، الگوريتم‌هاي مختلف و پارامترهاي مختلف از جمله روش‌هاي مورد استفاده در اين تحقيق براي ايجاد پراكندگي لازم در نتايج اوليه مي‌باشد.

[1] Initialization
[2] Barthelemy and Leclerc
[3] Features
[4] Fern and Brodley
[5] Resampling
[6] Dudoit and Fridlyand



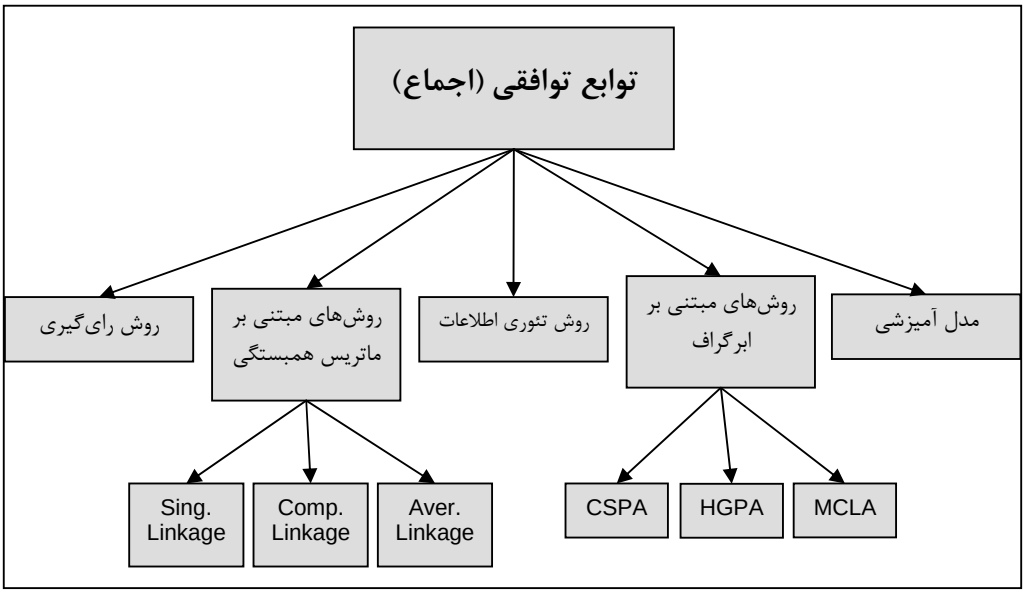
شكل (١-١) طبقه‌بندي روش‌هاي ايجاد پراكندگي در خوشه‌بندي تركيبي

١-٢-٢- تابع توافقي

پس از اينكه نتايج اوليه (تا حد ممكن پراكنده) توليد شد، معمولا با استفاده از يک تابع تركيب كننده اين نتايج تركيب مي‌شوند. يكي از متداول‌ترين روش‌هاي تركيب نتايج استفاده از ماتريس همبستگي است كه در بخش٣-٥ به طور مفصل تشريح خواهد شد. روش خوشه‌بندي تركيبي انباشت مدارک (EAC[1]) كه مبتني بر ماتريس همبستگي است، اولين بار توسط فرد و جين (٢٠٠٢) مطرح شد و خيلي زود به صورت يک روش متداول درآمد. امروزه روش‌هاي ديگري نيز مبتني بر ماتريس همبستگي

[1] Evidence Accumulation Clustering

ارايه شده‌اند (آيد و كمل، ۲۰۰۸). شكل (۱-۲) يک طبقه‌بندي كلي از توابع توافقي را نشان مي‌دهد.



شكل (۱-۲) طبقه‌بندي توابع توافقي در خوشه‌بندي تركيبي

در اين پايان‌نامه از يک روش مبتني بر ماتريس همبستگي براي تركيب نتايج استفاده شده است. همانگونه كه در شكل ۱-۲ مشاهده مي‌شود، معمولا در روش‌هاي مبتني بر ماتريس همبستگي از يک الگوريتم سلسله‌مراتبي براي استخراج نتايج نهايي از ماتريس همبستگي استفاده مي‌شود.

۱-۳- مشكلات پيش روي خوشه‌بندي تركيبي

تركيب خوشه‌بندي‌ها كار مشكل‌تري از تركيب رده‌بندي‌هاي با ناظر است. به عبارت بهتر، برخلاف مسئله رده‌بندي كه داراي ناظر[1] و يک مجموعه يادگيري[2] مي‌باشد، در خوشه‌بندي هيچ‌گونه شناختي نسبت به مجموعه داده وجود ندارد. عدم وجود ناظر و مجموعه يادگيري، ارايه روش‌هاي مدرن و هوشمند خوشه‌بندي داده‌ها كه داراي كارايي بالا باشند را بسيار مشكل نموده است. همچنين، در

[1] Supervisor
[2] Train

غياب داده آموزشي برچسب‌دار، ما با مشكل تناظر بين برچسب‌هاي خوشه در افرازهاي مختلف از يک تركيب مواجه هستيم.

بسياري از مطالعات در سال‌هاي اخير در اين زمينه استوار بوده است كه خوشه‌بندي‌هاي اوليه متنوع‌تري را براي ايجاد نتايج اوليه به كارگيرند (فيشر و بوهمن[1]، ۲۰۰۳؛ فرد و جين، ۲۰۰۳؛ تاپچي و همكاران، ۲۰۰۴؛ فرد و لورنسو، ۲۰۰۸؛ عظيمي و همكاران، ۱۳۸۴). مفهوم پراكندگي به طور گسترده‌اي در تحقيقات سال‌هاي اخير مورد استفاده قرار گرفته است (استرل و گاش، ۲۰۰۲؛ فرد و جين، ۲۰۰۵ و ۲۰۰۶؛ فرد و لورنسو، ۲۰۰۸). هدف اصلي در اكثر روش‌هاي اخير خوشه‌بندي تركيبي، تنها بررسي مجموعه داده از زواياي مختلف است و اين سوال كه "آيا پراكندگي بوجود آمده مفيد مي‌باشد يا نه؟" چندان مورد توجه قرار نگرفته است. در حقيقت به خاطر ماهيت بدون ناظر بودن مسئله خوشه‌بندي مطالعه اين امر با دشواري‌هاي زيادي روبروست. اگرچه نتايج تجربي نشان داده‌اند كه ايجاد پراكندگي در خوشه بندي‌هاي اوليه معمولا موجب بهبود خوشه‌بندي در اكثر مواقع مي‌شود (كانچوا و وايتاكر[2]، ۲۰۰۳)، عظيمي (۱۳۸۶) نشان داده است كه در بعضي مجموعه داده‌ها، پراكندگي بيشتر لزوما كمكي به افزايش دقت در نتايج نهايي نمي‌كند.

عامل ديگري كه معمولا براي بهبود عملكرد خوشه‌بندي تركيبي از آن استفاده شده است، كيفيت نتايج اوليه مي‌باشد. نشان داده شده است كه هر چه نتايج اوليه علاوه بر داشتن پراكندگي لازم، از كيفيت بالاتري برخوردار باشند، كيفيت خوشه‌هاي نهايي نيز بهتر خواهد بود (كانچوا و هاجيتودوروف[3]، ۲۰۰۴). اگرچه فرن و لين[4] (۲۰۰۸) نشان داده‌اند كه بهينه‌سازي همزمان دو عامل پراكندگي و كيفيت در نتايج اوليه خوشه‌بندي تركيبي مي‌تواند كارايي خوشه‌بندي تركيبي را به طور چشمگيري بهبود بخشد، تنظيم و مصالحه بين اين دو عامل مسئله‌اي است كه حل دقيق آن هنوز با دشواري‌هاي فراواني روبروست.

در اين تحقيق سعي شده است تا زيرمجموعه‌ي نسبتا بهينه‌اي از نتايج خوشه‌بندي‌هاي اوليه براي

[1] Fischer and Buhmann
[2] Kuncheva and Whitaker
[3] Kuncheva and Hadjitodorov
[4] Fern and Lin

شركت در تركيب نهايي يافت شود كه بتواند كارايي خوشه‌بندي تركيبي را بهبـود بخـشد. بـراي ايـن منظور يک چهارچوب كلي و چند روش مختلف پيشنهاد شده است.

فصل دوم از اين پايان‌نامه به بررسي اجمالي كارهاي صورت گرفته در زمينه خوشه‌بندي تركيبي مبتني بر زيرمجموعه‌اي از نتايج اوليه مي‌پردازد. نتايج اوليه مـورد بررسـي شـامل افرازهـاي اوليـه و همچنـين خوشه‌هاي اوليه مي‌باشد. روش پيشنهادي براي خوشه‌بندي تركيبي مبتنـي بـر زيرمجموعـه‌اي از نتـايج اوليه در فصل سوم تشريح شده است. فصل چهارم به بررسي و تفسير نتايج آزمايش‌هـا مـي‌پـردازد. در نهايت جمع‌بندي حاصل از اين پايان‌نامه در فصل پنجم آمده است. همچنـين، در ايـن فـصل كارهـاي آينده معرفي مي‌شوند.

فصل ۲:

مروري بر منابع

۲-۱- مقدمه

در اين فصل مروري بر كارهاي انجام شده و تحقيقات اخير در زمينه خوشه‌بندي تركيبي صورت گرفته است. به ويژه، تحقيقات در زمينه استفاده از زيرمجموعه‌اي از نتايج اوليه در تركيب نهايي مورد توجه قرار گرفته است. همچنين، استفاده از معيار پايداري خوشه براي ارزيابي عملكرد خوشه‌بندي‌هاي اوليه كه در سال‌هاي اخير به طور فزاينده‌اي مورد توجه محققان قرار گرفته است، مورد بررسي قرار مي‌گيرد.

۲-۲- مروري بر كارهاي انجام شده

روش‌هاي خوشه‌بندي تركيبي سعي مي‌كنند تا با تركيب افرازهاي[1] مختلف توليد شده از روش‌هاي خوشه‌بندي پايه، يک افراز مستحكم[2] از داده‌ها را توليد كنند (استرل و گاش، ۲۰۰۲؛ فرد و جين، ۲۰۰۵؛ آيد و كمل، ۲۰۰۵؛ كنچوا و حاجيتودوروف، ۲۰۰۴). در اكثر مطالعات اخير، همه افرازها با وزن برابر در تركيب نهايي حاضر مي‌شوند و همه خوشه‌هاي موجود در همه افرازها نيز با وزن برابر در تركيب نهايي شركت مي‌كنند (فرد وجين، ۲۰۰۶). استرل و گاش (۲۰۰۲) يک معيار براي انتخاب از ميان تركيبات ممكن ارايه كرده‌اند كه مبتني بر كيفيت كلي يک خوشه‌بندي است. براي اين كار، آنها ميزان ثبات بين افراز تركيبي و افرازهاي پايه را در نظر گرفته‌اند و با استفاده از يک قاعده تركيبي ثابت، يک معيار شباهت دو به دو[3] را روي فضاي ويژگي‌هاي d-بعدي به كار برده‌اند.

عظيمي (۱۳۸۶) از مفهوم پراكندگي[4] براي هوشمند نمودن خوشه‌بندي تركيبي استفاده كرده است. در اين روش كه به صورت پويا اقدام به انتخاب زيرمجموعه بهينه‌اي از نتايج اوليه در تركيب نهايي

[1] Partitions
[2] Robust
[3] Pairwise
[4] Diversity

مي‌كند، ابتدا يک خوشه‌بندي تركيبي ساده انجام مي‌شود. سپس اين روش ميزان شباهت تمام نتايج خوشه‌بندي‌هاي اوليه را نسبت به جواب به دست آمده ارزيابي مي‌كند و سعي در طبقه‌بندي[1] مجموعه داده‌ها به سه مجموعه داده راحت[2] ، معمولي[3] و سخت[4] مي‌كند. در اين طبقه‌بندي، مجموعه داده راحت به مجموعه داده‌اي اطلاق مي‌شود كه خوشه‌بندي‌هاي اوليه تفاوت چنداني با خوشه‌بندي تركيبي به دست آمده نداشته باشند. به اين معني كه هر خوشه‌بندي ساده بتواند تقريبا مانند خوشه‌بندي تركيبي نتايج مشابه‌اي ارايه كند. مجموعه داده معمولي به مجموعه داده اي اطلاق مي‌شود كه خوشه بندي‌هاي اوليه نه تفاوت چنداني و نه تشابه چنداني با نتايج خوشه‌بندي تركيبي به دست آمده دارند. مجموعه داده سخت به مجموعه داده‌اي اطلاق مي‌شود كه خوشه‌بندي‌هاي اوليه تشابه چنداني با خوشه‌بندي تركيبي به دست آمده نداشته باشند. اين رويداد نشان مي‌دهد كه داده‌هاي مجموعه مورد نظر كاملا داراي مرزهاي مشترک هستند و روش‌هاي ساده و معمولي خوشه‌بندي همانند روش‌هاي پيچيده و قدرتمند خوشه‌بندي تركيبي قادر به جداسازي نمونه‌ها نمي‌باشند. سپس كل نتايج خوشه‌بندي‌هاي اوليه به چهار زيرمجموعه متفاوت بر اساس ميزان تطبيق دقتشان با نتايج خوشه‌بندي تركيبي ساده تقسيم مي‌شوند و بر اساس رده[5] هر مجموعه داده (راحت، معمولي و سخت) اقدام به انتخاب يكي از اين زيرمجموعه‌ها براي تركيب و به دست آوردن نتيجه نهايي مي‌كنيم. نتايج تجربي[6] صورت گرفته در (عظيمي، ۱۳۸۶) نيز نشان داده‌اند كه تركيب خوشه بندي‌هاي اوليه با بيشترين كمترين و ميزان متوسطي از تطبيق با خوشه‌بندي تركيبي اوليه، نتيجه بهتري را به ترتيب، در مجموعه داده‌هاي راحت، سخت و متوسط مي‌دهد. روش فوق در هر مجموعه داده سعي مي‌كند تا نتايج خوشه‌بندي اوليه‌اي كه موجب منحرف شدن نتايج نهايي مي‌شود را از تركيب نهايي خارج كند و به اين ترتيب خوشه‌بندي‌هاي تركيبي اوليه‌اي را كه داراي دقت نسبتا مناسبي هستند، وارد تركيب نهايي كند.

[1] Classification
[2] Easy
[3] Intermediate
[4] Hard
[5] Class
[6] Experimental Results

در اكثر الگوريتم‌هاي پايه براي خوشه‌بندي تركيبي از نمونه‌برداري داده‌ها استفاده مي‌شود. مسئله اصلي در اين روش‌ها چگونگي ارزيابي خوشه و خوشه‌بندي (افراز) است (عليزاده و مينايي، ۱۳۸۷ب). بامگارتنر و همكاران (۲۰۰۰) يک روش مبتني بر بازنمونه‌برداري را براي بررسي اعتبارسنجي نتايج خوشه‌بندي فازي ارايه كرده‌اند. در چند سال اخير، پايداري خوشه به عنوان يک معيار ارزيابي خوشه مورد توجه زيادي قرار گرفته است (فرد و جين، ۲۰۰۶؛ لاو[1] و همكاران، ۲۰۰۴؛ شاميري و تيشباي[2]، ۲۰۰۷؛ لانژ[3] و همكاران، ۲۰۰۳). ايده‌هاي اوليه براي اعتبارسنجي خوشه با استفاده از بازنمونه‌برداري در (بريكنريج[4]، ۱۹۸۹) ارايه شده و بعدها در (فريدياند و دودوايت[5]، ۲۰۰۱؛ ليوين و دومني[6]، ۲۰۰۱) كامل‌تر شده است. راس و همكاران (۲۰۰۲a و ۲۰۰۲b) نيز يک روش مبتني بر بازنمونه‌برداري براي اعتبارسنجي خوشه ارايه كرده‌اند. عنصر اصلي در اين روش، كه در واقع كامل‌شده‌ي روش‌هاي پيشين مي‌باشد، پايداري خوشه است. معيار پايداري، ميزان همبستگي افرازهاي به دست آمده از دو نمونه‌برداري مستقل از مجموعه داده را اندازه‌گيري مي‌كند. هر چه ميزان پايداري براي يک خوشه‌بندي بيشتر باشد، به اين معني است كه اگر الگوريتم خوشه‌بندي چندين مرتبه ديگر روي آن نمونه‌ها به كار رود، نتايج مشابهي حاصل مي‌شود (راخلين و كاپونتو[7]، ۲۰۰۷؛ لوكسبرگ و بن-ديويد، ۲۰۰۵). همچنين، راس و لانژ[8] (۲۰۰۴) يک الگوريتم جديد براي خوشه‌بندي ارايه كرده‌اند كه مبتني بر انتخاب ويژگي مي‌باشد. در اين روش از معيار پايداري مبتني بر بازنمونه‌برداري داده‌ها، براي انتخاب پارامترهاي الگوريتم خوشه‌بندي استفاده شده است. چندين روش اعتبارسنجي خوشه[9] مبتني بر ايده استفاده از پايداري پيشنهاد شده است (لانژ و همكاران، ۲۰۰۴). بن هور[10] و همكاران (۲۰۰۲) نيز روشي براي محاسبه پايداري ارايه كرده‌اند كه بر مبناي شباهت بين نمونه‌ها در

[1] Law
[2] Shamiry and Tishby
[3] Lange
[4] Breckenridge
[5] Fridlyand and Dudoit
[6] Levine and Domany
[7] Rakhlin and Caponnetto
[8] Roth and Lange
[9] Cluster Validity
[10] Ben-Hur

خوشه‌بندي‌هاي مختلف عمل مي‌كند. در اين روش، ابتدا ماتريس همبستگي با استفاده از روش بازنمونه‌برداري به دست مي‌آيد. سپس ضريب جاكارد[1] به عنوان معيار پايداري بر اساس اين ماتريس محاسبه مي‌شود. همچنين، كاسترو و يانگ[2] (۲۰۰۳) روشي براي ارزيابي افرازهاي نهايي خوشه‌بندي ارايه كرده‌اند كه از ماشين بردار پشتيبان[3] استفاده مي‌كند. اين روش با شناسايي نويزها و داده‌هاي دورافتاده[4] به نتايج خوشه‌بندي داراي استحكام دست يافته است. مولر و رادك[5] (۲۰۰۶) از بازنمونه‌برداري نزديك‌ترين همسايه (NNR[6]) براي اعتبارسنجي نتايج خوشه‌بندي استفاده كرده‌اند. اين روش بازنمونه‌برداري اولين بار در تحليل سري‌هاي زماني به كار رفته است (برندسما و بوئيشند[7]، ۱۹۹۸). اينوكوچي[8] و همكاران (۲۰۰۶) معيار اعتبارسنجي هسته-محور[9] پيشنهاد كرده‌اند. هسته در اين روش به معني تابع مركزي استفاده شده در ماشين بردار پشتيبان مي‌باشد. در اين روش، دو شاخص مورد توجه قرار گرفته است: اولي مجموع مقادير كواريانس فازي داخل خوشه‌هاست و دومي شاخص مبتني بر هسته ژي-بن[10] (۱۹۹۱). در اين روش از اين دو شاخص براي ارزيابي نتايج خوشه‌بندي و همچنين، تعيين تعداد خوشه‌ها با مرزهاي غير خطي استفاده شده است. داس و سيل[11] (۲۰۰۷) روشي براي تعيين تعداد خوشه‌ها ارايه كرده‌اند كه از اعتبارسنجي خوشه‌ها براي تقسيم و ادغام آنها بهره مي‌برد. فرن و لين (۲۰۰۸) روشي براي خوشه‌بندي تركيبي پيشنهاد كرده‌اند كه از زيرمجموعه‌ي موثرتري از افرازهاي اوليه در تركيب نهايي استفاده مي‌كند. در اين روش اگر چه تعداد اعضاي شركت كننده در تركيب نهايي كمتر از يک خوشه‌بندي تركيبي كامل[12] است، به دليل انتخاب افرازهاي با كارايي بالاتر، نتايج نهايي بهبود مي‌يابند. پارامترهايي كه در اين روش مورد توجه

[1] Jaccard Coefficient
[2] Estivill-Castro and Yang
[3] Support Vector Machine
[4] Outliers
[5] Moller and Radke
[6] Nearest Neighbor Resampling
[7] Brandsma and Buishand
[8] Inokuchi
[9] Kernelized Validity Measure
[10] Xie - Beni
[11] Das and Sil
[12] Full Ensemble

قرار گرفته‌اند، عبارتند از: كيفيت و پراكندگي. اين روش سعي مي‌كند تا زيرمجموعه‌اي از افرازهايي از نتايج اوليه را وارد تركيب نهايي كند كه از بالاترين ميزان كيفيت برخوردار بوده و در عين حال نسبت به هم بيشترين پراكندگي را دارا باشند. در اين روش از معيار مجموع اطلاعات متقابل نرمال‌شده (SNMI[1]) (براي يک افراز در مقايسه با افرازهاي ديگر تركيب) براي اندازه‌گيري كيفيت يک افراز استفاده شده است. همچنين، معيار اطلاعات متقابل نرمال‌شده (NMI[2]) (بين تمام افرازهاي موجود در تركيب) براي اندازه‌گيري پراكندگي لازم براي تركيب به كار رفته است. فرن و لين (۲۰۰۸) نشان مي‌دهند كه روش آنها نسبت به خوشه‌بندي تركيبي كامل و يا روش انتخاب تصادفي از كارايي بهتري برخوردار است. لاو و بقيه (۲۰۰۴) يک روش خوشه‌بندي چندهدفي[3] ارايه كرده‌اند كه مبتني بر انتخاب خوشه‌هاي اوليه توليد شده توسط الگوريتم‌هاي مختلف خوشه‌بندي، در طي يک روال بهينه‌سازي مي‌باشد. در اين روش، بهترين مجموعه از توابع هدف براي بخش‌هاي مختلف از فضاي ويژگي انتخاب شده است. فرد و جين (۲۰۰۶) يک روش خوشه‌بندي تركيبي ارايه كرده‌اند كه در آن با استفاده از معيار پايداري خوشه، شباهت دو به دو آموزش داده مي‌شود. در اين روش، به جاي استفاده از معيارهاي ارزيابي مبتني بر افراز نهايي، افرازهاي حاصل از الگوريتم‌هاي پايه در نواحي مختلف از فضاي ويژگي d-بعدي مورد ارزيابي قرار گرفته است.

در اكثر روش‌هاي فوق از ماتريس همبستگي براي تجميع اطلاعات خوشه‌بندي‌هاي اوليه استفاده شده است. اگر چند ماتريس وجود داشته باشد، روش‌هاي مختلفي وجود دارند كه ماتريس همبستگي نهايي را از ادغام آنها توليد مي‌كنند. بروزير[4] (۱۹۹۰) يک روش براي ادغام چند ماتريس فاصله فرامتري[5] به يک ماتريس فاصله فرامتري ارايه نمود. شرط لازم و كافي براي يكتايي عمليات ادغام نيز مورد بررسي قرار گرفته است. همچنين، لاپوينته و لجندره[6] (۱۹۹۱) دو روش براي توليد

[1] Sum of Normalized Mutual Information
[2] Normalized Mutual Information
[3] Multiobjective
[4] Brossier
[5] Ultrametric Distance Matrices
[6] Lapointe and Legendre

دندوگرام‌هاي تصادفي ارايه كرده‌اند. آنها اين كار را با تصادفي كردن ماتريس كفنتيک وابسته[1] انجام داده‌اند. هر وارده $d_C(i,j)$ در ماتريس كفنتيک بيانگر سطحي است كه در آن نمونه‌هاي x_i و x_j براي اولين بار همديگر را در يک خوشه ملاقات كرده‌اند (جين و دوبس، ۱۹۸۸). بنفيلد[2] (۱۹۷۶) الگوريتمي براي محاسبه فاصله‌هاي فرامتري براي يک دندوگرام اتصال منفرد ارايه مي‌كند. اين الگوريتم يک ماتريس فاصله از ورودي مي‌گيرد و از درخت پوشاي كمينه‌ي[3] حاصل از آن استفاده مي‌كند. فاصله‌هاي فرامتري در اين الگوريتم، به ماكزيمم اتصالات زنجيري گفته مي‌شود كه هر كدام از جفت نقاط درخت را به هم متصل مي‌كند. در واقع فاصله‌هاي فرامتري سطوحي از دندوگرام هستند كه جفت نقاط در آن سطح به هم متصل شده‌اند.

۲-۳- نتيجه‌گيري

انتخاب زيرمجموعه‌اي از نتايج اوليه براي استفاده در تركيب نهايي، مسئله‌اي است كه اخيرا مطرح شده و با وجود ابهامات فراوان در جوانب اين مسئله، امروزه به صورت جدي دنبال مي‌شود. ايده‌هاي اوليه با ارايه يک معيار ارزيابي براي كيفيت خوشه آغاز شد (لانژ و همكاران، ۲۰۰۳). تعدادي از مطالعات انجام شده سعي در انتخاب زيرمجموعه مناسبي از افرازهاي اوليه دارند (عظيمي، ۱۳۸۶؛ فرن و لين، ۲۰۰۸) و تعدادي نيز زيرمجموعه‌اي از خوشه‌هاي اوليه را مورد آزمايش قرار داده‌اند (لاو و همكاران، ۲۰۰۴؛ فرد و جين، ۲۰۰۶). پاسخ به اين سوال كه "انتخاب چه زيرمجموعه‌اي از نتايج اوليه موجب بهترين كارايي خوشه‌بندي تركيبي مي‌شود؟" هنوز نياز به انجام تحقيقات گسترده‌تري دارد. تحقيقات در اين زمينه‌ها به حدي كم و در آغاز راه قرار دارد كه حتي در حالتي كه استفاده از زيرمجموعه‌اي از خوشه‌هاي اوليه مورد توجه قرار گرفته است، هنوز راه‌حلي براي ساخت ماتريس همبستگي در غياب تعدادي از خوشه‌ها مطرح نشده است. براي گريز از اين مشكل، روش پيشنهادي براي خوشه‌بندي تركيبي توسط فرد و جين (۲۰۰۶)، از نتايج خوشه‌بندي‌هاي تركيبي به عنوان نتايج

[1] Associated Cophenetic Matrix
[2] Banfield
[3] Minimum Spanning Tree

اوليه استفاده مي‌كند. در واقع يک خوشه‌بندي تركيبي انجام نمي‌شود، بلكه زيرمجموعه‌اي از نتايج چندين خوشه‌بندي تركيبي استفاده مي‌شود.

فصل ۳:

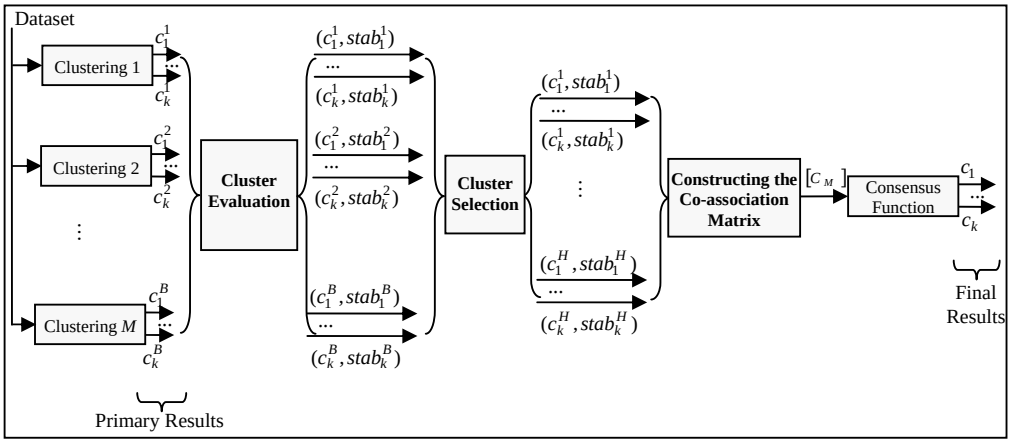
روش پیشنهادي

۳-۱- مقدمه

در فصل گذشته مروري بر كارهاي انجام شده در زمينه خوشه‌بندي تركيبي مبتني بر زيرمجموعه‌اي از نتايج اوليه انجام شد. در بيشتر كارهاي انجام شده، از معيار اطلاعات متقابل نرمال شده براي ارزيابي كيفيت خوشه‌بندي استفاده شده است. در اين فصل روش پيشنهادي براي خوشه‌بندي تركيبي مبتني بر زيرمجموعه‌اي از خوشه‌هاي اوليه ارايه شده است. همچنين، چند روش مختلف با كارايي بهتر نسبت به معيار اطلاعات متقابل نرمال شده پيشنهاد و تشريح شده است.

۳-۲- چهارچوب اصلي روش پيشنهادي

ايده اصلي در اين روش استفاده از زيرمجموعه‌اي از خوشه‌هاي اوليه به جاي كل خوشه‌ها در خوشه‌بندي تركيبي است. نماي كلي از چهارچوب كلي براي روال پيشنهادي در شكل ۳-۱ نشان داده شده است.



شكل (۳-۱) روال الگوريتم پيشنهادي براي خوشه‌بندي تركيبي

در اين روش ابتدا با استفاده از روش‌هاي ايجاد پراكندگي تعداد B خوشه‌بندي اوليه انجام مي‌شود.

اين كار مي‌تواند با استفاده از نمونه‌برداري از داده‌ها، استفاده از الگوريتم‌هاي مختلف خوشه‌بندي، استفاده از زيرمجموعه‌اي از ويژگي‌ها و يا انتخاب پارامترهاي مختلف براي يک الگوريتم خوشه‌بندي انجام شود. در اينجا از الگوريتم K-means براي توليد نتايج اوليه استفاده شده است. پراكندگي لازم در نتايج اوليه براي الگوريتم K-means نيز، با انتخاب تصادفي نقاط اوليه مراكز خوشه‌ها و همچنين با نمونه‌برداري فراهم شده است.

در مرحله بعد خوشه‌هاي به دست آمده مورد ارزيابي قرار مي‌گيرند تا كيفيت هر خوشه مشخص شود. پس از اينكه پايداري هر خوشه محاسبه شد، در گام بعد، عمل انتخاب خوشه‌ها با توجه به مقدار پايداري خوشه انجام مي‌شود. در اين تحقيق چندين روش براي انتخاب خوشه‌ها پيشنهاد شده است كه در بخش ۳-۴ تشريح شده‌اند.

در گام بعدي خوشه‌هاي انتخاب شده با هم تركيب شده و خوشه‌هاي نهايي از آنها به دست مي‌آيد. روش‌هاي مختلفي براي تركيب خوشه‌بندي‌هاي اوليه و به دست آوردن خوشه‌هاي نهايي وجود دارد. تفاوتي كه در اينجا وجود دارد اين است كه در اين روش ممكن است كه از هر خوشه‌بندي اوليه، تنها تعدادي از خوشه‌ها موجود باشند. در اين پايان‌نامه دو روش جديد براي تركيب نتايج پيشنهاد شده است. اين روش‌ها كه توانايي ساخت ماتريس همبستگي براي نمونه‌ها -در شرايطي كه تنها تعدادي از خوشه‌ها موجود هستند- را دارند، در بخش ۳-۵ تشريح شده‌اند. پس از ساخت ماتريس همبستگي، مي‌توان با استفاده از يكي از الگوريتم‌هاي سلسله‌مراتبي خوشه‌هاي نهايي را به دست آورد.

از آنجايي كه خوشه‌بندي تركيبي به فرايندي پيچيده شامل چندين مرحله مي‌باشد، پيچيدگي محاسباتي بالاتي داشته و در مجموع فرايندي زمان‌بر و برون‌خط[1] است. از همين رو، در تحقيقاتي كه در اين زمينه صورت مي‌گيرد معمولا روي پيچيدگي محاسباتي روش‌هاي پيشنهادي بحث خاصي صورت نمي‌گيرد (فرد و لورنسو، ۲۰۰۸؛ فرن و لين، ۲۰۰۸).

[1] Offline

۳-۳- ارزيابي خوشه

در اين مرحله خوشه‌هاي به دست آمده مورد ارزيابي قرار مي‌گيرند تا كيفيت هر خوشه مشخص شود. از آن جايي كه ميزان برازندگي[1] يک خوشه در ميان كل نقاط داده معني‌دار است، تابع برازندگي خوشه يعني $g_j(C_i,D)$ علاوه بر پارامتر اول خود يعني خوشه C_i به مجموعه داده D نيز وابسته است. يک تابع برازندگي بايد خصوصيات زير را داشته باشد (لاو و همكاران، ۲۰۰۴):

- بايد با تابع f_j كه توسط الگوريتم خوشه‌بندي خاص A_j بهينه مي‌شود، ارتباط منطقي داشته باشد. به عبارت ديگر، مقدار بيشتر براي $g_j(C_i,D)$ به اين معني باشد كه خوشه C_i نسبت به تابع f_j و متناظرا نسبت به الگوريتم خوشه‌بندي خاص A_j، برازنده‌تر (بهينه‌تر) است.
- بايد نسبت به توابع خوشه‌بندي مختلف قابل مقايسه باشد. به عبارت ديگر، اگر $g_j(C_i,D) > g_l(C_i,D)$، آنگاه بايد نتيجه گرفت كه كيفيت خوشه C_i با توجه به تابع f_j از تابع f_l بهتر است. يعني كيفيت خوشه C_i در خوشه‌بندي j-ام از كيفيت خوشه C_i در خوشه‌بندي l-ام بهتر مي‌باشد.
- در نهايت مقدار تابع برازندگي خوشه بايد نسبت به خوشه‌هاي مختلف قابل مقايسه باشد. به عبارت ديگر، $g_j(C_i,D) = g_j(C_l,D)$، بايد نتيجه بدهد كه خوشه‌هاي C_i و C_l نسبت به تابع f_j ميزان برازندگي برابري دارند. يعني اين دو خوشه در خوشه‌بندي j-ام از كيفيت برابري برخوردارند.

يكي از معيارهايي كه مي‌تواند به عنوان تابع برازندگي خوشه در نظر گرفته شود، معيار پايداري خوشه (لانژ و همكاران، ۲۰۰۳) است. پايداري خوشه, اثر آشفتگي[2] در نتايج خوشه‌بندي‌هاي مختلف را انعكاس مي‌دهد. يكي از مهم‌ترين روش‌هاي ايجاد آشفتگي در خوشه‌بندي استفاده از بازنمونه‌برداري است كه مي‌تواند به دو شكل رايج با جايگذاري و يا بدون جايگذاري انجام شود.

براي ارزيابي خوشه از معيارهاي پايداري مختلفي استفاده شده است. اگر چه همه روش‌هاي پيشنهادي

[1] Goodness
[2] Perturbation

براي ارزيابي خوشه مبتني بر اطلاعات متقابل نرمال شده هستند، هر كدام از اين روش‌ها يكي از معايب آن را جبران مي‌كند. تشريح كامل روش‌هاي پيشنهادي ارزيابي خوشه در اين بخش آمده است.

۳-۳-۱- اطلاعات متقابل نرمال شده

در اين روش خوشه‌ها بر اساس معيار پايداري مبتني بر اطلاعات متقابل نرمال شده (NMI) ارزيابي مي‌شوند (عليزاده و مينايي، ۱۳۸۷ج).

براي شناسايي خوشه‌هاي پايدارتر نياز به مكانيزمي است تا بتواند پايداري را براي هر خوشه از يک خوشه‌بندي، مستقل از خوشه‌هاي ديگر به دست آمده از آن خوشه‌بندي، حساب كند. براي اين كار، فرض كنيد كه مي‌خواهيم پايداري خوشه C_i را محاسبه كنيم. در اين روش ابتدا با نمونه‌برداري مجموعه داده‌هاي جديدي درست مي‌شود و خوشه‌بندي‌هاي مختلفي روي آن صورت مي‌گيرد. سپس، سعي مي‌شود تا به اين سوال كه "آيا اين خوشه، در اين خوشه‌بندي‌ها هم ظاهر شده است يا نه؟" پاسخ داده شود. براي اين كار يک معيار شباهت بين آن خوشه (C_i) و خوشه‌بندي اوليه ($P(D)$) پيشنهاد مي‌شود كه با $sim(C_i,P(D))$ نشان داده مي‌شود. با استفاده از اين معيار، شباهت آن خوشه را با خوشه‌بندي‌هاي مختلف حاصل از نمونه‌برداري محاسبه مي‌شود. سپس ميانگين اين معيارهاي شباهت، به عنوان ميزان پايداري اين خوشه $g_i(C_i,D)$ برگردانده مي‌شود. در واقع $sim(C_i,P(D))$ ميزان اعتبار خوشه C_i را درخوشه‌بندي P روي مجموعه داده D مشخص مي‌كند. شبه كد مربوط به اين روال در شكل ۳-۲ نشان داده شده است.

براي محاسبه $sim(C_i,P(D))$ كه ميزان شباهت بين خوشه C_i و نتيجه خوشه‌بندي $P(D)$ است، به صورت زير عمل مي‌شود. ابتدا تمام نمونه‌هاي ديگر متعلق به مجموعه داده D كه در خوشه C_i قرار ندارند، به صورت يک خوشه مستقل D/C_i نمايش داده مي‌شود. حال يک خوشه‌بندي شامل دو خوشه C_i و D/C_i ايجاد شده است كه آن را P_1 مي‌ناميم $P_1=\{C_i, D/C_i\}$. اكنون خوشه‌بندي $P(D)$ كه روي داده‌هاي نمونه‌برداري شده اعمال شده است، نيز بايد به صورت دو خوشه‌اي ارايه شود تا در

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
$g_j(C_i,D) := average$ of $score[l]$;

شكل (۳-۲) الگوريتم محاسبه پايداري خوشه C_i به عنوان تابع برازندگي

نهايت نتايج حاصل از اين دو خوشه‌بندي طي فرآينـدي بـا هـم منطبـق شـوند. بـراي ايـن منظـور همـه خوشه‌ها در $P(D)$ به دو خوشه C^* و D/C^* تقسيم مي‌شوند. خوشه C^* از اجتماع همه خوشه‌هـايي كـه بيش از ۵۰% از نمونه‌هايشان در خوشه C_i وجود دارند، تشكيل مي‌شود و مابقي خوشه‌ها نيز در خوشـه D/C^* قرار مي‌گيرند. اين خوشه‌بندي را P_2 مي‌ناميم $P_2=\{C^*, D/C^*\}$. حال از اطلاعات متقابل[1] نرمال شده (NMI) كه معيار متداول براي ارزيابي شباهت بين دو افراز (نتيجه خوشه‌بندي) مي‌باشد (استرل و گاش، ۲۰۰۲؛ فرد و جين، ۲۰۰۲؛ دودوايـت و فريديانـد، ۲۰۰۳)، بـراي انـدازه‌گيـري شـباهت بـين دو خوشه‌بندي P_1 و P_2 استفاده مي‌شود. از آن جايي كه معيار اطلاعات متقابل نرمال نشده (MI)، وابـسته به اندازه خوشه‌هاست، معمولا از معيار NMI استفاده مي‌شود. رابطه NMI بين دو خوشه‌بندي P_1 و P_2 به صورت زير محاسبه مي‌شود.

(۱-۳)

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

كه اطلاعات متقابل، $MI(P_1,P_2)$ از رابطه ۳-۲ به دست مي‌آيد.

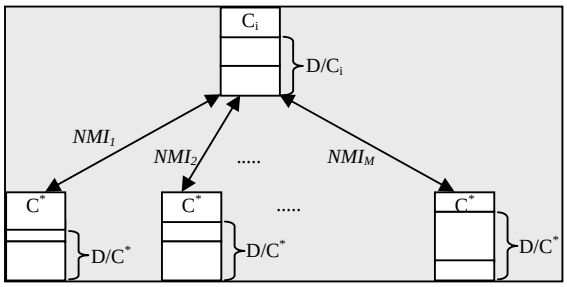
[1] Mutual Information (MI)

$$MI(P_1,P_2)=\sum_{i=0}^{1}\sum_{j=0}^{1}\frac{r_{ij}}{m^2}\log\frac{mp_{ij}}{r_{ij}} \qquad (۲-۳)$$

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

که در این رابطه، p_{11} نشان‌دهنده تعداد نمونه‌هاي مشترک موجود در C^* و C_i است. p_{10} نشان‌دهنده تعداد نمونه‌هاي مشترک موجود در D/C^* و C_i است. p_{01} نشان‌دهنده تعداد نمونه‌هاي مشترک موجود در C^* و D/C_i است. p_{00} نشان‌دهنده تعداد نمونه‌هاي 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 نمونه‌هاي موجود در C_i و C^* هستند.

شکل۳-۳ نماي کلي از این روش محاسبه پایداري خوشه را نشان مي‌دهد. با توجه به شکل ۳-۳، NMI_i نشان‌دهنده میزان شباهت خوشه‌بندي P_1 و P_2 مي‌باشد. همچنین بیانگر میزان پایداري خوشه C_i در خوشه‌بندي i-ام نیز مي‌باشد. این مقدار با توجه به الگوریتم شکل ۳-۲ در $sim(C_i,P(D))$ و سپس در $score[i]$ ذخیره مي‌گردد.



شکل (۳-۳) محاسبه پایداري خوشه C_i با روش مبتني بر NMI

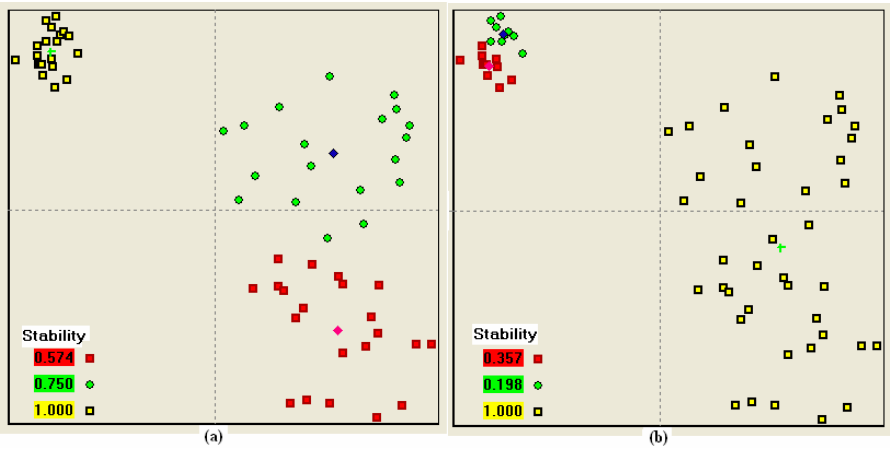
پایداري کل از میانگین کل این پایداري‌ها تشکیل مي‌شود.

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

که M تعداد خوشه‌بندي‌ها در مجموعه مرجع مي‌باشد. این روش، در واقع خوشه‌هایي را که بیشترین تکرار را در خوشه‌بندي‌هاي مختلف دارند، به عنوان خوشه‌هاي پایدارتر معرفي مي‌کند.

۳-۳-۲- روش ماكزيمم

شكل ۳-۴ دو افراز اوليه را نشان مي‌دهد كه ميزان پايداري براي هر كدام از خوشه‌هاي به دست آمده هم محاسبه شده است. در اين مثال الگوريتم K-means به عنوان الگوريتم خوشه‌بندي اوليه انتخاب شده است و تعداد خوشه‌هاي اوليه برابر با سه نيز به عنوان پارامتر آن از قبل مشخص شده است. همچنين، در اين مثال تعداد افرازهاي موجود در مجموعه مرجع[1] برابر با ۴۰ مي‌باشد. در ۳۶ افراز نتايجي مشابه با شكل ۳-۴(a) و در ۴ حالت باقيمانده نيز نتايجي مشابه با شكل ۳-۴(b) حاصل شده است.



شكل (۳-۴) دو افراز اوليه با تعداد سه خوشه. (*a*) خوشه‌بندي درست (*b*) خوشه‌بندي نادرست

از آن جايي كه در مجموعه مرجع در ۹۰% مواقع، داده‌هاي متراكم گوشه بالا-چپ از شكل ۳-۴ در يک خوشه مجزا گروه‌بندي شده‌اند، بنابراين اين خوشه بايد مقدار پايداري بالايي را به خود اختصاص دهد. اگرچه اين مقدار نبايد دقيقا برابر با يک باشد (چون در همه موارد اين خوشه درست تشخيص داده نشده است)، مقدار پايداري با روش متداول اطلاعات متقابل نرمال شده مقدار يک را بر مي‌گرداند. از آن جايي كه ادغام دو خوشه سمت راست تنها در ۱۰% موارد مانند شكل ۳-۴(b) اتفاق افتاده است، خوشه حاصل بايد مقدار پايداري كمي به دست آورد. اگر چه خوشه حاصل از

[1] Reference set

ادغام دو خوشه سمت راستي، به ندرت (۱۰% موارد) در مجموعه مرجع ديده شده است، مقدار پايداري براي اين خوشه نيز برابر با يک به دست مي‌آيد. در اينجا مشكل روش متداول محاسبه پايداري با استفاده از اطلاعات متقابل نرمال شده ظاهر مي‌شود. از آنجايي كه معيار اطلاعات متقابل نرمال شده يک معيار متقارن است، مقدار پايداري خوشه بزرگ ادغامي سمت راست (با ۱۰% تكرار) دقيقا برابر با ميزان پايداري خوشه متراكم گوشه بالا-چپ (با ۹۰% تكرار) به دست مي‌آيد. به عبارت ديگر در مواردي كه داده‌هاي دو خوشه مكمل يكديگر باشند، يعني اجتماع داده‌هاي آنها شامل كل مجموعه داده شود و اشتراک داده‌هاي آنها نيز تهي باشد، مقدار پايداري براي هر دو به يک اندازه برابر به دست مي‌آيد.

از ديدگاه ديگر، اين اتفاق زماني رخ مي‌دهد كه تعداد خوشه‌هاي تشكيل‌دهنده مجموعه C^* در خوشه‌بندي مرجع عددي بيشتر از يک باشد. هر زمان كه C^* با ادغام دو يا بيشتر از خوشه‌ها به دست آيد، منجر به نتايج نادرست در مقدار پايداري مي‌شود.

در اين پايان‌نامه چندين راهكار براي رفع اين مشكل پيشنهاد شده است:

- انتخاب بزرگترين خوشه به عنوان مجموعه C^*، از بين تمام خوشه‌هاي موجود در مجموعه مرجع كه شرط شباهت (بالاي ۵۰% نمونه‌هايشان متعلق به خوشه C_i باشد) را ارضا مي‌كنند. با استفاده از اين روش كه از اين به بعد با نام روش ماكزيمم از آن ياد مي‌كنيم، مشكل فوق هرگز پيش نمي‌آيد. چون همان طور كه پيش‌تر نيز اشاره شد، اين مشكل تنها زماني به وجود مي‌آيد كه مجموعه C^* از ادغام دو يا بيشتر از خوشه‌هاي مجموعه مرجع تشكيل شود.
- اعمال يک جريمه براي هر بار ادغام خوشه‌هاي تشكيل‌دهنده مجموعه C^*، با استفاده از جمع وزن‌دار در محاسبه پايداري، به صورتي كه تعداد ادغام‌ها در وزن‌ها اعمال شود. از اين به بعد، از اين روش با نام پايداري وزن‌دار نام خواهيم برد. روابط ۳-۴ و ۳-۵ به ترتيب نحوه محاسبه پايداري وزن‌دار و همچنين، محاسبه وزن‌ها را نشان مي‌دهد. $n_{merge(i)}$ تعداد خوشه‌هاي ادغام شده براي ساخت خوشه C^* است.

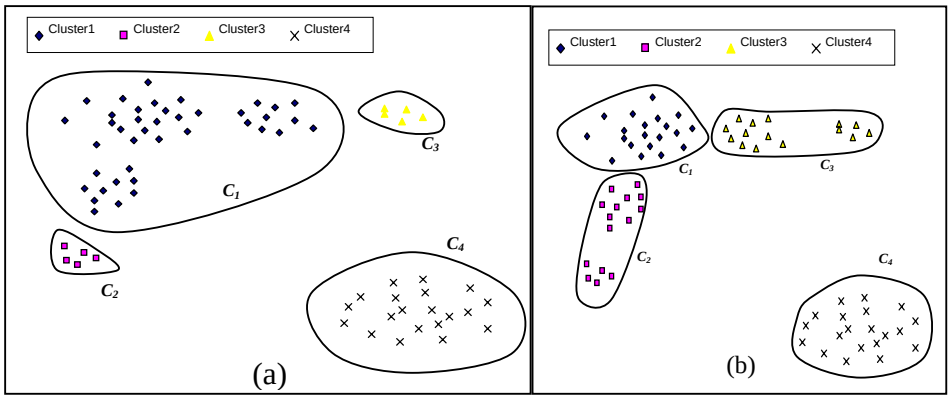
$$Stability(C_i) = \frac{1}{M}\sum_{i=1}^{M} w_i NMI_i \quad (٤-٣)$$

$$w_i = \frac{1}{n_{merge(i)}} \quad (٥-٣)$$

- روش عليزاده-مشکي-مينايي[1] (AMM) که از يک معيار نامتقارن استفاده مي‌کند. اين روش در بخش ۳-۳-۳ تشريح شده است.
- روش اطلاعات متقابل نرمال اصلاح‌شده[2] که در بخش ۳-۳-۴ تشريح شده است.

۳-۳-۳- استفاده از معيار نامتقارن AMM

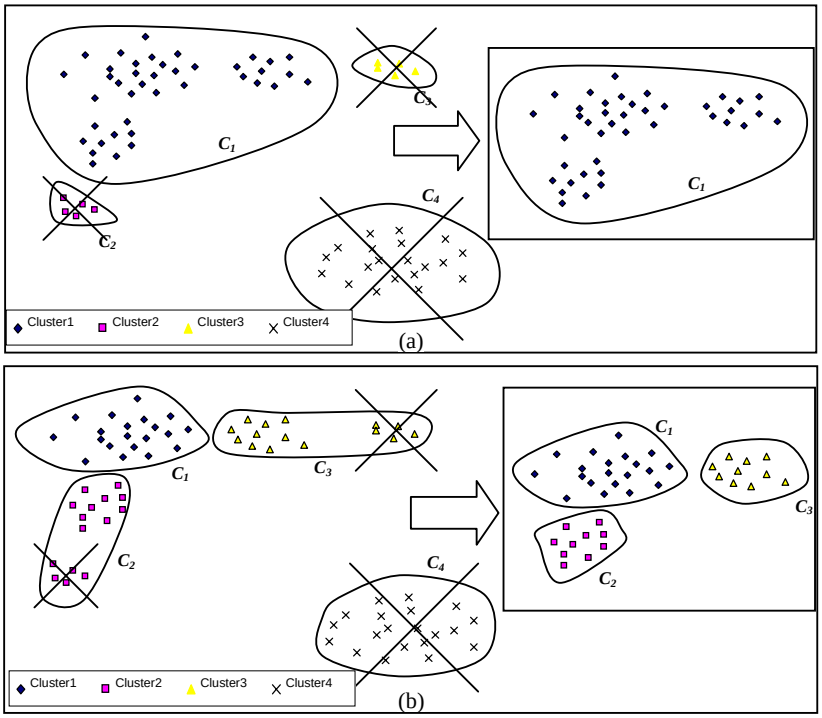
در اينجا يک روش جديد پيشنهاد شده است که تقارن را در روش اطلاعات متقابل نرمال شده از بين مي‌برد. فرض کنيد که مسئله ارزيابي معيار AMM براي خوشه C_1 در شکل ۳-۵(a) مي‌باشد.



شکل (۳-۵) محاسبه معيار *AMM* براي خوشه C_1 از افراز شکل (*a*)نسبت به افراز شکل (b)

[1] Alizadeh-Moshki-Minaei (AMM)

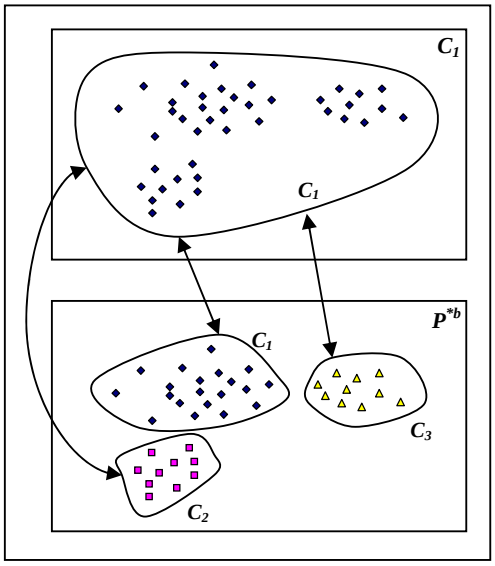
[2] Edited Normalized Mutual Information (ENMI)



شكل (۳-۶) روال پيشنهادي براي محاسبه معيار *AMM*. (*a*) حذف تمام خوشه‌هاي افراز P^a به جز C_1 (b) محاسبه P^{b*} از نمونه‌هاي متناظر با خوشه C_1 از P^a.

در اين روش سعي شده است تا تقارن موجود در روش اطلاعات متقابل از بين برود. براي اين كار همه خوشه‌ها در افراز P^a (شكل ۳-۵(a)) به جز خوشه C_1 كنار گذاشته مي‌شوند. همچنين، همه خوشه‌ها در افراز P^b (شكل ۳-۵(b)) كه شامل هيچ نمونه‌ي متناظري از اين خوشه نيستند، كنار گذاشته مي‌شوند. همچنين، خوشه‌هاي موجود در افراز P^b كه شامل تعدادي از نمونه‌هاي متناظر خوشه C_1 هستند، از وجود ديگر نمونه‌ها پاک‌سازي مي‌شوند. اين روال در شكل ۳-۶ نشان داده شده است.

در اين مرحله اطلاعات متقابل بين خوشه‌هاي باقيمانده از دو افراز P^a و P^b محاسبه مي‌شود. اين روال در شكل ۳-۷ نشان داده شده است. از آن جايي كه در اين محاسبات خوشه‌هاي غير از خوشه مورد نظر ناديده گرفته مي‌شوند، اين روش مشكل تقارن را نخواهد داشت.



شكل (۳-۷) محاسبه AMM بين خوشه C_1 از P^a و خوشه‌هاي باقيمانده از افراز P^b.

همه كارهاي گذشته براي محاسبه پايداري مبتني بر رابطه كلي اطلاعات متقابل نرمال شده مي‌باشند كه به صورت رابطه ۳-۶ آمده است.

(٦-٣)

$$NMI(P^a,P^b)=\frac{-2\sum_{i=1}^{k_a}\sum_{j=1}^{k_b}n_{ij}^{ab}\log\left(\frac{n_{ij}^{ab}.n}{n_i^a.n_j^b}\right)}{\sum_{i=1}^{k_a}n_i^a\log\left(\frac{n_i^a}{n}\right)+\sum_{j=1}^{k_b}n_j^b\log\left(\frac{n_j^b}{n}\right)}$$

كه n تعداد كل نمونه‌ها و n_{ij}^{ab} تعداد نمونه‌هاي مشترک بين خوشه‌هاي $C_i^a\in P^a$ و $C_j^b\in P^b$ مي‌باشد. در همه روش‌هاي گذشته براي ارزيابي يک خوشه، ابتدا مي‌بايست تا يک افراز از آن ايجاد شود تا بتوان از رابطه ۳-۶ استفاده كرد. تبديل يک خوشه به يک افراز خود مي‌توانست باعث بروز مشكلاتي شود، نظير مشكل تقارن كه در بخش‌هاي قبلي به طور مفصل تشريح شده است. در اينجا به جاي تغيير مسئله به صورتي كه بتوان از معيار اطلاعات متقابل نرمال شده استفاده كرد، معيار اطلاعات متقابل نرمال شده به گونه‌اي تغيير داده شده است، تا بتوان اطلاعات متقابل نامتقارن را بين يک خوشه و يک افراز به دست آورد (براي ارزيابي يک خوشه). با توجه به شكل ۳-۷ در اين معيار جديد، مقدار اطلاعات بين هر خوشه C_i از P^a با افرازهايي كه از آن خوشه تشكليل شده است، محاسبه مي‌شوند و

اين كار مستقل از نمونه‌هاي خوشه‌هاي ديگر است. يعني يک طرف فقط C_i را داريم و طرف ديگر خوشه‌هاي منتج شده از نمونه‌هاي اين خوشه در افرازديگر. بنابراين، n_{ij}^{ab} كه تعداد نمونه‌هاي مشترک بين خوشه‌هاي $C_i^a \in P^a$ و $C_j^b \in P^b$ در رابطه ۳-۶ مي‌باشد، به n_j^b در رابطه جديد (۳-۷) تغيير مي‌كند. همچنين k_a كه تعداد خوشه‌هاي افراز P^a است، برابر با تنها يک خوشه خواهد بود (يعني تنها همان خوشه C_i). بنابراين معيار AMM به صورت رابطه ۳-۷ بين هر خوشه C_i از P^a و افراز P^b به اين صورت به دست مي‌آيد.

(۳-۷)

$$AMM(C_i, P^{b*}) = \frac{-2\sum_{i=i}^{i}\sum_{j=1}^{k_{b*}} n_j^{b*} \log\left(\frac{n_j^{b*}.n}{n_i^a.n_j^{b*}}\right)}{\sum_{i=i}^{i} n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_{b*}} n_j^{b*} \log\left(\frac{n_j^{b*}}{n}\right)}$$

كه P^{b*} افراز نمونه‌هاي متناظر با خوشه C_i در خوشه‌بندي P^b است. با ساده سازی n_j^{b*} در *log* صورت و همچنين، حذف سيگماي داراي انديس i از صورت و مخرج (چون تنها به ازاي C_i اعمال مي‌شود) رابطه به شكل ۳-۸ تغيير خواهد كرد.

(۳-۸)

$$AMM(C_i, P^{b*}) = \frac{-2\sum_{j=1}^{k_{b*}} n_j^{b*} \log\left(\frac{n}{n_i^a}\right)}{n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_{b*}} n_j^{b*} \log\left(\frac{n_j^{b*}}{n}\right)}$$

از آنجايي كه *log* صورت در رابطه ۳-۸، مستقل از j مي‌باشد، مي‌تواند از داخل سيگما بيرون بيايد. بنابراين رابطه نهايي براي معيار اطلاعات نامتقارن AMM به صورت رابطه ۳-۹ خواهد بود.

(۳-۹)

$$AMM(C_i^a, P^{b*}) = \frac{-2\log\left(\frac{n}{n_i^a}\right)\sum_{j=1}^{k_{b*}} n_j^{b*}}{n_i^a \log\left(\frac{n_i^a}{n}\right) + \sum_{j=1}^{k_{b*}} n_j^{b*} \log\left(\frac{n_j^{b*}}{n}\right)}$$

كه n تعداد كل نمونه‌هاي مجموعه داده و n_{ij}^{ab} تعداد نمونه‌هاي مشترک بين خوشه‌هاي $C_i^a \in P^a$ و

$C_j^b \in P^{b*}$ مي‌باشد. همچنين، k_{b*} تعداد خوشه‌هاي موجود در افراز P^{b*} مي‌باشد.

در اينجا ميانگين AMM، AAMM[1] به عنوان ميزاني براي محاسبه پايداري يک خوشه پيشنهاد شده است. رابطه ۳-۱۰ پايداري خوشه C_i را با معيار AAMM نشان مي‌دهد.

$$AAMM(C_i) = \frac{1}{M}\sum_{j=1}^{M} AMM(C_i^a, P_j^{b*}) \qquad (۱۰-۳)$$

كه P_j^{b*} نشان‌دهنده j-امين افراز از مجموعه مرجع مي‌باشد.

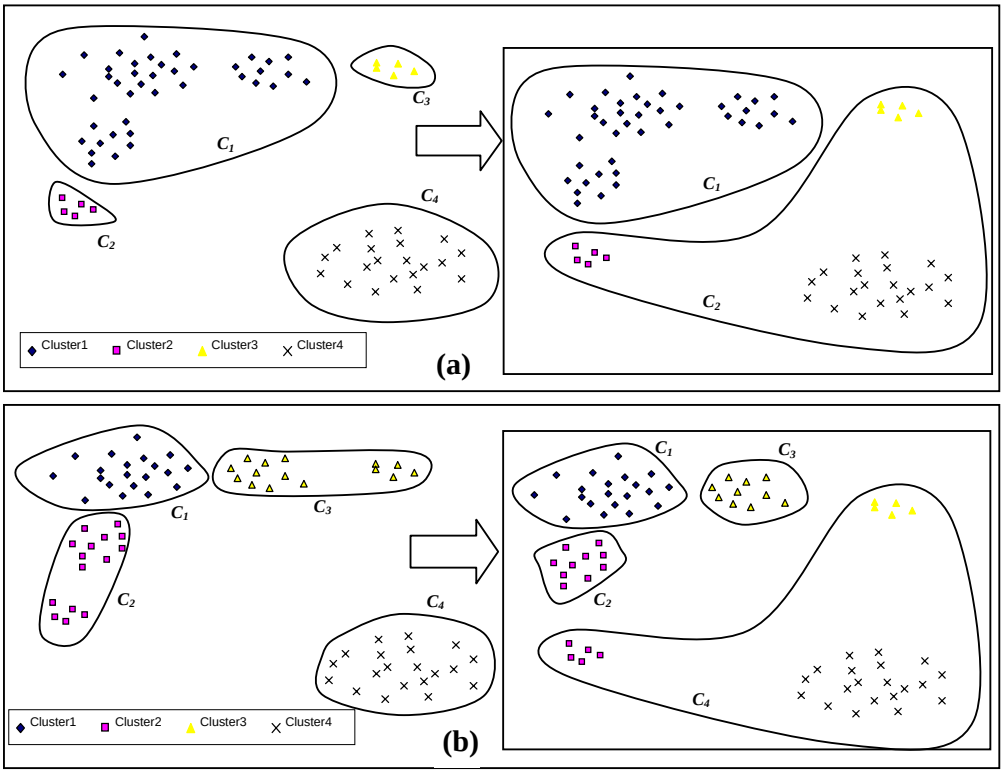
۳-۳-۴- اطلاعات متقابل نرمال اصلاح شده[2]

اين روش مشابه روش AMM مي‌باشد، با اين تفاوت كه در روش AMM خوشه‌هايي از خوشه‌بندي مرجع كه هيچ داده‌ي متناظري با C_1 نداشته باشند، در مرحله اول حذف مي‌شوند. در صورتي كه در اين روش هيچ نمونه‌اي از داده‌ها حذف نمي‌شود. به عبارت ديگر، مجموعه تمام نمونه‌هاي غيرمتناظر با نمونه‌هاي خوشه C_1 به عنوان يک خوشه ديگر در محاسبه ميزان پايداري شركت مي‌كنند. اين روال در شكل ۳-۸ به تصوير كشيده شده است.

۳-۴- انتخاب زيرمجموعه‌اي از خوشه‌هاي اوليه

پس از اينكه پايداري هر خوشه محاسبه شد، در گام بعد، عمل انتخاب خوشه‌ها با توجه به مقدار پايداري خوشه انجام مي‌شود. چندين روش انتخاب خوشه در اين پايان‌نامه پيشنهاد شده است: اعمال آستانه، انتخاب تطبيقي، تنظيم پراكندگي و كيفيت. هر كدام از اين روش‌ها به طور مفصل در اين بخش آمده‌اند.

[1] Average AMM
[2] Edited NMI



شكل (۳-۸) روال پيشنهادي براي محاسبه اطلاعات متقابل نرمال اصلاح شده. (a) ايجاد دو خوشه از افراز P^a (b) استخراج P^{b*}.

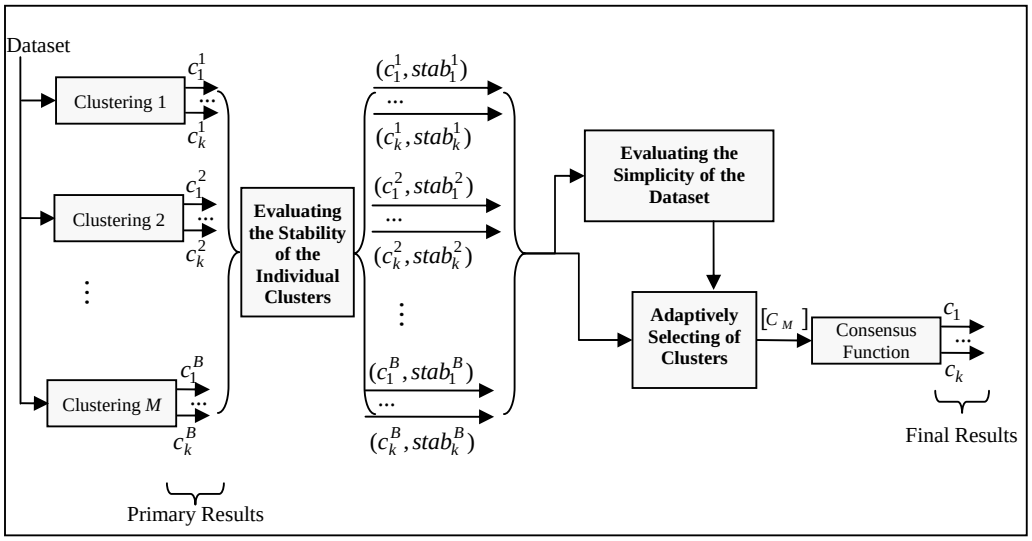
۳-۴-۲- اعمال آستانه

در اين روش خوشه‌هاي پايدارتر از خوشه‌بندي‌هاي اوليه شناسايي مي‌شوند و ماتريس همبستگي نهايي، تنها از اين خوشه‌هاي پايدار تشكيل مي‌شود. يک خوشه پايدار، خوشه‌اي است كه اگر آن روش خوشه‌بندي را چند بار ديگر هم، روي آن مجموعه داده (يا روي مجموعه‌هاي مختلف حاصل از نمونه‌برداري از آن مجموعه داده) اجرا كنيم، با احتمال زياد اين خوشه باز هم ديده خواهد شد. به عبارت ديگر، خوشه‌هاي پايدار به خوشه‌هايي اطلاق مي‌شود كه در خوشه‌بندي‌هاي مختلف روي زيرمجموعه‌هاي به دست آمده از نمونه‌برداري‌هاي مختلف بيشترين تكرار را داشته باشند. يعني با تغييرات جزيي در مجموعه داده، آن خوشه‌ها باز هم تكرار شوند. اين كار به سادگي با اعمال يک مقدار آستانه انجام مي‌شود. در اين تحقيق مقدار آستانه به گونه‌اي انتخاب شده است تا همواره ۳۳% از

نتايج خوشه‌هاي اوليه در ترکيب نهايي شرکت کنند.

۳-۴-۳- انتخاب تطبيقي

از آنجايي که ممکن است استفاده از خوشه‌هاي پايدارتر منجر به نتايج نهايي بهتر نشود، در اين بخش روش جديدي براي استفاده از زيرمجموعه‌اي از خوشه‌هاي اوليه پيشنهاد شده است که به صورت تطبيقي با مجموعه داده عمل مي‌کند. عمل انتخاب زيرمجموعه‌اي از خوشه‌ها در اين روش با توجه به مقدار پايداري خوشه‌ها و ميزان سادگي مجموعه داده به صورت تطبيقي صورت مي‌گيرد. اين بخش به صورت مفصل به چگونگي روش پيشنهادي براي محاسبه ميزان سادگي مجموعه داده مي‌پردازد. همچنين يک روال مشخص براي انتخاب زيرمجموعه‌اي از خوشه‌هاي اوليه با توجه به ميزان سادگي مجموعه داده پيشنهاد گرديده است. نماي کلي از روش پيشنهادي در شکل ۳-۹ نشان داده شده است.



شکل (۳-۹) روال پيشنهادي براي خوشه‌بندي ترکيبي با انتخاب تطبيقي خوشه‌ها.

◘ محاسبه ميزان سادگي مجموعه داده

پس از اينکه معيار پايداري براي تمام خوشه‌ها به دست آمد، براي محاسبه پايداري يک خوشه‌بندي مستقل P، رابطه زير تعريف مي‌شود.

(۱۱-۳)

$$Stability(P) = \frac{1}{N}\sum_{i=1}^{k}|C_i|Stability(C_i)$$

كه $|C_i|$ تعداد اعضاي خوشه i، N تعداد كل نمونه‌هاي موجود در اين افراز، و k تعداد خوشه‌هاست. رابطه فوق، يک رابطه وزن‌دار براي محاسبه پايداري يک خوشه‌بندي است. در گام بعدي، معيار سادگي براي مجموعه داده با استفاده از پايداري خوشه‌بندي‌هاي اوليه تعريف مي‌شود. رابطه ۳-۱۲ براي محاسبه ميزان سادگي مجموعه داده پيشنهاد شده است.

(۱۲-۳)

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

هر چه اين مقدار زيادتر باشد، نشان‌دهنده آن است كه مجموعه داده مورد بررسي، يک مجموعه داده راحتي بوده است، چرا كه خوشه‌هايي كه توسط الگوريتم‌هاي اوليه خوشه‌بندي پيدا مي‌شوند، داراي پايداري بالايي مي‌باشند و خوشه‌بندي‌هاي اوليه مي‌توانند جواب‌هاي نسبتا مناسبي ارايه دهند. برعكس، اگر مجموعه داده مورد بررسي، يک مجموعه داده سخت باشد، خوشه‌هاي ايجاد شده توسط الگوريتم‌هاي پايه خوشه‌بندي پايداري كمي خواهند داشت و در خوشه‌بندي‌هاي مختلف تغييرات زيادي از خود نشان خواهند داد. در نتيجه معيار سادگي مجموعه داده نيز مقدار كمي خواهد بود.

در اين دسته‌بندي مجموعه داده‌ها به سه رده سخت، متوسط و راحت تقسيم شده‌اند. در اين رده‌بندي، مجموعه داده‌هاي با مقدار سادگي كمتر از ۰,۵، در بازه [۰,۵-۰,۵۵] و بالاي ۰,۵۵ به ترتيب سخت، متوسط و راحت رده‌بندي شده‌اند. همچنين زيرمجموعه‌هاي مشخصي از نتايج خوشه‌بندي‌هاي اوليه براي هر رده از مجموعه داده‌ها پيشنهاد شده است. براي اين كار بعد از محاسبه معيار سادگي، نتايج همه M خوشه‌بندي اوليه بر مقدار پايداري‌شان مرتب مي‌شوند. سپس ۳ زيرمجموعه متفاوت از نتايج خوشه‌بندي‌هاي اوليه مورد بررسي قرار مي‌گيرند كه عبارتند از:

- زيرمجموعه اول: شامل ۳۳% از نتايج خوشه‌بندي‌هاي اوليه كه كمترين ميزان پايداري را دارا مي‌باشند.

- زيرمجموعه دوم: شامل ۳۳% از نتايج خوشه‌بندي‌هاي اوليه كه ميزان پايداري متوسطي را دارا مي‌باشند.
- زيرمجموعه سوم: شامل ۳۳% از نتايج خوشه‌بندي‌هاي اوليه كه بيشترين ميزان پايداري را دارا مي‌باشند.

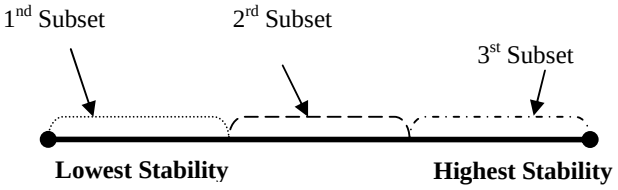
در گام بعدي با توجه به رده هر مجموعه داده يكي از اين زيرمجموعه‌ها پيشنهاد مي‌شود.

◘ نحوه تخصيص زيرمجموعه‌اي از خوشه‌ها براي مجموعه داده مشخص

انتخاب اينكه براي هر رده از مجموعه داده‌ها استفاده از كدام زيرمجموعه از خوشه‌هاي اوليه مناسب‌تر است و مي‌تواند منجر به بهبود عملكرد خوشه‌بندي تركيبي شود، يكي از مراحل حساس در روش پيشنهادي است. در واقع هدف اين است كه با توجه به مقدار سادگي مجموعه داده در مورد حضور يک زيرمجموعه مشخص از خوشه‌ها تصميم گرفته شود. يعني بايد تابع *f* با توجه به رابطه ۳-۱۳ مشخص شود.

$$f(\text{Simplicity of Dataset}) = \text{Special Subset of Clusters} \qquad (۱۳-۳)$$

براي اين منظور آزمايش‌ها به صورت زير طراحي شد. ابتدا تعداد ۱۵ مجموعه داده (كه برچسب واقعي نمونه‌هايشان در دسترس مي‌باشد) كه با توجه به رده‌بندي پيشنهادي جزو مجموعه داده‌هاي سخت قرار گرفتند، با هر كدام از سه زيرمجموعه از خوشه‌هاي اوليه امتحان شدند و جواب خوشه‌بندي تركيبي مشخص شد. سپس با مقايسه نتايج حاصل با برچسب واقعي نمونه‌ها، مقدار كارايي استفاده از هر زيرمجموعه (به صورت با ناظر) محاسبه شد. اين آزمايش نشان مي‌دهد كه رفتار مجموعه داده‌هاي سخت در مقابل انتخاب زيرمجموعه‌هاي مشخصي از خوشه‌ها چگونه است. اين كار به طور مشابه براي مجموعه داده‌هاي متوسط و راحت نيز انجام شد. مجموعه داده‌هاي مورد آزمايش طي فرآيندي به صورت دستي توليد شده‌اند. شكل ۳-۱۰ انتخاب اين سه زيرمجموعه از نتايج خوشه‌بندي‌هاي اوليه را براي مجموعه داده‌هاي مختلف نمايش مي‌دهد.



شكل (۳-۱۰) زيرمجموعه‌هاي انتخابي از خوشه‌هاي اوليه مرتب شده بر اساس مقدار پايداري

۳-۴-۴- تنظيم پراكندگي و كيفيت

نشان داده شده است كه هر چه كيفيت خوشه‌هاي اوليه بالاتر باشد، معمولا نتايج نهايي نيز از كيفيت بالاتري برخوردارند. همچنين، نشان داده شده است كه پراكندگي نتايج اوليه نيز به عنوان يک پارامتر مهم ديگر نقش مهمي در نتايج نهايي بازي مي‌كند. در اين بخش يک روش جديدي براي انتخاب خوشه‌ها پيشنهاد شده است كه از هر دو معيار كيفيت و پراكندگي استفاده مي‌كند. در اين روش سعي شده است تا يک موازنه منطقي از اين دو معيار حاصل شود. كيفيت خوشه‌هاي اوليه در اين روش، با استفاده از معيار AMM ارزيابي شده است. براي حصول به اين امر از چند راهكار استفاده شده است.

- راهكار اول

در اين روش، ابتدا خوشه‌هاي اوليه بر اساس كيفيت رتبه‌دهي مي‌شوند. در مرحله بعد به ترتيب از خوشه‌هاي بهتر، آنهايي انتخاب مي‌شوند كه ميزان پراكندگي‌شان با خوشه‌هاي انتخاب شده، بيشتر از يک مقدار آستانه باشد. شكل ۳-۱۱ الگوريتم روش پيشنهادي را نشان مي‌دهد.

مزيت اين روش اين است كه خوشه‌هاي شبيه هم زياد تكرار نخواهند شد. با اين روش خوشه‌هاي انتخابي ضمن اينكه خوشه‌هاي پايداري هستند(به خاطر رتبه‌دهي و انتخاب خوشه‌ها با شروع از پايدارترين خوشه)، پراكندگي آنها نسبت به هم نيز مقدار بالايي مي‌باشد.

تعين مقدار آستانه يكي از مشكلاتي است كه پيش روي اين روش وجود دارد و با توجه به بدون ناظر بودن مسئله خوشه‌بندي، استفاده از اين روش را در عمل محدود خواهد كرد. در اينجا مقدار آستانه به طور خودكار طوري تنظيم شده است كه همواره ۳۳% از خوشه‌هاي اوليه وارد تركيب نهايي شوند.

```
S := {};          //Subset of Selected Clusters
T                 //Sorted Total Primary Clusters
Initialization
  T := Sort total clusters according to AMM;
  S := {the most stable cluster};
For i := 2 to length (T)
  current :=  i-th cluster of T;
  similar := Find the most similar cluster in S with the current
  If distance(current,similar)>th
        Add current into S;
  End If;
End For;
Return S;
```

شكل (۳-۱۱) راهكار اول براي انتخاب خوشه‌هاي با كيفيت و پراكنده

◘ راهكار دوم

اين روش نيز كه الگوريتم آن در شكل ۳-۱۲ نشان داده شده است، مشابه روش قبل عمل مي‌كند.

```
S := {};          //Subset of Selected Clusters
T                 //Sorted Total Primary Clusters
Initialization
  T := Sort total clusters according to AMM;
  S := {the most stable cluster};
For i := 2 to length (T)
  current :=  i-th cluster of T;
  For j := 1 to length(S)
        temp := j-th cluster of S
        If similarity(current,temp)>th1
              Add temp into similar;
        End If;
  End For;
  If mean_distance(current,similar)>th2
        Add current into S;
  End If;
End For;
Return S;
```

شكل (۳-۱۲) راهكار دوم براي انتخاب خوشه‌هاي با كيفيت و پراكنده

تفاوت آنها در اين است كه در اينجا تعدادي از شبيه‌ترين‌ها را مورد بررسي قرار مي‌دهيم. به اين معني كه خوشه‌هايي از مجموعه S كه بيشتر از يک حدي به خوشه مورد نظر شبيه هستند را به عنوان مجموعه مشابه در نظر مي‌گيريم. سپس ميانگين فاصله خوشه مورد نظر را با اين مجموعه مي‌سنجيم. اگر اين فاصله از مقدار آستانه بيشتر بود، يعني پراكندگي لازم وجود دارد. در نتيجه، اين خوشه انتخاب مي‌شود.

- راهكار سوم

اگرچه مي‌توان با تنظيم پارامترها در دو راهكار اول و دوم به نتايج قابل توجهي دست يافت، ولي تنظيم اين پارامترها خود يكي از مشكلات اين روش‌ها مي‌باشد. در اين بخش روش ديگري پيشنهاد شده است كه مستقل از هر پارامتري معمولا منجر به نتايج قابل قبولي مي‌شود. در اين روش يک خوشه‌بندي روي خوشه‌هاي اوليه صورت مي‌گيرد. به عبارت ديگر، در اين روش، شباهت‌ها و تفاوت‌هاي خوشه‌هاي اوليه با خوشه‌بندي آنها به دست مي‌آيد. تعداد خوشه‌هاي اعمال شده براي اين خوشه‌بندي برابر با ۳۳% تعداد كل خوشه‌هاي اوليه در نظر گرفته شده است. فضاي ويژگي براي خوشه‌بندي خوشه‌ها شامل يک ماتريس $n\times d$ است كه n تعداد كل خوشه‌هاست و d تعداد كل نمونه‌ها مي‌باشد. مثلا اگر بخواهيم خوشه‌هاي اوليه حاصل از اجراي ۱۰۰ بار الگوريتم K-means (با k=3) را روي مجموعه داده Iris با ۱۵۰ نمونه داده خوشه‌بندي نماييم، اين فضا يک ماتريس ۱۵۰×۳۰۰ خواهد بود. در مرحله بعد پايدارترين خوشه از هر گروه براي شركت در ماتريس همبستگي و در نتيجه در تركيب نهايي انتخاب مي‌شود. همچنين اين نكته نيز در نظر گرفته شده است كه اگر گروهي شامل تنها يک خوشه باشد، به اين معني است كه آن خوشه يک نتيجه دورافتاده و پرت ارايه كرده است و در نتيجه از نتايج حذف مي‌شود. شكل ۳-۱۳ الگوريتم روش پيشنهادي را نشان مي‌دهد.

از ويژگي‌هاي اين روش اين است كه نياز به مرتب‌سازي كل خوشه‌هاي اوليه بر اساس پايداري نيست. همچنين، عدم وجود پارامتر اين روش را كاملا كاربردي و قابل استفاده مي‌نمايد. اين در حالي است كه از هر دو عامل پراكندگي و كيفيت خوشه‌ها در انتخاب خوشه‌ها استفاده شده است.

```
S := {};              //Subset of Selected Clusters
G := Apply a clustering technique over all primary clusters
For i := 1 to length(G)
  current := i-th group of G;
  If size(current)>1
        Add the most stable cluster of in the current into S;
  End If;
End For;
Return S;
```

شكل (۳-۱۳) راهكار سوم براي انتخاب خوشه‌هاي با كيفيت و پراكنده

۳-۵- ساخت ماتريس همبستگي از زيرمجموعه‌اي از خوشه‌ها

در اين مرحله خوشه‌هاي انتخاب شده با هم تركيب شده و خوشه‌هاي نهايي از آنها به دست مي‌آيد. روش‌هاي مختلفي براي تركيب خوشه‌بندي‌هاي اوليه و به دست آوردن خوشه‌هاي نهايي وجود دارد. تفاوتي كه در اينجا وجود دارد اين است كه در اين روش ممكن است كه از هر خوشه‌بندي اوليه، تنها تعدادي از خوشه‌ها موجود باشند. از آن جايي كه استفاده از روش انباشت مدارک (EAC) نمي‌تواند شباهت بين جفت نمونه‌ها را در حضور تنها تعدادي از خوشه‌ها به درستي تشخيص دهد، در اين پايان‌نامه دو روش جديد براي تركيب نتايج پيشنهاد شده است. اين روش‌ها كه توانايي استخراج ماتريس همبستگي براي نمونه‌ها -در شرايطي كه تنها تعدادي از خوشه‌ها موجود هستند- را دارند، در بخش‌هاي بعدي تشريح شده است. پس از ساخت ماتريس همبستگي، مي‌توان با استفاده از يكي از الگوريتم‌هاي سلسله‌مراتبي نظير اتصال منفرد يا اتصال ميانگين[1]، خوشه‌هاي نهايي را استخراج كرد.
اين نكته قابل يادآوري است كه در اينجا تنها يک ماتريس استخراج مي‌شود و خوشه‌هاي نهايي نيز به سادگي با اعمال يكي از الگوريتم‌هاي سلسله‌مراتبي از اين ماتريس به دست مي‌آيند.

[1] Single Linkage(SL) or Average Linkage(AL)

۳-۵-۱- انباشت مدارک توسعه یافته

در روش انباشت مدارک (EAC) نتايج m خوشه‌بندي روي داده‌هاي نمونه‌برداري شده در ماتريس همبستگي $n \times n$ ذخيره مي‌شوند. هر داده ورودي از اين ماتريس در روش انباشت مدارک، به صورت رابطه ۳-۱۴ محاسبه مي‌شود.

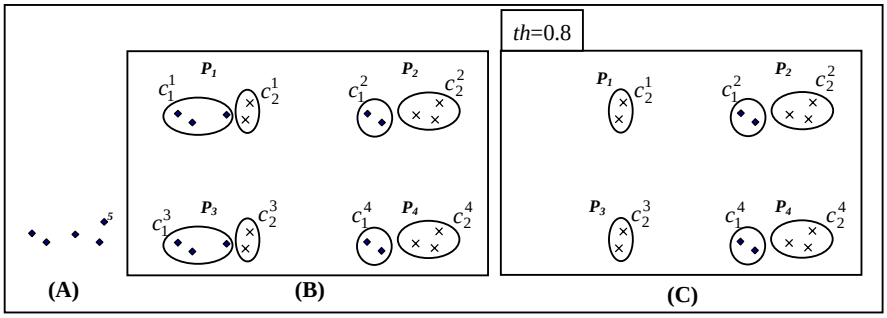
$$C(i, j) = \frac{n_{i,j}}{m_{i,j}} \qquad (۱٤-۳)$$

كه $n_{i,j}$ تعداد دفعاتي است كه جفت نمونه‌هاي i و j با هم در يک خوشه گروه‌بندي شده‌اند و $m_{i,j}$ تعداد نمونه‌برداري‌هايي است كه هر دوي اين جفت نمونه‌ها به طور همزمان در آن ظاهر شده‌اند.
از آن جايي كه پس از آستانه‌گيري در روش پيشنهادي، تنها تعدادي از خوشه‌هاي اوليه در دسترس مي‌باشند، روش EAC نمي‌تواند روابط بين جفت‌نمونه‌ها را تشخيص دهد. بنابراين براي تركيب نتايج با استفاده از ماتريس همبستگي بايد يک معياري براي نشان‌دادن همبستگي نمونه‌ها تعريف شود كه بتواند شباهت بين نمونه‌ها را با حضور تنها زيرمجموعه‌اي از خوشه‌هاي اوليه به درستي استخراج و محاسبه كند. ما روش خود را براي ساخت ماتريس همبستگي در شرايطي كه تعدادي از خوشه‌ها حذف شده‌اند، انباشت مدارک توسعه يافته (EEAC) مي‌ناميم. هر داده ورودي از ماتريس همبستگي در روش EEAC به صورت رابطه ۳-۱۵ تعريف مي‌شود.

$$C(i, j) = \frac{n_{i,j}}{\max(n_i, n_j)} \qquad (۱۵-۳)$$

n_i تعداد دفعاتي است كه نمونه i در خوشه‌هاي انتخاب شده ظاهر شده است. به طور مشابه n_j نيز، تعداد دفعاتي است كه نمونه j در خوشه‌هاي انتخاب شده ظاهر شده است. همچنين، $n_{i,j}$ تعداد دفعاتي است كه جفت نمونه‌هاي i و j با هم در يک خوشه از خوشه‌هاي انتخاب شده ظاهر شده‌اند. بديهي است كه با در نظر گرفتن تعداد خوشه‌هاي ثابت در خوشه‌بندي‌هاي اوليه همواره n_i و n_j كمتر از تعداد كل افرازهاي اوليه و همچنين، تعداد كل خوشه‌هاي ممكن مي‌باشد. يعني $n_i, n_j \le B \le k \times B$.

براي روشن‌تر شدن بحث، مثال زير را در نظر بگيريد. فرض كنيد ۵ نمونه مطابق شكل ۳-۱۴(A) وجود دارند كه چهار خوشه‌بندي اوليه P_1 تا P_4 به عنوان خوشه‌بندي‌هاي اوليه روي آن صورت گرفته است (شكل ۳-۱۴(B)).



شكل (۳-۱۴) محاسبه پايداري خوشه C_i با روش مبتني بر *NMI*. A) مجموعه داده شامل ۵ نمونه، B) نتايج چهار خوشه‌بندي اوليه C) خوشه‌هاي باقيمانده پس از آستانه‌گيري

همچنين، فرض كنيد كه مقادير پايداري خوشه‌هاي توليد شده به صورت زير باشند:

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

اگر مقدار آستانه براي انتخاب خوشه‌ها برابر با ۰,۸ باشد، خوشه‌هاي اول از افرازهاي اول و سوم حذف مي‌شوند. اكنون ماتريس همبستگي بايد با استفاده از بقيه خوشه‌ها درست شود. يعني شكل ۳-۱۴(C).

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

ماتريس همبستگي قبل و بعد از آستانه‌گيري به ترتيب، به صورت روابط ۳-۱۶ و ۳-۱۷ خواهد بود:

(۱۶-۳)

$$C_{before} = \begin{bmatrix} 1 & 1 & 0.5 & 0 & 0 \\ 1 & 1 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 \end{bmatrix}$$

در اين ماتريس نمونه سوم مي‌تواند با احتمال ۵۰% به هر كدام از دو خوشه بچسبد. اگر بتوان اطلاعات اضافه‌تري به اين ماتريس اضافه كرد، به گونه‌اي كه خوشه پايدارتر داراي وزن بيشتري در مقادير همبستگي بين نمونه‌هايش شود، مي‌توان به عملكرد بهتر خوشه‌بندي تركيبي اميدوار بود.

(۱۷-۳)

$$C_{after} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 1 & 1 \\ 0 & 0 & 0.5 & 1 & 1 \end{bmatrix}$$

با مشاهده اين دو ماتريس مي‌توان دريافت كه چگونه حذف خوشه‌هاي ناپايدار مي‌تواند باعث بهبود ماتريس همبستگي شود. از آن جايي كه خوشه مربوط به نمونه‌هاي {۱،۲،۳} داراي مقدار پايداري پاييني مي‌باشند، حذف آنها موجب روشن‌تر شدن ماتريس همبستگي مي‌شود. اكنون يک الگوريتم سلسله‌مراتبي ساده نيز مي‌تواند خوشه‌هاي موجود در ماتريس همبستگي (بعد از انتخاب) را استخراج كند.

پس از اينكه ماتريس همبستگي با روش EEAC ساخته شد، در مرحله بعد از يک تابع توافقي براي استخراج خوشه‌هاي نهايي از اين ماتريس استفاده مي‌شود. معمولا از يكي از الگوريتم‌هاي سلسله‌مراتبي براي استخراج خوشه‌هاي نهايي از ماتريس همبستگي استفاده مي‌شود. در اين مقاله از الگوريتم سلسله‌مراتبي اتصال منفرد[1] استفاده شده است.

[1] Single Linkage

۳-۵-۲- اشتراک به اجتماع[1]

اين روش نيز مشابه روش انباشت مدارک توسعه يافته عمل مي‌کند. ايده اصلي پشت اين روش شمارش تمام حالت‌هاي ممکني است که نمونه‌هاي i و j نسبت به هم در خوشه‌هاي انتخاب شده داشته‌اند. در واقع تفاوت اصلي اين دو روش در مخرج کسر مي‌باشد. هر داده ورودي از ماتريس همبستگي در روش اشتراک به اجتماع (ItoU) به صورت رابطه ۳-۱۸ تعريف مي‌شود.

$$C(i,j) = \frac{\bigcap(n_i, n_j)}{\bigcup(n_i, n_j)} = \frac{n_{i,j}}{n_i + n_j - n_{i,j}} \qquad (۱۸-۳)$$

n_i تعداد دفعاتي است که نمونه i در خوشه‌هاي انتخاب شده ظاهر شده است. به طور مشابه n_j نيز، تعداد دفعاتي است که نمونه j در خوشه‌هاي انتخاب شده ظاهر شده است. همچنين، $n_{i,j}$ تعداد دفعاتي است که جفت نمونه‌هاي i و j با هم در يک خوشه از خوشه‌هاي انتخاب شده ظاهر شده‌اند.

[1] Intersection to Union

فصل ۴:

نتایج و تفسیر آنها

۴-۱- مقدمه

در اين فصل نتايج تجربي براي ارزيابي روش پيشنهادي بـا اسـتفاده از چنـدين مجموعـه داده گـزارش شده است. مجموعه داده‌هاي مورد استفاده شامل دو نوع كلي مي‌باشـند. نـوع اول مجموعـه داده‌هـاي استاندارد يوسي‌آي[1] مي‌باشد كه تقريبا نتايج تمام مطالعات اخير دنيا در زمينه خوشه‌بندي با اسـتفاده از اين مجموعه داده‌ها گـزارش مـي‌شـوند. نـوع دوم مجموعـه داده‌هـاي مـورد اسـتفاده در ايـن تحقيـق، مجموعه داده‌هاي مصنوعي مي‌باشند كه صـرفا در راسـتاي همـين تحقيـق بـه صـورت دسـتي طراحـي شده‌اند. توصيف بيشتر در مورد مجموعه داده‌هاي استاندارد مورد استفاده در بخش ۴-۲ آمـده اسـت. همچنين، نتايج آزمايش‌ها كه نشان از كارايي نسبتا بالاي روش‌هاي پيشنهادي در مواجهـه بـا مجموعـه داده‌هاي مختلف مي‌باشد، در اين فصل ارايه شده است.

۴-۲- مجموعه داده

روش پيشنهادي بر روي ۱۰ مجموعه داده استاندارد و ۴۵ مجموعه داده مـصنوعي[2] (كـه بـه بـراي ايـن تحقيق طراحي شده‌اند) مورد آزمايش قرار گرفته است. براي انجـام آزمـايش‌هـا سـعي شـده اسـت تـا مجموعه داده‌ها از لحاظ تعداد كلاس‌ها، تعداد ويژگي‌ها و همچنين تعـداد نمونـه‌هـا از حـداكثر تنـوع برخوردار باشند تا نتايج آزمايش‌ها تـا حـد ممكـن داراي اسـتحكام و قابـل تعمـيم باشـد. جـدول ۴-۱ اطلاعات مختصري از مجموعه داده‌هاي استاندارد مورد استفاده را در اختيار مي‌گذارد.

از آنجـايي كـه مجموعـه داده‌هـاي Wine و Iris از معـروف‌تـرين و پركـاربردترين مجموعـه داده‌هـا مي‌باشند، توضيح مختصري از آنها به صـورت زيـر آمـده اسـت. همچنـين، بـه خـاطر مـصنوعي‌بـودن مجموعه داده نيم‌حلقه‌ها، اين مجموعه داده نيز توصيف شده است. براي اطلاعات بيـشتر در مـورد هـر

[1] UCI repository
[2] Artificial

كدام از مجموعه داده‌هاي جدول ۴-۱ (غير از نيم حلقه‌ها) مـي‌تـوان بـه (نيـومن[1] و همكـاران، ۱۹۹۸) رجوع كرد.

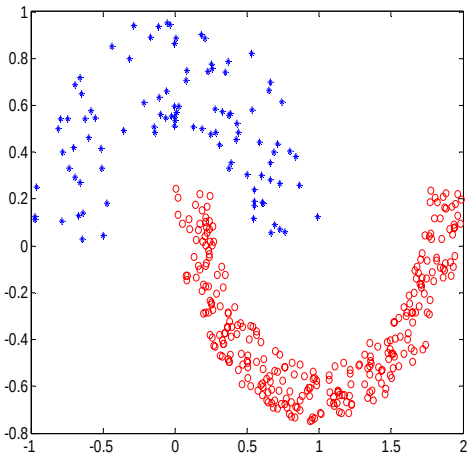
جدول (۴-۱) خلاصه‌اي از مشخصه‌هاي مجموعه داده‌هاي استاندارد مورد استفاده

	Class	Features	Samples
Glass	6	9	214
Breast-Cancer	2	9	683
Wine	3	13	178
Bupa	2	6	345
Yeast	10	8	1484
Iris	3	4	150
SAHeart	2	9	462
Ionosphere	2	34	351
Halfrings	2	2	400
Galaxy	7	4	323

مجموعه داده Wine كه در (ابرهارد[2] و همكاران، ۱۹۹۲) به طور مفـصل توصـيف شـده اسـت، شـامل مقدار تركيب شيميايي شراب‌هايي (آب انگور) است كه در يک محل مشابه، اما با سه نوع (سه نوع از موجودات ذره‌بيني خاكهاي زراعتي) مختلف درست شده‌اند. الگوها با سيزده جزء سازنده (ويژگي‌ها) توصيف شده‌اند كه در هر سه نوع مشروب وجـود دارنـد. در مجمـوع ۱۷۸ نمونـه وجـود دارد كـه در دسته‌هاي ۵۹تايي، ۷۱تايي و ۴۸تايي قرار گرفته‌اند. مجموعه داده Iris شـامل ۱۵۰ نمونـه در ۳ كـلاس ۵۰ تايي است كه هر كلاس به يک نوع گياه زنبق و سوسن رجوع مي‌كنـد. يـک كـلاس بـه صـورت خطي از بقيه دو كلاس تفكيک‌پذير است، و هر نمونه چهار ويژگي با مقدار پيوسته دارد.

مجموعه داده نيم‌حلقه‌ها كه يک مجموعه داده مصنوعي مي‌باشد -همـانطور كـه در شـكل ۴-۱ نـشان داده شده است- شامل دو خوشه مي‌باشـد كـه خوشـه‌هـا بـا ۱۰۰ نقطـه و ۳۰۰ نقطـه، نـامتوازن هـستند. الگوريتم K-means به تنهايي قـادر بـه تـشخيص دو خوشـه طبيعـي نيـست، بـه خـاطر اينكـه صـورت پيش‌فرض خوشه‌ها را به شكل ابركره تصور مي‌كند.

[1] Newman
[2] Aeberhard



شكل (۴-۱) مجموعه داده "نيم‌حلقه‌ها" با ٤٠٠ الگو (۱۰۰-۳۰۰ در هر كلاس)

تعدادي از نتايج آزمايش‌ها بر روي ويژگي‌هاي نرمال شده از اين مجموعه داده‌ها گزارش شده است. براي عمليات نرمال‌سازي، هر كدام از ويژگي‌هاي اين مجموعه داده‌ها با ميانگين صفر و واريانس يک، N(0,1)، نرمال شده‌اند.

براي همه اين مجموعه داده‌ها، تعداد خوشه‌ها و برچسب واقعي نمونه‌ها از قبل معلوم هستند. بنابراين، درصد نمونه‌هايي كه درست تشخيص داده شده‌اند، به عنوان معيار كارايي روش خوشه‌بندي مورد استفاده قرار گرفته است. در واقع بعد از حل مسئله تناظر بين برچسب‌هاي به دست آمده و خوشه‌هاي واقعي، مي‌توان نرخ خطا را تعيين كرد.

۴-۳- نتايج آزمايش‌ها

در اين بخش نتايج به كارگيري روش پيشنهادي روي مجموعه داده‌هاي مختلف و پارامترهاي مورد استفاده گزارش شده است. روش پيشنهادي در محيط MATLAB (ver 7.1) پياده‌سازي و مورد آزمايش قرار گرفته است. نتايج آزمايش‌ها روي ميانگين ۱۰ بار اجراي مستقل برنامه گزارش شده

است. عملكرد روش‌هاي مختلف خوشه‌بندي با استفاده از فرايند بازبرچسب‌گذاري[1] بين خوشه‌هاي به دست آمده و كلاس‌هاي واقعي و مقايسه آنها محاسبه شده است.

در تمامي روش‌هاي مورد استفاده از الگوريتم K-means به عنوان الگوريتم پايه استفاده شده است. تعداد نتايج اوليه توليد شده نيز در تمام روش‌ها ثابت و برابر با ۱۲۰ مي‌باشد. در واقع اين تعداد با دستكاري پارامتر k از الگوريتم K-means به دست آمده است. به اين صورت كه چهار گروه ۳۰ تايي از نتايج اوليه، با در نظر گرفتن تعداد خوشه‌هاي مورد استفاده توسط اين الگوريتم با اندازه‌هاي k، k+1، k+2 و k+3 حاصل شده است. همچنين، براي ايجاد پراكندگي بيشتر در نتايج اوليه از نمونه‌برداري بدون جاگذاري با نرخ ۵۰% استفاده شده است. همچنين، براي استخراج نتايج نهايي از ماتريس همبستگي از روش اتصال منفرد[2] استفاده شده است.

۴-۳-۱- اعمال آستانه براي انتخاب خوشه

در روش اعمال آستانه براي انتخاب خوشه، همواره ۳۳% بالاي نتايج (نتايج با كيفيت بالاتر) با عمل آستانه‌گيري انتخاب مي‌شوند. در اين روش، كيفيت خوشه‌هاي اوليه مهم‌ترين عاملي است كه باعث بهبود عملكرد خوشه‌بندي تركيبي مي‌شود. ۸ سطر اول از جدول ۴-۲ نتايج روش‌هاي پيشنهادي را روي مجموعه داده‌هاي استاندارد با اعمال چنين آستانه‌اي نشان مي‌دهد. دو سطر آخر از اين جدول نتايج روش‌هاي خوشه‌بندي انباشت مدارک (فرد و جين، ۲۰۰۵) و روش تطبيقي مطرح شده توسط عظيمي (۱۳۸۶) مي‌باشند. بهترين نتايج در هر ستون به صورت ضخيم نشان داده شده است. مقايسه نتايج با روش‌هاي سايرين نشان مي‌دهد كه روش‌هاي پيشنهادي در همه موارد بهتر از روش‌هاي پيشين عمل مي‌كنند.

لازم به ذكر است كه مجموعه داده‌هايي كه به صورت نرمال مورد استفاده قرار گرفته‌اند، در جدول

[1] Relabling
[2] Single Linkage

۲-۴ علامت .N[1] قبل از نام آن مجموعه داده آمده است.

جدول (۴-۲) نتايج روش‌هاي پيشنهادي روي مجموعه داده‌هاي استاندارد مورد آزمايش

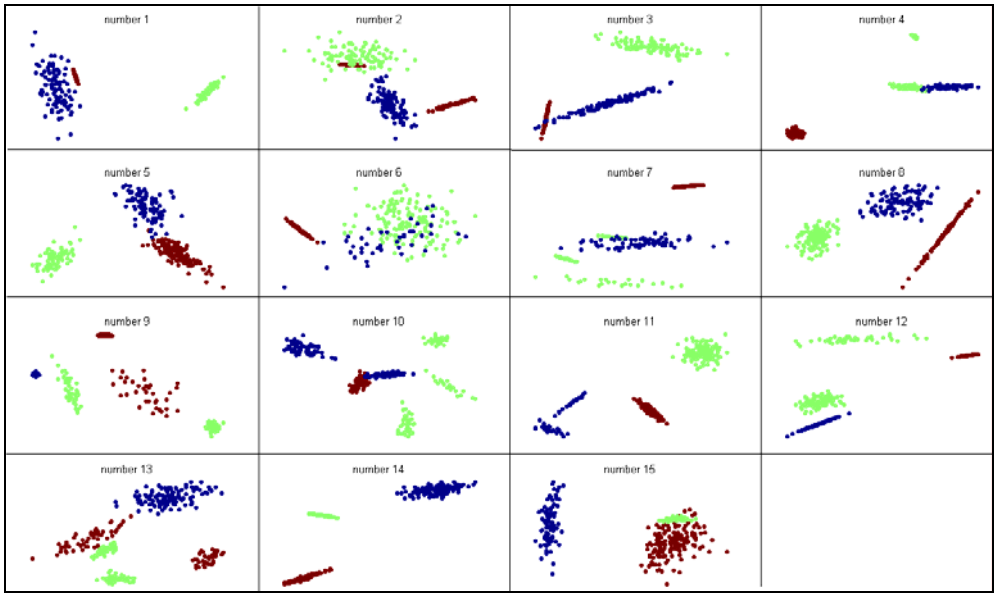
روش ارزيابي خوشه	روش ساخت ماتريس همبستگي	مجموعه داده‌هاي استاندارد										
		N. Breast Cancer	Iris	N. Bupa	N. SAHeart	Ionosphere	N. Glass	Halfrings	N. Galaxy	N. Yeast	Wine	N. Wine
NMI	ItoU	95.02	88.67	54.78	63.42	70.09	44.86	74.50	29.41	42.86	70.22	96.63
	EEAC	95.73	76.13	54.33	63.36	70.60	**47.76**	74.48	31.27	42.93	69.38	85.17
MAX	ItoU	96.93	**90.00**	54.78	**64.50**	**71.51**	44.86	87.25	29.41	48.45	71.35	97.75
	EEAC	96.49	84.87	**57.42**	63.87	57.75	44.35	74.55	29.85	**51.27**	70.00	94.44
AMM	ItoU	95.43	88.00	54.73	63.42	**71.51**	44.39	74.50	29.69	48.52	70.73	96.63
	EEAC	95.46	**90.00**	55.07	63.85	70.66	45.79	54.00	30.65	53.10	70.23	96.63
ENMI	ItoU	96.78	**90.00**	55.07	**64.50**	**71.51**	45.79	**88.25**	30.03	50.47	70.23	**98.32**
	EEAC	96.93	88.67	54.78	63.20	71.23	43.93	**88.00**	30.65	50.47	70.23	97.19
D&Q	1	97.66	97.33	55.36	68.83	72.93	50.47	87.25	35.29	56.67	72.47	98.31
	2	97.07	91.33	55.36	68.02	74.36	53.74	76.50	33.44	54.33	71.35	98.31
	3	**97.05**	**90.00**	55.00	63.83	70.91	47.20	83.25	**31.36**	50.18	**71.42**	97.75
Adaptive		95.43	88.00	54.73	63.42	**71.51**	44.39	74.50	29.69	48.52	70.73	96.63
EAC (Full Ens.)		95.17	89.33	54.49	63.20	70.66	46.26	74.50	30.96	44.21	70.22	96.63
Azimi		96.91	89.33	54.75	56.06	70.74	45.05	67.70	29.97	43.40	60.95	96.63

۴-۳-۲- روش تطبيقي براي انتخاب خوشه

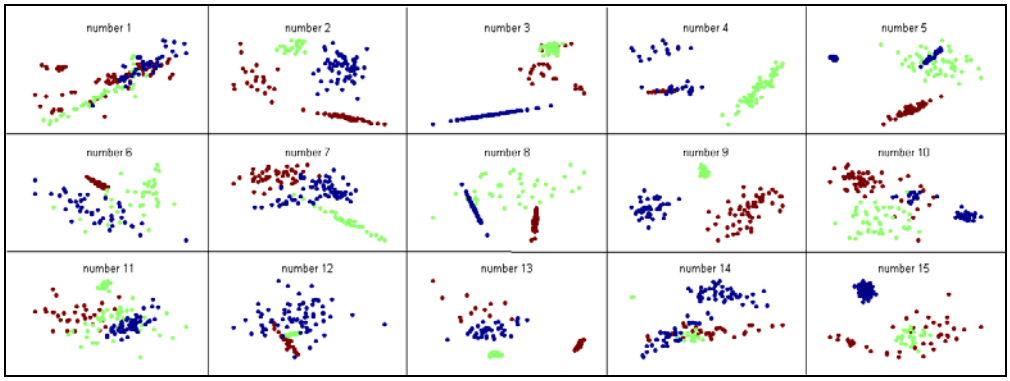
در اين روش، سعي بر آن است كـه رابطـه‌اي بـين ميـزان سـادگي مجموعـه داده و زيـرمجموعـه‌اي از خوشه‌ها كه استفاده از آنها در تركيب نتيجه بهتري مي‌دهد، پيدا كنيم. بدين منظور، تعدادي مجموعـه داده به صورت خودكار توليد شد. توليد كننده‌ي خودكار مجموعه داده، يـک عـدد بـه عنـوان ميـزان سادگي مورد انتظار مي‌گيرد و بر طبق آن مجموعه داده‌اي توليد مي‌كند كـه سـادگي آن كـم و بـيش مانند آنچه انتظار مي‌رود مي‌باشد. تمـام مجموعـه داده‌هـا بـه صـورت سـه رده‌اي توليـد شـده‌انـد. هـر مجموعه داده شامل ۳۰۰ داده مي‌باشد كه به صورت تصادفي در اين سه رده توزيع مي‌شوند. براي هـر رده ابتدا يک ميانگين در نظر گرفته مي‌شود. سپس، داده‌ها به صورت تـصادفي بـا اسـتفاده از يکـي از توزيع‌هاي نرمال گوسين، توزيع k، توزيع يكنواخت در يک بعد(توليد خوشه ميله‌اي) و يـا در دو بعـد

[1] Normal

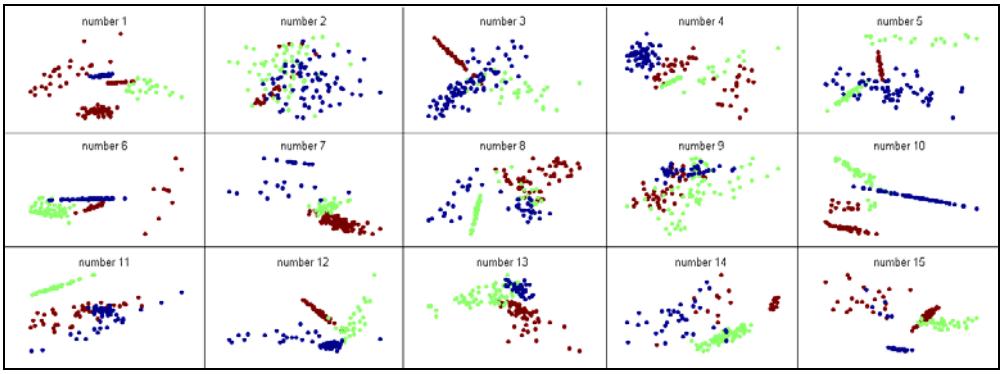
خوشه‌ها را تشكليل مي‌دهند. شكل‌هـاي ۴-۲، ۴-۳ و ۴-۴ بـه ترتيـب ۱۵ مجموعـه داده‌ي راحـت، ۱۵ مجموعه داده‌ي متوسط و ۱۵ مجموعه داده‌ي سخت را نشان مي‌دهند. بـه ازاي هـر مجموعـه داده، هـر بار يک زيرمجموعه از خوشه‌ها براي تركيب نتايج خوشه‌بندي استفاده شد و دقت به دست آمده ثبـت گرديد (منظور از دقت، همان ميزان تطابق نتايج خوشه‌بندي با برچسب‌هاي واقعي است).



شكل (۴-۲) مجموعه داده با ميانگين ميزان سادگي ۰,۵۹ از رده راحت



شكل (۴-۳) مجموعه داده با ميانگين ميزان سادگي ۰,۵۲ از رده متوسط



شكل (۴-۴) مجموعه داده با ميانگين ميزان سادگي ۰,۴۸ از رده سخت

براي ۱۵ مجموعه داده‌ي راحت، ميـانگين دقـت بـه ازاي سـه زيرمجموعـه تفـاوت معنـاداري را نـشان نمي‌دهد. همان طور كه در جدول ٤-٣ مشاهده مي‌شود، ميانگين دقت به ازاي سه زير مجموعه‌ي اول تا سوم به ترتيب۷۹,٦۳، ۷۹,۱٤ و ۷۹,۸٦ است كه بسيار نزديک به هم هستند. به عنوان يک نتيجه كلي مي‌توان گفت «براي مجموعه داده‌هـاي راحـت اگرچـه اسـتفاده از زيرمجموعـه‌هـاي مختلـف تفـاوت زيادي نمي‌كند، بهتر است از زيرمجموعه‌ي سوم استفاده شود».

جدول (۴-۳) نتايج آزمايش‌هاي روش تطبيقي براي مجموعه داده راحت

شماره مجموعه داده	ميزان سادگي	نرخ تشخيص با استفاده از زيرمجموعه اول	نرخ تشخيص با استفاده از زيرمجموعه دوم	نرخ تشخيص با استفاده از زيرمجموعه سوم
1	0.56	94.70	97.67	96.47
2	0.61	61.53	81.27	58.67
3	0.59	88.83	95.27	92.67
4	0.60	69.13	73.00	85.27
5	0.56	88.53	98.00	98.33
6	0.68	81.60	87.33	87.33
7	0.53	77.90	68.57	76.00
8	0.58	90.40	83.03	95.57
9	0.63	72.67	42.33	50.23
10	0.62	60.30	54.67	68.33
11	0.59	81.33	84.20	63.00
12	0.58	76.07	73.73	84.70
13	0.59	77.43	74.43	72.00
14	0.58	84.67	84.67	84.00
15	0.53	89.33	88.90	85.30
Average	0.59	79.63	79.14	**79.86**

براي ۱۵ مجموعه داده‌ي متوسط، ميانگين دقت به ازاي سه زير مجموعه بـا هـم متفـاوت اسـت. همـان طور كه در جدول ٤-٤ مشاهده مي‌شود، ميانگين دقـت بـه ازاي سـه زيـر مجموعـه‌ي اول تـا سـوم بـه ترتيب٦٩,٢٣، ٧٤,٢٩ و ٧٧,٥١ است. در اينجا به عنوان يک نتيجه كلي مي‌توان گفت «بـراي مجموعـه داده‌هاي متوسط بهتر است از زيرمجموعه‌ي سوم استفاده شود».

جدول (۴-۴) نتايج آزمايش‌هاي روش تطبيقي براي مجموعه داده متوسط

شماره مجموعه داده	ميزان سادگي	نرخ تشخيص با استفاده از زيرمجموعه اول	نرخ تشخيص با استفاده از زيرمجموعه دوم	نرخ تشخيص با استفاده از زيرمجموعه سوم
1	0.48	54.13	49.60	49.73
2	0.56	59.80	55.47	65.73
3	0.51	87.73	86.47	79.40
4	0.51	55.07	68.60	64.27
5	0.55	74.73	90.00	90.73
6	0.53	79.67	67.73	79.73
7	0.55	57.13	91.93	81.27
8	0.54	72.07	89.20	89.33
9	0.57	100.00	93.87	100.00
10	0.54	66.33	75.20	92.67
11	0.40	41.13	63.73	63.93
12	0.44	72.47	63.20	68.80
13	0.57	92.00	92.00	92.00
14	0.53	59.53	64.00	64.33
15	0.53	66.67	63.33	80.67
Average	0.52	69.23	74.29	**77.51**

براي ۱۵ مجموعه داده‌ي سخت، ميانگين دقت به ازاي سه زير مجموعه با هم متفاوت است. همان طور كـه در جـدول ٤-٥ مـشاهده مـي‌شـود، ميـانگين دقـت بـه ازاي سـه زيـر مجموعـه‌ي اول تـا سـوم بـه ترتيب٥٩,٤٠، ٦٨,٣٦ و ٦٨,٥٨ است. به عنوان يک نتيجه كلي مي‌توان گفت «براي مجموعه داده‌هـاي سخت مي‌توان از دو زير مجموعه‌ي دوم و سوم استفاده شـود، امـا بهتـر اسـت از زيرمجموعـه‌ي سـوم استفاده شود».

جدول (۴-۵) نتايج آزمايش‌هاي روش تطبيقي براي مجموعه داده سخت

شماره مجموعه داده	ميزان سادگي	نرخ تشخيص با استفاده از زيرمجموعه اول	نرخ تشخيص با استفاده از زيرمجموعه دوم	نرخ تشخيص با استفاده از زيرمجموعه سوم
1	0.36	50.93	60.47	55.40
2	0.51	54.73	62.80	56.40
3	0.50	70.87	77.67	74.13
4	0.49	56.07	64.07	66.07
5	0.49	46.67	77.07	77.27
6	0.44	72.47	63.20	68.80
7	0.40	41.13	63.73	63.93
8	0.47	57.53	54.33	47.93
9	0.48	54.13	49.60	49.73
10	0.55	58.53	72.00	72.00
11	0.48	68.93	63.73	82.80
12	0.51	46.33	75.87	68.00
13	0.52	68.60	83.07	88.00
14	0.52	84.73	89.47	89.67
15	0.36	50.93	60.47	55.40
Average	0.48	59.40	68.36	**68.58**

با توجه به نتايج دست آمده، همواره استفاده از زيرمجموعه‌ي سوم نتيجه بهتري خواهد داد. از آنجايي كه ميزان سادگي مجموعه داده نقشي در تعيين زيرمجموعه‌ي مناسب ندارد، بنابر اين نتيجه استفاده از روش تطبيقي مشابه با روش آستانه‌گيري خواهد بود. اين نتايج در سطر دوازدهم از جدول ۴-۲ آمده است.

لازم به ذكر است كه در اين روش، از معيار AMM براي ارزيابي خوشه‌هاي اوليه و تعيين ميزان سادگي مجموعه داده استفاده شده است. همچنين، روش اشتراک به اجتماع نيز براي ساخت ماتريس همبستگي به كار گرفته شده است.

۴-۳-۳- روش تنظيم پراكندگي و كيفيت

نتايج اجراي اين روش با راهكارهاي اول تا سوم به ترتيب درسطرهاي نهم تا يازدهم از جدول ۴-۲ آمده است. عنوان اين روش به صورت D&Q[1] در اين جدول نشان داده شده است. از آنجايي كه مقادير آستانه براي راهكارهاي اول و دوم به صورت دستي براي هر مجموعه داده انتخاب مي‌شود،

[1] Diversity & Quality

اين روش‌ها منجر به بهبود چشمگيري در نتـايج مـي‌شـوند. اگرچـه نتـايج ايـن دو روش بـسيار بـا بقيـه متفاوت مي‌باشد، اتكا به اين نتايج كار چندان درستي به نظر نمي‌رسد. تنظيم خودكار پارامترهـاي ايـن دو راهكار (مقادير آستانه) مي‌تواند از جمله كارهاي آينده در اين زمينه باشد.

راهكار سوم كه از خوشه‌بندي خوشه‌هاي اوليه استفاده مي‌كنـد، از هـيچ پـارامتري اسـتفاده نمـي‌كنـد؛ بنـابراين نتـايج كـاملا قابـل اتكـايي توليـد مـي‌كنـد. اگرچـه توليـد نتـايج اوليـه در ايـن روش بـا نـرخ نمونه‌برداري ۳۰% انجام شده است (برخلاف همه روش‌هاي ديگر كه با نرخ ۵۰% نمونه‌برداري را انجام مي‌دهند)، اين نتايج در مقايسه با ساير روش‌ها، در اكثر مواقع چـشمگير و عـالي اسـت. بـه طـور كلـي مي‌توان اين روش را بهترين روش در بين تمام روش‌هاي پيشنهادي قلمداد كرد.

لازم به ذكر است كه معيار ارزيابي خوشه در روش تنظيم پراكندگي و كيفيت، معيار AMM مي‌باشد. همچنين، از روش اشتراک به اجتماع براي ساخت ماتريس همبستگي استفاده شده است.

فصل ۵:

جمع‌بندي و كارهاي آينده

۵-۱- جمع‌بندي

در اين پايان‌نامه چند روش براي بهبود كارايي خوشه‌بندي تركيبي پيشنهاد شد. اين روش‌ها مبتني بر به‌كارگيري زيرمجموعه‌اي از نتايج اوليه مي‌باشند. براي انتخاب زيرمجموعه‌اي از نتايج اوليه، نياز به تعريف معيار ارزيابي مي‌باشد كه در اين تحقيق چند معيار براي ارزيابي خوشه‌ها پيشنهاد شده و مورد آزمايش قرار گرفته‌اند. در واقع چهارچوب پيشنهادي براي خوشه‌بندي تركيبي شامل دو مرحله اساسي است كه مرحله اول از آن ارزيابي خوشه‌هاي اوليه مي‌باشد.

معيار اطلاعات متقابل نرمال شده، متداول‌ترين معياري است كه در سال‌هاي اخير براي مقايسه دو افراز مختلف استفاده شده است. در اين تحقيق از اين معيار به گونه‌اي كه در بخش ۳-۳-۱ اشاره شد، براي ارزيابي كيفيت (پايداري) خوشه‌ها استفاده شده است.

همچنين، در ۳-۳-۲ نشان داده شد كه حالت‌هايي نيز وجود دارند كه روش اطلاعات متقابل نرمال شده به درستي نمي‌تواند خوشه‌ها را مورد ارزيابي قرار دهد. بنابراين، چند روش جديد براي ارزيابي خوشه پيشنهاد شد كه كمبود روش اطلاعات متقابل را نيز جبران مي‌كنند. اين روش‌ها شامل روش ماكزيمم، استفاده از معيار AMM، و اطلاعات متقابل نرمال اصلاح شده مي‌باشند. هر كدام از اين روش‌ها با جزييات كامل به ترتيب در بخش‌هاي ۳-۳-۲، ۳-۳-۳ و ۳-۳-۴ تشريح شده‌اند.

دومين مرحله اساسي از چهارچوب پيشنهادي براي خوشه‌بندي تركيبي ارايه روالي براي انتخاب زيرمجموعه‌اي از خوشه‌ها با در دست داشتن مقدار پايداري هر كدام از خوشه‌ها مي‌باشد. براي اين مرحله چند رويكرد پيشنهاد شده است. اعمال آستانه و انتخاب بهترين نتايج اوليه ساده‌ترين و در عين حال يک روش موثر براي انتخاب خوشه‌هاي با كيفيت براي شركت در تركيب نهايي مي‌باشد. اين روش با مرتب‌سازي خوشه‌ها بر اساس معيار پايداري مورد نظر و اعمال آستانه انجام شده است.

انتخاب تطبيقي روش ديگري است كه در اين تحقيق براي انتخاب زيرمجموعه‌اي از خوشه‌هاي اوليه پيشنهاد شده است. در اين روش، ابتدا ميزان سادگي مجموعه داده تعيين مي‌شود. سپس، با استفاده از ايجاد چند مجموعه داده مصنوعي و تعيين زيرمجموعه‌اي با بهترين كارايي براي هر رده از مجموعه

داده‌ها به يک نتيجه منطقي در مورد اختصاص يک زيرمجموعه خاص به ازاي هر مجموعه داده مي‌رسيم.

همچنين، تنظيم پراكندگي و كيفيت روش ديگري است كه در اينجا براي انتخاب خوشه‌ها پيشنهاد شده است. در اين روش سعي شده است تا هر دو عامل پراكندگي و كيفيت نتايج اوليه مورد توجه قرار گيرند. براي اين كار از سه راهكار مختلف استفاده شده است. در اين روش‌ها تنها كيفيت مورد نظر نمي‌باشد و يک خوشه با كيفيت بالا تنها زماني اجازه شركت در تركيب نهايي را خواهد داشت كه خوشه‌هاي مشابه آن در تركيب نبوده باشند. با اين كار از افزونگي خوشه‌هاي كاملا مشابه در تركيب جلوگيري خواهد شد. همچنين راهكار سوم از اين روش نسبت به داده‌هاي دورافتاده نسبتا مقاوم است.

نتايج تجربي انجام شده بر روي ۱۰ مجموعه داده استاندارد نشان از كارايي بالاي روش‌هاي پيشنهادي در مقايسه با ساير روش‌هاي خوشه‌بندي تركيبي دارد. همچنين، نتايج نشان مي‌دهند كه اگرچه در روش‌هاي پيشنهادي تنها ۳۳% از نتايج اوليه در تركيب نهايي استفاده مي‌شوند، كارايي اين روش‌ها حتي از روش تركيب كامل[1] هم بالاتر است. به طور دقيق‌تر روش‌هاي اطلاعات نرمال اصلاح شده (ENMI) در تركيب با روش اشتراک به اجتماع براي ساخت ماتريس همبستگي، بيشترين بهبود د رنتايج را ايجاد مي‌كند. مقدار متوسط اين بهبود روي كليه ۱۱ مجموعه آزمايش‌شده نسبت به روش‌هاي تركيب كامل و روش عظيمي به ترتيب برابر است با ۲,۳% و ۴,۵%. همچنين، نتايج روش تنظيم پراكندگي و كيفيت (D&Q) با راهكار سوم كه از خوشه‌بندي خوشه‌هاي اوليه استفاده مي‌كند، نزديک به روش فوق مي‌باشد. مقدار متوسط بهبودهای اين روش نسبت به دو روش تركيب كامل و روش عظيمی نيز به ترتيب برابر است با ۲٪ و ۴,۲٪.

۵-۲- كارهاي آينده

از جمله كارهاي آينده‌اي كه مي‌توان انجام داد، ارايه راهكاري خودكار براي تعيين بهينه مقدار آستانه

[1] Full Ensemble

در روش‌هاي پيشنهادي مي‌باشد. اگرچه در اين تحقيق اين مقدار برابر با انتخاب ۳۳% از جواب‌هاي با مقدار پايداري بالا انتخاب شده است، شايد بتوان با تغيير و بهينه‌سازي اين مقدار باعث بهبود بيشتر كارايي خوشه‌بندي تركيبي پيشنهادي شد.

همچنين، تعيين خودكار پارامترها در راهكارهاي اول و دوم از روش تنظيم پراكندگي و كيفيت مي‌تواند اين روش‌ها را به روش‌هاي بسيار پرقدرت در خوشه‌بندي تركيبي بدل كند.

از جمله كارهاي ديگري كه مي‌توان در ادامه اين تحقيق به آن پرداخت، جستجوي بيشتر براي ارايه معيارهاي مناسب براي ارزيابي يک خوشه و يا يک افراز از داده‌ها مي‌باشد. اگر بتوان يک معيار يا يک مجموعه از معيارهاي دقيق براي ارزيابي برازندگي خوشه و افراز يافت كه مستقل از نوع خوشه‌ها و دادگان عمل كند، مي‌توان با استفاده از آنها، از روش‌هاي جستجوي هوشمند نيز بهره گرفت. با داشتن يک معيار برازندگي مستقل، مي‌توان با استفاده از الگوريتم‌هاي تكاملي نظير الگوريتم ژنتيک، يک جستجوي گسترده‌اي را در فضاي جستجوي مسئله انجام داد. با اين روش قادر خواهيم بود تا روش‌هاي مطرح شده در اين تحقيق را با استفاده از الگوريتم‌هاي تكاملي بهينه كنيم.

مراجع

مراجع

Aeberhard, S., Coomans D. and de Vel, O. (1992). "***Comparison of Classifiers in High Dimensional Settings***", *Tech. Rep. no. 92-02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland*.

Alizadeh H., Amirgholipour S.K., Seyedaghaee N.R. and Minaei-Bidgoli B. (2009a), ***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. (in press).*

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

Ayad H. and 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. LNCS 3541.

Ayad H.G. and 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.

Barthelemy J.P. and 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., and Jarmasz M. (2000), ***Resampling as a Cluster Validation Technique in fMRI,*** *JOURNAL OF MAGNETIC RESONANCE IMAGING* 11: pp. 228–231.

Banfield C.F. (1976), ***"Ultrametric Distances for a Single Linkage Dendrogram",*** *JSTOR: Applied Statistics, Statistical Algorithms,* Vol. 25, No. 3, pp. 313-315.

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.

Brandsma T. and 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.*

Brossier G. (1990). ***Piecewise hierarchical clustering,*** *Journal of Classification, Springer New York,* Vol. 7, No. 2, pp. 197-216.

Das A.K. and Sil J. (2007), ***Cluster Validation using Splitting and Merging Technique***, *in proc. of Int. Conf. on Computational Intelligence and Multimedia Applications, ICCIMA.*

Duda R.O., Hart P.E., and Stork D.G. (2001), ***Pattern Classification***, *second ed. Wiley*.

Dudoit S. and Fridlyand, J. (2003), *"Bagging to improve the accuracy of a clustering procedure", Bioinformatics*, 19 (9), pp. 1090-1099.

Estivill-Castro V. and Yang J. (2003), ***Cluster Validity Using Support Vector Machines,*** *DaWaK 2003,* LNCS 2737, pp. 244–256.

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. and 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. and Lin W. (2008), ***Cluster Ensemble Selection****, SIAM International Conference on Data Mining (SDM08).*

Fischer B. and Buhmann J.M., (2003), ***Bagging for path-based clustering****, IEEE Transactions on Pattern Analysis and Machine Intelligence,* pp.1411–1415.

Fred, A. and 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. and Jain A.K., (2003), ***Robust data clustering****, in: Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR ,USA,* vol. II, pp. 128–136.

Fred A.L. and Jain A.K. (2005). ***Combining Multiple Clusterings Using Evidence Accumulation****. IEEE Trans. on Pattern Analysis and Machine Intelligence,* 27(6):835–850.

Fred A. and Jain A.K. (2006), ***Learning Pairwise Similarity for Data Clustering****, In Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06).*

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

Fridlyand J. and 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.

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

Inokuchi R., Nakamura T. and 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.K. and Dubes R.C. (1988), ***Algorithms for Clustering Data***. *Prentice Hall.*

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

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

Kuncheva L.I. and Whitaker C. J., (2003), ***Measures of diversity in classifier ensembles****, Machine Learning.*

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

Lapointe F.J. and Legendre P. (1991). ***The generation of random ultrametric matrices representing dendrograms***. *Journal of Classification, Springer New York,* vol. 8, No. 2, pp 177-200.

Law M.H.C., Topchy A.P., and 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. and Ben-David S. (2005), ***Towards a statistical theory of clustering,*** *Technical report, PASCAL workshop on clustering, London.*

Man Y. and 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. and 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. and Minaei-Bidgoli B. (2008), "***Neural Network Ensembles using Clustering Ensemble and Genetic Algorithm**", Intl. Conf. on Convergence and hybrid Information Technology, ICCIT08, Nov. 11-13, IEEE CS*.

Moller U., Radke D. (2006), ***A Cluster Validity Approach based on Nearest-Neighbor Resampling**, In Proc. of the 18th Int. Conf. on Pattern Recognition (ICPR'06).*

Newman C.B.D.J., Hettich S. and Merz C. (1998), ***UCI repository of machine learning databases**, http://www.ics.uci.edu/~mlearn/MLSummary.html.*

Parvin H., Alizadeh H. and Minaei-Bidgoli B. (2009a), ***A New Method for Constructing Classifier Ensembles**, International Journal of Digital Content: Technology and its Application, JDCTA, ISSN: 1975-9339, (in press).*

Parvin H., Alizadeh H. and Minaei-Bidgoli B. (2009b), ***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.

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

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

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

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

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

Strehl A. and 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. and Punch, W.F. (2003), ***"Combining Multiple Weak Clusterings"**, Proc. 3d IEEE Intl. Conf. on Data Mining,* pp. 331-338.

Topchy A., Minaei-Bidgoli B., Jain A.K. and Punch W.F., (2004), ***Adaptive Clustering ensembles**, In Proc. Intl. Conf on Pattern Recognition,ICPR'04, Cambridge, UK,* pp.272-275.

Xie X.L., Beni G. (1991), ***A Validity measure for Fuzzy Clustering**, IEEE Trans. on Pattern Analysis and Machine Intelligence,* Vol.13, No.4, pp. 841–846.

داودي س. ر.، محمدي م. و رحماني ع.، (۱۳۸۵)، "*خوشه‌بندي تركيبي توزيع شده به روش* ***K-means*** "، دوازدهمين كنفرانس بين‌المللي انجمن كامپيوتر، تهران، ايران، ۱ تا ۳ اسفند.

عليزاده ح. ، مينايي بيدگلي ب.، (۱۳۸۷الف)، *ايجاد پراكندگي در خوشه‌بندي تركيبي*، كنفرانس ملي مهندسي نرم‌افزار و كاربردهاي آن، لاهيجان، ايران.

عليزاده ح.، مينايي بيدگلي ب.، (۱۳۸۷ب)، *بررسي روش‌هاي ارزيابي خوشه‌بندي*، كنفرانس ملي مهندسي نرم‌افزار و كاربردهاي آن، لاهيجان، ايران.

عليزاده ح.، مينايي بيدگلي ب.، (۱۳۸۷ج)، *يک روش جديد براي خوشه‌بندي تركيبي با استفاده از خوشه‌هاي پايدار*، پنجمين كنفرانس بين‌المللي مديريت فناوري اطلاعات و ارتباطات، تهران، ايران.

عظيمي ج. ،معاني ر. و مزيني ن. ، (۱۳۸۴)، *خوشه‌بندي تركيبي بهبود داده شده*، يازدهمين كنفرانس بين‌المللي كامپيوتر، انجمن كامپيوتر ايران، پژوهشگاه دانش‌هاي بنيادي، پژوهشكده علوم كامپيوتر، ۴ تا ۶ بهمن، تهران، ايران.

عظيمي ج.، (۱۳۸۶)، *بررسي پراكندگي در خوشه‌بندي تركيبي*، پايان نامه كارشناسي ارشد، دانشگاه علم و صنعت ايران، خرداد.

پیوست‌ها

Abstract:

Data clustering is an important step in data mining which excavates hidden patterns in unsupervised data. On account of complexity of the problem and weakness of the base clustering algorithms, most of researches are focused on ensemble methods. Diversity in primary results and also quality of the primary results are two vital factors which affect the final results. Although both of these two factors are considered and investigated in the late researches on clustering ensembles; there are some of questions which are still vague. Using a subset of primary results can be better that total primary results or not? Which subset of the primary results can cause to improve the performance of clustering ensemble? How to evaluate the primary results? This thesis tries to give a reasonable answer to these questions. Here, a framework for improving the efficiency of the clustering ensemble is proposed which is based on using a subset of primary clusters. Furthermore, some new methods for each step of this framework are suggested. For evaluating each individual cluster some new methods are proposed which are inspired from mutual information. Moreover, two new techniques are presented to construct the co-association matrix from only a subset of primary clusters. Experimental results over several standard data sets show that the presented schemes can significantly enhance the efficiency of the primary and even the full ensemble results. The average enhancement over 11 examined cases is 2.3% in comparison with the full ensemble. In addition, several artificial data sets are produced and examined during empirical studies.

Keywords:

Clustering Ensemble, Cluster Evaluation, Mutual Information, Subset of Primary Results, Evidence Accumulation Clustering, Co-association Matrix.

Iran University of Science and Technology
Computer Engineering Department

Clustering Ensemble Based on a Subset of Primary Results

A Thesis Submitted in Partial Fulfillment of the Requirement for the Degree of Master of Science in Computer Engineering

By:
Hosein Alizadeh

Supervisor:
Dr. Behrouz Minaei-Bidgoli

March 2009