ISSN: 2745-3673
http://www.Rcsj.ir

مجله علمی پژوهش در علوم رایانه (سال ششم)
شماره ۲۱ / بهار ۱۴۰۰ / ص ۱۳-۱

روشی جهت خوشه بندی اجماعی وزن دار شده توسط رابطه گری

رقیه مجد آبادی[۱]، علیرضا دهقانی[۲]

[۱] دانشجوی کارشناس ارشد مهندسی کامپیوتر گرایش نرم افزار.

[۲] گروه مهندسی کامپیوتر، واحد مرودشت،دانشگاه آزاد اسلامی، مرودشت، ایران.

نام نویسنده مسئول:
علیرضا دهقانی

تاریخ دریافت: ۱۳۹۹/۱۱/۴
تاریخ پذیرش: ۱۴۰۰/۱/۱۶

چکیده

خوشه بندی داده که جزء یادگیری بدون نظارت است، مسئله‌ای بسیار چالش برانگیز و مهم است. هدف خوشه‌بندی، پارتیشن بندی مجموعه‌ای از اشیای بدون برچسب به گروه‌ها یا خوشه‌های همگن می‌باشد. پارتیشن بندی اطلاعات به معنای طبقه بندی اطلاعات یا تقسیم بندی برخی نمونه ها در خوشه ها است به طوری که نمونه‌های داخل هر خوشه حداکثر شباهت به یکدیگر و حداکثر فاصله از سایر خوشه ها را دارند. الگوی پیشنهادی برای خوشه بندی اجماعی در این پایان نامه، حول وابستگی و اطمینان خوشه ها متمرکز می‌شود. ابتدا اندازه داده محاسبه می‌شود. سپس پارتیشن پایه باتوجه به یک خوشه بندی پایه ایجاد می شود که براساس این پارتیشن معیار وابستگی و میزان اطمینان برای هر خوشه محاسبه می شود. پس از آن ماتریس شباهت براساس رابطه وابستگی و اطمینان محاسبه می شود. آنگاه در مرحله بعدی پارتیشن‌های اجماع برای معیار وابستگی و اطمینان به ترتیب محاسبه می شوند که الگوریتم‌های خوشه بندی سلسله مراتبی تجمعی، به عنوان تابع توافق استفاده شده است. در مرحله آخر روش پیشنهادی جهت عملکرد باتوجه به معیارهای ارزیابی همچون دقت، صحت، شاخص رند و سایر معیارها ارزیابی شد. باتوجه به مجموعه داده ها تمامی معیارها عملکردی خوبی را نشان داده‌اند. در مقایسه با روش خوشه بندی **k-means** عملکرد بهتری داشته و می‌توان گفت این روش با روش های جدید مطابقت دارد و عملکردی مشابه‌ای داشته است.

واژگان کلیدی: داده کاوی، خوشه بندی اجماعی، خوشه بندی وزن دار.

ISSN: 2745-3673
http://www.Rcsj.ir

بیان مساله

فنون داده‌کاوی موجب انقلابی بزرگ در کسب و کارهای بزرگ شده‌اند. داده‌کاوی، قابلیت جستجوی پیچیده داده‌ای است که از الگوریتم‌های پیچیده‌ای استفاده می‌کند تا الگوها و همبستگی بین داده‌ها را کشف کند. داده‌کاوی، داده و دانشی را که در انبارهای داده مدفون است یافته و استخراج می‌کند. داده کاوی در واقع بخشی از فرآیند کشف دانش در پایگاه داده محسوب می‌شود و داده‌هایی را استخراج می‌کند که علم آمار ناتوان از تحلیل آنهاست. واژه کشف دانش در پایگاه داده‌ها در اوایل دهه ۸۰ در مراجعه به مفهوم کلی، گسترده، سطح بالا و به دنبال جستجوی دانش در اطلاعات شکل گرفته است. با شکل گرفتن این مفهوم، داده کاوی به عنوان یکی از مراحل فرآیند کشف دانش معرفی شد و به دلیل گستردگی، مباحث زیادی را به خود اختصاص فؤایند KDD عبارت است از: پاکسازی و یکپارچه سازی داده، ایجاد یک انبار داده مشترک برای تمام منابع، داده کاوی و بصری سازی نتایج تولید شده. در این فرآیند داده کاوی به عنوان مرحله سوم فرآیند کشف دانش معرفی می‌شود. اما این مفهوم بسیار گستره و پیچیده است. داده کاوی فرآیندی تحلیلی برای کاوش داده های طراحی شده است که در جستجوی الگوهای سازگار یا روابط سیستماتیک بین متغیرهاست و سپس به تایید این یافته ها با استفاده از الگوهای تشخیص داده شده می پردازد[۱]. داده کاوی الگوهای حاوی اطلاعات را در داده های موجود جستجو می کند. الگوریتم های داده کاوی می‌توانند جنبه پیش بینی داشته باشند و یا داده‌ها را توصیف کنند. داده کاوی توصیفی به دنبال یافتن اگرها در فعالیت های گذشته است و داده کاوی پیش بینانه با نگاه به سابقه، رفتار آینده را پیش بینی می کند.

یادگیری ماشینی[1] زیر مجموعه ای از داده کاوی است که تکنیک های خودکار برای یادگیری را برای پیش بینی های دقیق براساس مشاهدات گذشته مطالعه می کند. یادگیری ماشین از دو نوع تکنیک استفاده می کند: یادگیری نظارت شده (طبقه بندی و رگرسیون)، که یک مدل را بر روی داده های ورودی و خروجی شناخته شده آموزش می دهد تا بتواند خروجی های آینده را پیش بینی کند، و یادگیری بدون نظارت (خوشه بندی[2]) که الگوهای پنهان یا ساختارهای ذاتی را در ورودی می یابد.

خوشه بندی داده که جزء یادگیری بدون نظارت است، مسئله‌ای بسیار چالش برانگیز و مهم است. هدف خوشه‌بندی، پارتیشن بندی (بخش بندی) مجموعه‌ای از اشیای بدون برچسب به گروه‌ها یا خوشه‌های همگن می‌باشد[۲]. خوشه‌ها در مسائل دنیای واقعی می‌توانند در شکل، اندازه، درجه کمیت داده و درجه جداسازی متفاوت ظاهر گردند. تکنیک‌های خوشه بندی به تعریف پارامتر تشابه میان هر جفت الگو نیازمندند. اگر هیچ دانش قبلی درباره شکل و ساختار خوشه نداشته باشیم، انتخاب یک روش خوشه بندی مخصوص آسان نخواهد بود[۲].

تحقیقات درباره خوشه بندی‌ها در چند سال اخیر مثل همه زیرزمینه‌ها در تشخیص الگو و طبقه بندی به روش‌های ترکیبی تمایل نشان داده‌اند[۳]. روش‌های گروه بندی خوشه سعی دارند با ترکیب اطلاعات چندین پارتیشن بندی داده اولیه، راه‌حل خوشه بندی بهتر و قدرتمندتری بیابند. گروه‌های خوشه بندی مشکل، استخراج یک پارتیشن اجتماع از یک مجموعه خوشه بندی اولیه هستند. پارتیشن اجتماع باید نمایان‌ترین پارتیشن برای همه خوشه بندی‌های اولیه در گروه باشد. پارتیشن اجتماع وظیفه بهینه سازی یک تابع با هدف معین را برعهده دارد.

به عبارت دیگر گروه‌بندی خوشه‌بندی، که رویکردی در مساله خوشه‌بندی است، نتایج خوشه‌بندی چندگانه را برای دستیابی به خوشه‌های نهایی بدون دسترسی به ویژگی‌ها یا الگوریتم هایی که خوشه‌بندی را به دست می‌آورند، ترکیب می‌کند. ترکیب خوشه ها توسط یک الگوریتم اجماع انجام می شود. رویکرد گروه‌های خوشه‌ای سعی در بهبود کیفیت و استحکام نتایج خوشه‌بندی دارد[۴]. علاوه بر این، گروه خوشه بندی می تواند به برخی خصوصیات مانند تازگی، ثبات و مقیاس پذیری دست یابد. برخی از برنامه های گروه خوشه بندی در بیو انفورماتیک ، پردازش تصویر و بازاریابی وجود دارد. از آنجا که این مجموعه خوشه بندی فقط باید به جای داده های خود به دسته بندی های پایه دسترسی پیدا کند ، یک رویکرد مناسب برای حفظ حریم خصوصی و استفاده مجدد از دانش فراهم می کند.

[1] Machin learning
[2] Clustering

ISSN: 2745-3673
http://www.Rcsj.ir

عموماً دو گام برای گروه بندی خوشه وجود دارد: (الف) تولید چند پارتیشن بندی ضعیف، (ب)تجمیع پارتیشن بندی اولیه بدست آمده است. گام اول، تولید چند پارتیشن بندی ضعیف است. چون هر پارتیشن بندی اولیه یک جنبه پنهان از داده را آشکار می‌سازد، گروه آن‌ها می‌تواند ایرادات هر کدام از آن‌ها پوشش دهد. بنابراین نیاز است که نتایج اولیه دارای حداکثر تنوع ممکن باشند تا اطلاعات بیشتری درباره الگوهای اصلی داده ارائه دهند. روش‌های بسیاری جهت ایجاد تنوع لازم برای نتایج اولیه پیشنهاد شده است استفاده از الگوریتم‌های خوشه بندی مختلف ساده‌ترین راهکار می‌باشد.

هدف از این پایان نامه ارائه روشی جهت مجموعه گروه‌های خوشه ای، که چندین پارتیشن داده اصلی یا خوشه ها را به یک راه حل خوشه‌ای بهتر که معمولاً به عنوان پارتیشن اجماع نامیده می شود، ترکیب می کند. وابستگی خوشه های مختلف در طول تولید پارتیشن اجماع استفاده می شود. برای پیشنهاد ماتریس انجمنی وزنی[3]، باید مکانیسم وزنی معنی داری تعریف شود. در مکانیزم وزن، یک وزن به هر خوشه با توجه به ارزش خود اختصاص داده می‌شود. تعیین ارزش یک خوشه از تعیین مقدار یک کلاس چالش برانگیز است، زیرا نمی توان یک خوشه را مطابق برچسب ها ارزیابی کرد. بنابراین، ما نیاز داریم تا مقدار خوشه را با توجه به قابلیت اطمینان آن تعریف کنیم.

یکی از مهمترین اشکال در رویکردهای مبتنی بر ماتریس همبستگی، عدم توانایی آنها برای در نظر گرفتن وزن خوشه ها است. بنابراین، ما در این پایان نامه ماتریس همبستگی مشترک وزنی را براساس قابلیت اطمینان اول و دوم پیشنهاد خواهیم کرد. قابلیت اطمینان، از مهمترین ویژگی‌های یک سیستم است. برای تخصیص قابلیت اطمینان به اجزای آیتم‌ها ابتدا باید وابستگی اجزا همچنین قدرت وابستگی بین اجزا جهت تخصیص قابلیت اطمینان دقیق به اجزا مشخص گردد که یکی از راه های شناسایی اجزای مستقل و وابسته استفاده از الگوریتم های خوشه بندی است. هدف معمول خوشه بندی با استفاده از قابلیت اطمینان کاهش وابستگی های خارجی، افزایش وابستگی های داخلی با تغییر محل اجزا در ماتریس است.

اهمیت و ضرورت تحقیق

مشکل کشف پارتیشن اجماع در گروه خوشه ای به جای مشکل کشف طبقه بندی اجماع در گروه طبقه بندی، یک کار چالش برانگیزتر است. یک کار چالش برانگیز دیگر در مجموعه خوشه‌ها در مقایسه با طبقه بندی ترکیبی، مطابقت خوشه ها در پارتیشن های مختلف گروه است[۲و۳]. دو عامل مهم در تولید نسل گروه‌ها عبارتند ازعامل اول: کیفیت خوشه و عامل دوم: تنوع بین گروه. عامل اول بدان معنی است که کیفیت خوشه ها در این گروه از اهمیت بالایی برخوردار است. در نتیجه، باید آن را در طول مشکل تولید گروه بررسی کرد. مکانیسم‌های مکانیکی برای تضمین آن وجود دارد: ۱) برای تولید مجموعه‌ای از خوشه های با کیفیت بالا، ۲) برای تولید تعدادی از خوشه ها صرف نظر از ویژگی های آنها، و سپس فیلتر کردن آنها با توجه به ویژگی های خود، و ۳) برای تولید یک عملکرد اجماع که از پارتیشن اجماع با توجه به وزن خوشه هایی که با توجه به خصوصیات آنها به خوشه ها اختصاص می یابد، استخراج می شود. عامل دوم به این معنی است که خوشه های گروه باید متنوع باشند. مجدداً دو رویکرد برای تضمین آن وجود دارد: ۱)برای تولید یک گروه با تنوع سطح بالا بین گروه و ۲) برای تولید یک گروه از خوشه ها صرف نظر از تنوع آنها، و سپس انتخاب زیر مجموعه ای از آنها تا آنجا که ممکن است متنوع باشد[۴].

هر روش گروه خوشه ای سعی می کند هنگام استخراج پارتیشن اجماع از مجموعه، معیار خاصی را بهینه کند. اما گروه-های خوشه سنتی، تمام اعضای مجموعه را با اهمیت برابر در ساخت پارتیشن اجماع در نظر می گیرند. این بدان معناست که هر پارتیشن یا خوشه اساسی به طور معادل در الگوریتم گروه خوشه شرکت می کند. در واقع، آنها توجه به اهمیت اعضای آن را در نظر نمی گیرند. اما بدیهی است که برخی از خوشه‌ها با کیفیت بیشتری سزاوار تأکید بیشتر هستند و برخی از خوشه‌ها نیازمند کیفیت کمتری در زمان ایجاد پارتیشن اجماع هستند.

باتوجه به مباحث فوق نوآوری این پایان نامه (الف) یک معیار برای ارزیابی کیفیت هر خوشه دلخواه پیشنهاد می‌شود، ب) مکانیسمی برای ارزیابی کیفیت محاسبه شده یک خوشه به یک مقدار وزن معنی دار، و (ج) رویکردی برای استفاده از مقادیر

[3] weighing co-association matrix

ISSN: 2745-3673
http://www.Rcsj.ir

وزن خوشه های اساسی است که با استفاده از ماتریس همبستگی وزنی محاسبه می‌شود. این ماتریس فاصله گره ها را محاسبه و دخیره می‌کند.

فرضیات تحقیق

فرضیات:

۱.استفاده از مقادیر وزن برای محاسبه فاصله گره ها برای ایجاد ماتریس همبستگی و خوشه بندی کارا است.

۲.قابلیت اطمینان جهت رابطه خوشه‌های مختلف در طول تولید پارتیشن اجماع استفاده می‌شود.

اهداف

اهداف تحقیق به صورت زیر بیان می شود:

۱- محاسبه قابلیت اطمینان هر خوشه براساس فاصله گره ها

۲- محاسبه کیفیت هر خوشه باتوجه به فاصله گره ها

از آنجایی که خوشه بندی یکی از الگوریتم های پرکاربرد داده کاوی است می‌توان از این پژوهش در تمامی صنایع که به دنبال خوشه بندی داده های خود هستند، استفاده نمود.

مجموعه داده

مجموعه داده مورد استفاده در این پژوهش از ۳ پایگاه داده واقعی از پایگاه اطلاعاتی مخزن یادگیری ماشین UCI [۶] است.

جدول(۴-۱) اطلاعات مجموعه داده

مجموعه داده	تعدادنمونه	تعداد ویژگی:تعداد خوشه های هدف
سرطان سینه (BC)	۶۸۳	۹:۲
عنبیه(I)	۱۵۳	۴:۳
یونسفر(IoS)	۳۵۱	۳۴:۲

در ابتدا، برای تشکیل هر آزمون آزمایشی، کل پایگاه داده های معیار در ابتدا با این هدف استاندارد می‌شوند که هر مشخصه در هر مجموعه داده مورد استفاده در فاصله [۰ ، ۱] ترسیم شود. این بدان معنی است که قبل از انجام هر پیشرفت، مرحله پیش پردازش باید مانند معادله (۴-۱) انجام شود.

$$\ddot{D}_{jk} = \frac{(D_{jk} - \min_{k\in\{1,\dots,D_{1:}\}} D_{jk})}{\max_{k\in\{1,\dots,D_{1:}\}} D_{jk} - \min_{k\in\{1,\dots,D_{1:}\}} D_{jk}}$$

رابطه(۴-۱)

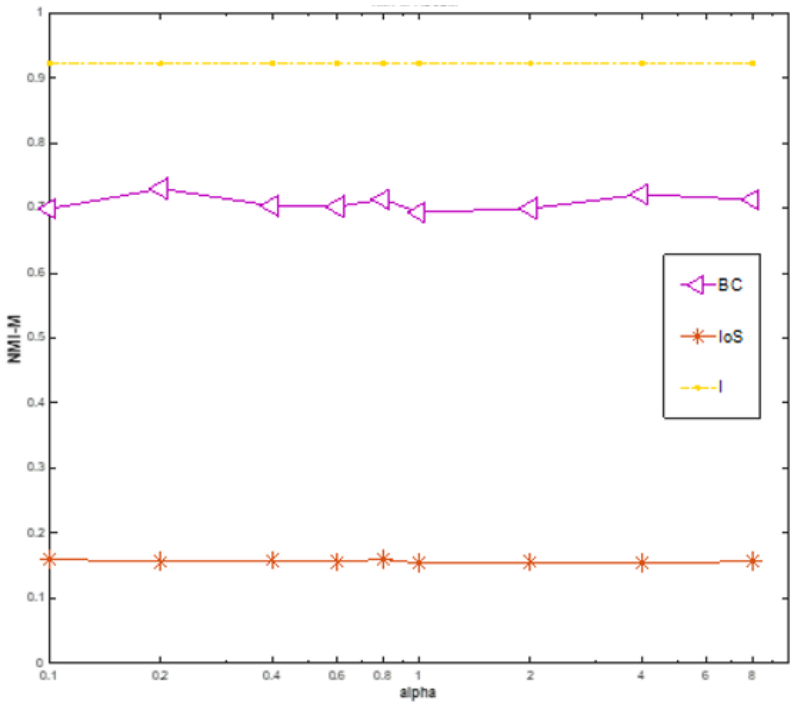
در این رابطه D_{jk} ویژگی (j,k) از دیتاست ورودی است. j در این جا شمارنده نمونه ها و k شمارنده ویژگی ها است. در این رابطه مقدار ماکزیمم و مینیمم ها بر روی سطر ماتریس دیتا و بر روی یک نمونه انجام می‌شود.

نتایج شبیه سازی

در شبیه‌سازی اول و به منظور نشان دادن عملکرد الگوریتم با تغییر پارامترهای α و B و با در نظر گرفتن **NMI** به عنوان معیار عملکرد، الگوریتم را برای مقادیر مختلف این پارامتر ها اجرا کرده و **NMI** حاصل را در شکل زیر نشان داده ایم.

ISSN: 2745-3673
http://www.Rcsj.ir

شبیه‌سازی ها با میانگین گرفتن ۳۰ شبیه‌سازی نشان داده شده است در شکل اول با تغییر α عملکرد الگوریتم نشان داده شده است

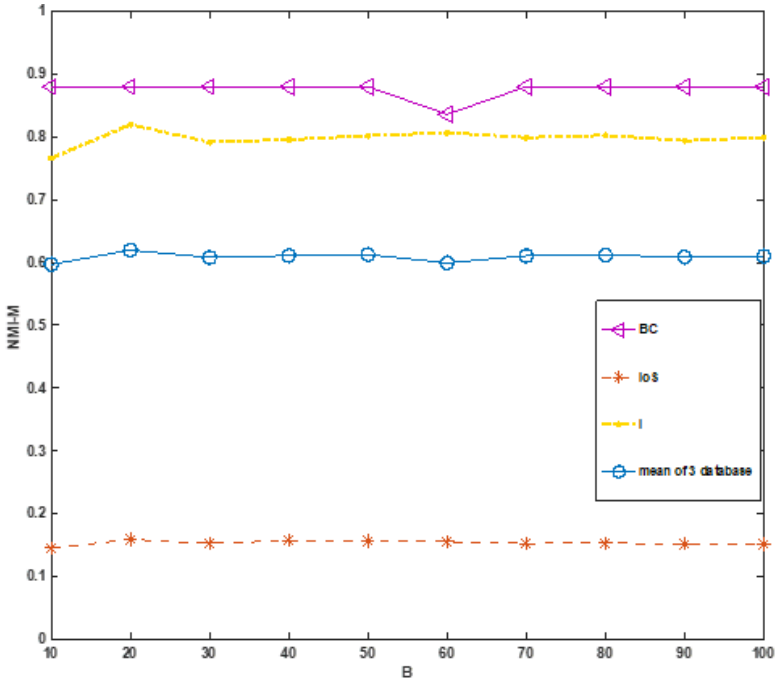


شکل(۴-۱) عملکرد الگوریتم با تغییر پارامترهای α

متغیر a تأثیر مهمی در رابطه بین قابلیت اطمینان و غیرقابل اعتماد بودن دارد. حتی یک افزایش اندک در درجه غیرقابل اعتماد بودن منجر به کاهش قابل توجهی در سطح قابلیت اطمینان، برای مقادیر کم متغیر α می شود. شکل (۴-۱) تأثیر متغیر α را نشان می دهد. همانطور که در شکل نشان داده شده است، پیشنهاد می شود که مقداری باید به متغیر α در محدوده ۰.۳ تا ۰.۷ داده شود. سپس از مقدار ۰.۴ برای متغیر α برای ادامه استفاده می کنیم.

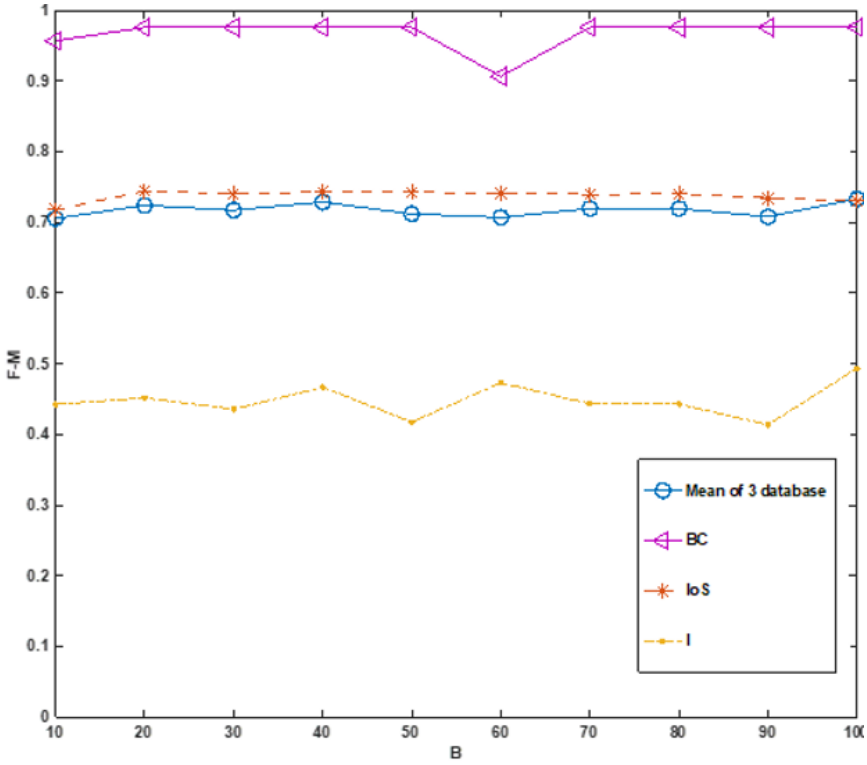
در شبیه‌سازی بعدی پارامتر **B** یا همان تعداد پارتیشن‌ها تغییر می‌کند که به صورت شکل زیر حاصل می‌شود:

ISSN: 2745-3673
http://www.Rcsj.ir



شکل(۴-۲) عملکرد الگوریتم با تغییر پارامترهای B

تعدادی آزمایش برای تعیین بهترین اندازه اجماع برای متغیر B انجام شده است و عملکرد گروه خوشه بندی ما از نظر مقادیر مختلف برای B در شکل (۴-۲)نشان داده شده است. تأثیر متغیر اندازه اجماع در شکل فوق تجزیه و تحلیل شده است. اختصاص مقداری در حدود ۴۰ به متغیر سایز اجماع با توجه به نتایج ارائه شده در شکل پیشنهاد می شود.

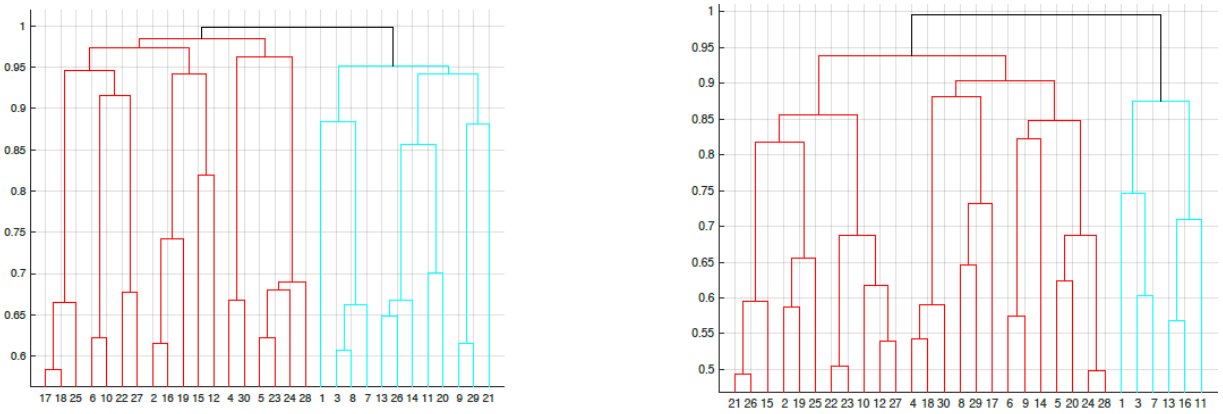


شکل(۴-۳) میانگین شبیه‌سازی ۳۰ بار اجرای الگوریتم

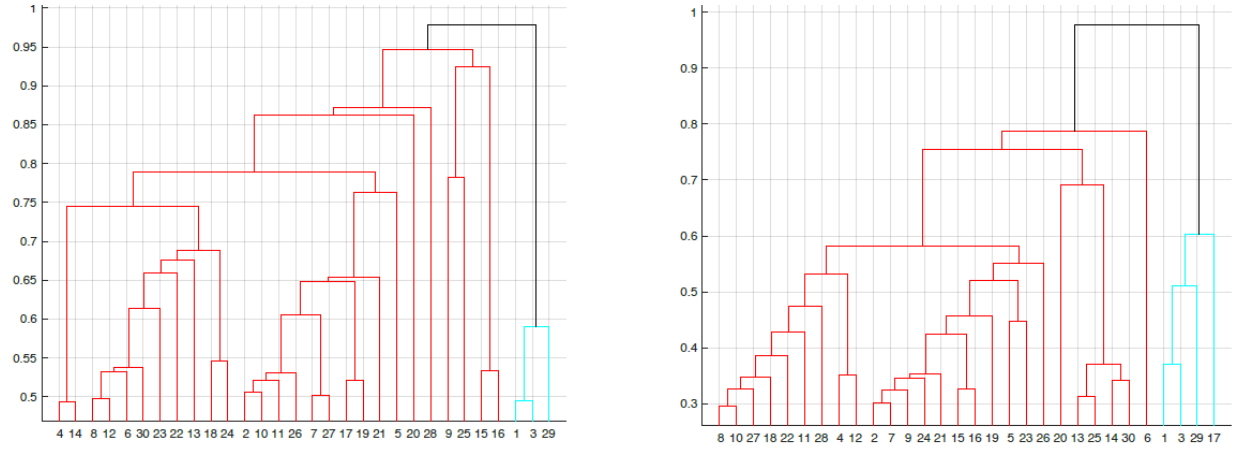
ISSN: 2745-3673
http://www.Rcsj.ir

تمام شبیه‌سازی ها از میانگین شبیه‌سازی ۳۰ بار اجرای الگوریتم بدست آمده و تقریبا در تمام شبیه‌سازی ها نتایج خوبی بدست آمده است.

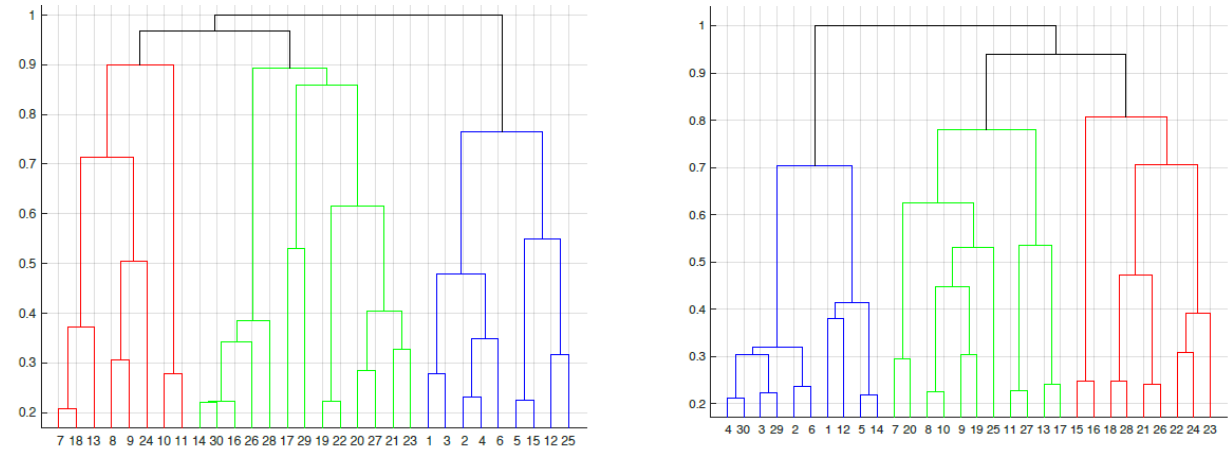
دندوگرام خوشه‌بندی اجماعی برای هر یک از دیتاست ها در شکل های زیر آورده شده است. برای بهتر نشان دادن شکل‌ها تنها ۳۰ گره خروجی برای هر کدام نشان داده شده است.



شکل (۴-۴)الف) دندوگرام با ۲خوشه ALSMدر مجموعه داده BC، ب) دندوگرام با ۲خوشه ALFM



شکل (۴-۵)الف) دندوگرام با ۲خوشه ALSMدر مجموعه داده IOS، ب) دندوگرام با ۲خوشه ALFM



شکل (۴-۶)الف) دندوگرام با ۲خوشه در مجموعه داده IRIS، ب) دندوگرام با ۲خوشه ALFM

ISSN: 2745-3673
http://www.Rcsj.ir

ارزیابی

با توجه به معیارهای ارزیابی تعریف شده در فصل پیشین، روش پیشنهادی ارزیابی و نتایج در جداول زیر ارائه شده است:

جدول(۴-۲) نتایج خوشه بندی اجماعی براساس معیار ارزیابی NMI-M

Name	ALFM	ALSM
BC	0.7915	0.7980
IOS	0.1138	0.1299
I	0.7697	0.7697

جدول(۴-۳) نتایج خوشه بندی اجماعی براساس معیار ارزیابیF-M

Name	ALFM	ALSM
BC	0.9749	0.9761
IOS	0.7506	0.7548
I	0.9293	0.9293

جدول(۴-۴) نتایج خوشه بندی اجماعی براساس معیار ارزیابیARI-M

Name	ALFM	ALSM
BC	0.8744	0.8799
IOS	0.1583	0.1727
I	0.8680	0.8680

جدول(۴-۵) نتایج خوشه بندی اجماعی براساس معیار ارزیابیA-M

Name	ALFM	ALSM
BC	0.9678	0.9693
IOS	0.7009	0.7094
I	0.9533	0.9533

جدول(۴-۶) نتایج خوشه بندی اجماعی براساس معیار ارزیابیS-M

Name	ALFM	ALSM
BC	0.9617	0.9640
IOS	0.7022	0.6978
I	0.9200	0.9200

جدول(۴-۷) نتایج خوشه بندی اجماعی براساس معیار ارزیابیSp-M

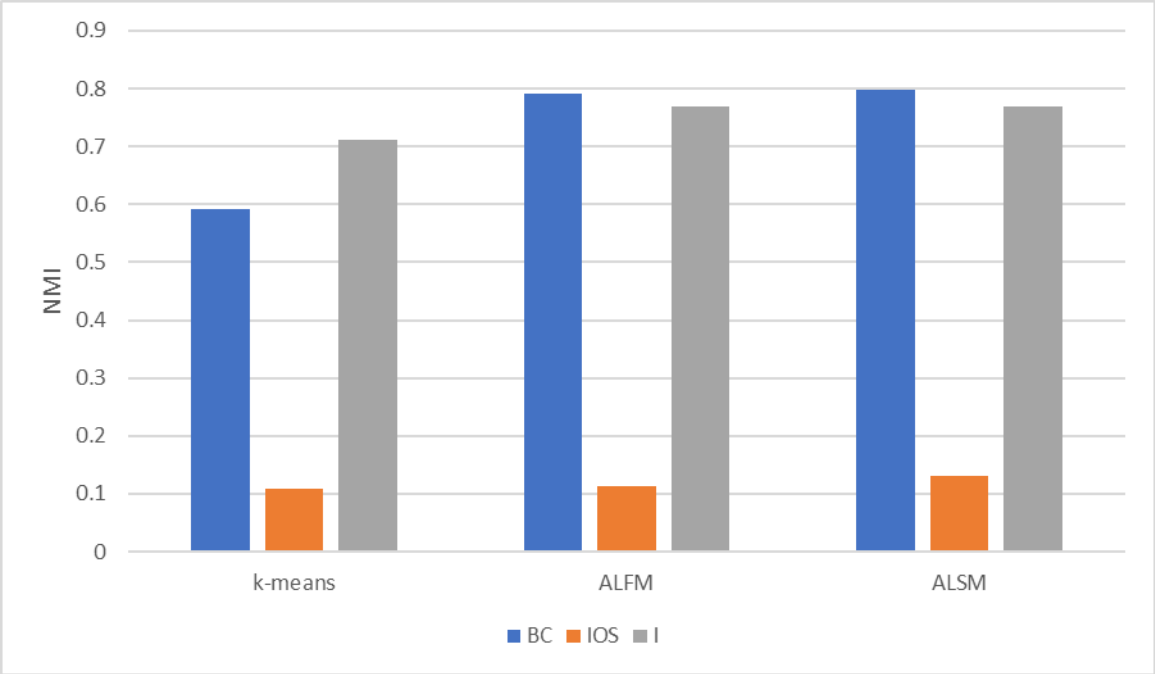
Name	ALFM	ALSM
BC	0.9791	0.9791
IOS	0.6984	0.6984
I	0.9700	0.9700

جدول(۴-۸) نتایج خوشه بندی اجماعی براساس معیار ارزیابیP-M

Name	ALFM	ALSM
BC	0.9884	0.9885
IOS	0.8061	0.8220
I	0.9388	0.9388

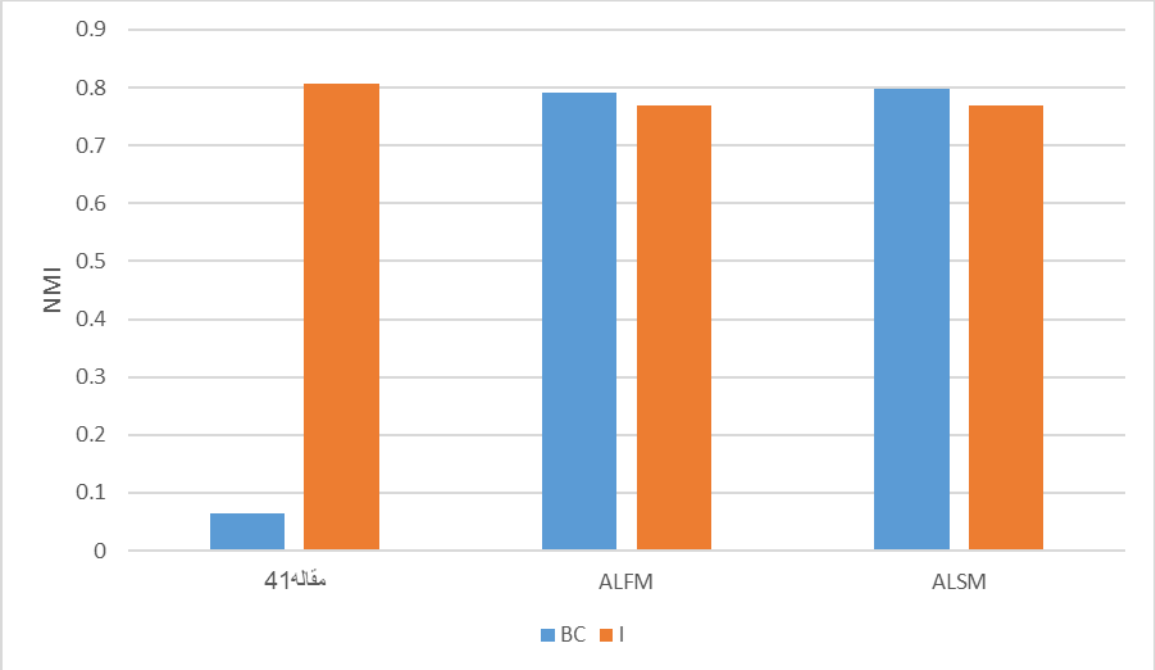
ISSN: 2745-3673
http://www.Rcsj.ir

همچنین در شکل(۴-۷) مقایسه شاخص NMI برای روش پیشنهادی و الگوریتم خوشه بندی k-means نشان داده است که عملکرد روش پیشنهادی این پایان نامه بهتر از خوشه بندی k-means است.



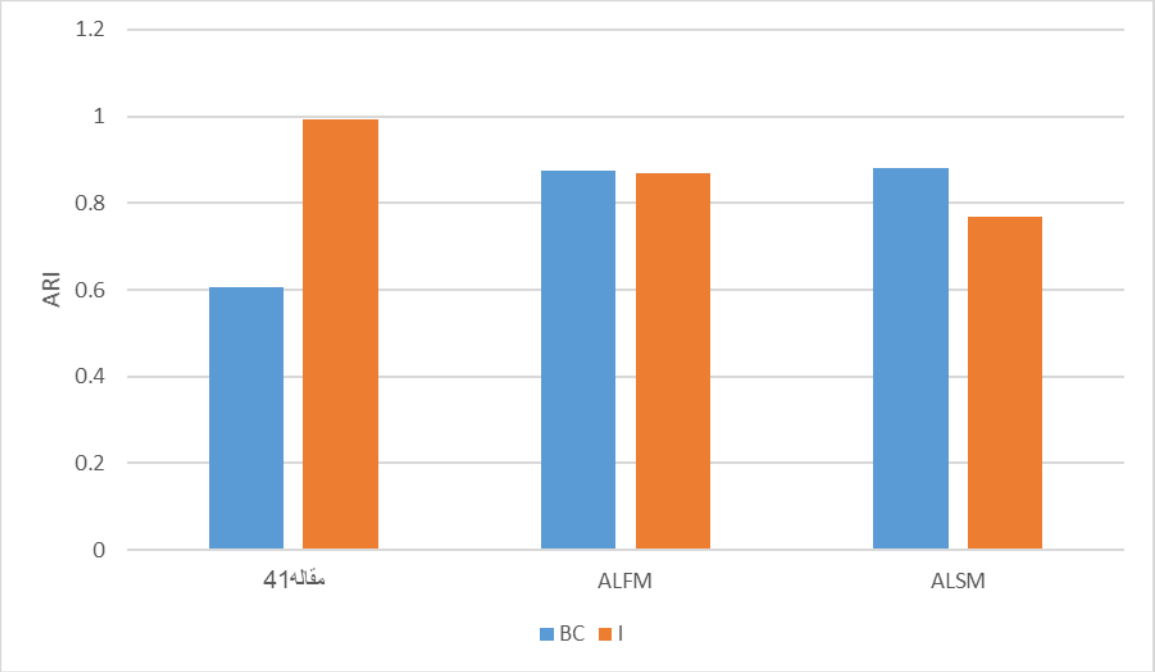
شکل(۴-۷) مقایسه شاخص NMI برای روش پیشنهادی و k-means

در شکل(۴-۸) روش پیشنهادی با مقاله[۴۱](بای و همکاران۲۰۲۰) برای دو شاخص NMI و ARI مقایسه شده که عملکرد روش پیشنهادی برای مجموعه I نسبت به مقاله[۴۱] با اینکه کمتر از آن است ولی نتیجه قابل قبولی را ارائه کرده و برای مجموعه داده BC عملکرد روش پیشنهادی بهتر از مقاله عنوان شده است.



شکل(۴-۸)الف) مقایسه شاخص NMI برای روش پیشنهادی و مقاله[۴۱]

ISSN: 2745-3673
http://www.Rcsj.ir



شکل(۴-۸)ب) مقایسه شاخص ARI برای روش پیشنهادی و مقاله[۴۱]

نتیجه گیری

این پایان نامه از طریق استفاده از قابلیت اطمینان تقریبی خوشه ها برای دستیابی به تقسیم اجماع، رویکرد خوشه بندی اجماعی وزن دار را معرفی می کند. یک رویکرد خوشه بندی اجماعی را پیشنهاد می کند که شامل چند مرحله است. قابلیت اطمینان هر خوشه در ابتدا از طریق یک محاسبه آنتروپی و یک تغییر نمایی محاسبه می شود، که نشان دهنده مقدار توزیع خوشه در خوشه های مختلف یک پارتیشن در یک مجموعه مرجع است. قابلیت اطمینان خوشه های مختلفی که در تولید پارتیشن اجماعی استفاده می شود. روشی در این پایان نامه پیشنهاد شده که با توجه به قابلیت اطمینان آن بتواند در هر خوشه ای استفاده شود که این بر اساس تجمع وزنی خوشه است. پس از آن، راه حل پیشنهادی برای گروه خوشه بندی در ۳ پایگاه داده در دنیای واقعی ارزیابی می شود. ارزیابی تجربی نشان می دهد که راه حل پیشنهادی عملکرد بهتری نسبت به رویکردهای خوشه بندی ساده دارد. بنابراین، مطالعات گسترده در مورد پایگاه داده های دنیای واقعی نشان می دهد که روش پیشنهادی می تواند با رویکردهای جدید مطابقت داشته باشد یا عملکرد بهتری داشته باشد.

اجرای ایده قابلیت اطمینان برای خوشه اجماعی وزن دار است. شایان ذکر است که قابلیت اطمینان در واقع یک نوع عدم قطعیت است. از نظر تجربی نشان داده شده است که رویکرد گروه خوشه بندی پیشنهادی از نظر اندازه گیری عملکرد، پیچیدگی زمان و استحکام بهتر از رویکردهای گروه خوشه بندی ساده دارد.

ISSN: 2745-3673
http://www.Rcsj.ir

منابع و مراجع

[۱] قره نژاد سحر.لزوم حفظ مشتریان بیمه با استفاده از ابزارهای داده کاوی،دانشجوی کارشناسی ارشد مدیریت فناوری اطلاعات، تازه های جهان بیمه. شماره ۱۵۰ و۱۵۱.

[2] Abbasi, S. O., Nejatian, S., Parvin, H., Rezaie, V., & Bagherifard, K. (2019). Clustering ensemble selection considering quality and diversity. *Artificial Intelligence Review*, *52*(2), 1311-1340.

[3] Yu, Z., Chen, H., You, J., Han, G., & Li, L. (2013). Hybrid fuzzy cluster ensemble framework for tumor clustering from biomolecular data. *IEEE/ACM transactions on computational biology and bioinformatics*, *10*(3), 657-670.

[4] Akbari, E., Dahlan, H. M., Ibrahim, R., & Alizadeh, H. (2015). Hierarchical cluster ensemble selection. *Engineering Applications of Artificial Intelligence*, *39*, 146-156.

[5] Sivakumar, A., & Gunasundari, R. (2017). A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining. *International Journal of Pure and Applied Mathematics*, *117*(20), 785-794.

[6] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

[7] Mirza, S., Mittal, S., & Zaman, M. (2016). A Review of Data Mining Literature. International Journal of Computer Science and Information Security (IJCSIS), 14(11).

[8] Sivakumar, A., & Gunasundari, R. (2017). A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining. International Journal of Pure and Applied Mathematics, 117(20), 785-794.

[9] Kuwil, F. H., Shaar, F., Topcu, A. E., & Murtagh, F. (2019). A new data clustering algorithm based on critical distance methodology. Expert Systems with Applications, 129, 296-310.

[10] Kleinberg, J. M. (2003). An impossibility theorem for clustering. In Advances in neural information processing systems (pp. 463-470).

[11] García-Escudero, L. A., Gordaliza, A., Matrán, C., & Mayo-Iscar, A. (2010). A review of robust clustering methods. Advances in Data Analysis and Classification, 4(2-3), 89-109.

[12] Coretto, P., & Hennig, C. (2010). A simulation study to compare robust clustering methods based on mixtures. Advances in Data Analysis and Classification, 4(2-3), 111-135.

[13] Fred, A. L., & Jain, A. K. (2002, August). Data clustering using evidence accumulation. In Object recognition supported by user interaction for service robots (Vol. 4, pp. 276-280). IEEE.

[14] Strehl, A., & Ghosh, J. (2002). Cluster ensembles---a knowledge reuse framework for combining multiple partitions. Journal of machine learning research, 3(Dec), 583-617.

[15] Topchy, A., Jain, A. K., & Punch, W. (2003, November). Combining multiple weak clusterings. In Third IEEE international conference on data mining (pp. 331-338). IEEE.

[16] Gullo, F., Tagarelli, A., & Greco, S. (2009, April). Diversity-based weighting schemes for clustering ensembles. In Proceedings of the 2009 SIAM International Conference on Data Mining (pp. 437-448). Society for Industrial and Applied Mathematics.

[17] Fred, A., & Lourenço, A. (2008). Cluster ensemble methods: from single clusterings to combined solutions. In Supervised and unsupervised ensemble methods and their applications (pp. 3-30). Springer, Berlin, Heidelberg.

[18] Mirzaei, A. (2009). Combining hierarchical clusterings with emphasis on retaining the structural contents of the base clusterings. Computer Engineering & IT Department, Amir-kabir University of Technology, Tehran.

ISSN: 2745-3673
http://www.Rcsj.ir

[19] Friedman, J. H., & Meulman, J. J. (2004). Clustering objects on subsets of attributes (with discussion). Journal of the Royal Statistical Society: Series B (Statistical Methodology), 66(4), 815-849.
[20] Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M., & Papadopoulos, D. (2007). Locally adaptive metrics for clustering high dimensional data. Data Mining and Knowledge Discovery, 14(1), 63-97.
[21] Al-Razgan, M., & Domeniconi, C. (2006, April). Weighted clustering ensembles. In Proceedings of the 2006 SIAM International Conference on Data Mining (pp. 258-269). Society for Industrial and Applied Mathematics.
[22] Domeniconi, C., & Al-Razgan, M. (2009). Weighted cluster ensembles: Methods and analysis. ACM Transactions on Knowledge Discovery from Data (TKDD), 2(4), 1-40.
[23] Li, T., & Ding, C. (2008, April). Weighted consensus clustering. In Proceedings of the 2008 SIAM International Conference on Data Mining (pp. 798-809). Society for Industrial and Applied Mathematics.
[24] Gullo, F., Tagarelli, A., & Greco, S. (2009, April). Diversity-based weighting schemes for clustering ensembles. In Proceedings of the 2009 SIAM International Conference on Data Mining (pp. 437-448). Society for Industrial and Applied Mathematics.
[25] Huang, D., Lai, J., & Wang, C. D. (2016). Ensemble clustering using factor graph. Pattern Recognition, 50, 131-142.
[26] Liu, H., Liu, T., Wu, J., Tao, D., & Fu, Y. (2015, August). Spectral ensemble clustering. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 715-724).
[27] Ren, Y., Domeniconi, C., Zhang, G., & Yu, G. (2017). Weighted-object ensemble clustering: methods and analysis. Knowledge and Information Systems, 51(2), 661-689.
[28] Mojarad, M., Nejatian, S., Parvin, H., & Mohammadpoor, M. (2019). A fuzzy clustering ensemble based on cluster clustering and iterative fusion of base clusters. *Applied Intelligence*, *49*(7), 2567-2581.
[29] Hailin, L., & Miao, W. (2020). Fuzzy clustering based on feature weights for multivariate time series. *Knowledge-Based Systems*, 105907.
[30] Soufiane, K., Imene, H., Manel, A., & Tarek, K. M. (2019, March). Clustering Ensemble Approach Based on Incremental Learning. In *Proceedings of the 9th International Conference on Information Systems and Technologies* (pp. 1-7).
[31] Liang, W., Zhang, Y., Xu, J., & Lin, D. (2019). Optimization of Basic Clustering for Ensemble Clustering: An Information-Theoretic Perspective. *IEEE Access*, *7*, 179048-179062.
[32] Yang, H., Peng, H., Zhu, J., & Nie, F. (2020). Co-Clustering Ensemble Based on Bilateral K-Means Algorithm. *IEEE Access*, *8*, 51285-51294.
[33] Latifi Pakdehi, A., & Daneshpour, N. (2019). Cluster ensemble selection using voting. *Signal and Data Processing*, *15*(4), 17-30.
[34] Zhang, M. (2019). Weighted Clustering Ensemble: A Review. *arXiv preprint arXiv:1910.02433*.
[35] Harakawa, R., Takimura, S., Ogawa, T., Haseyama, M., & Iwahashi, M. (2019). Consensus Clustering of Tweet Networks via Semantic and Sentiment Similarity Estimation. *IEEE Access*, *7*, 116207-116217.
[36] Vahidi Ferdosi, S., & Amirkhani, H. (2020). Weighted Ensemble Clustering for Increasing the Accuracy of the Final Clustering. Signal and Data Processing, 17(2), 100-85.
[37] Vega-Pons, S., & Ruiz-Shulcloper, J. (2011). A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence, 25(03), 337-372.

ISSN: 2745-3673
http://www.Rcsj.ir

[38] Topchy, A., Jain, A. K., & Punch, W. (2003, November). Combining multiple weak clusterings. In Third IEEE international conference on data mining (pp. 331-338). IEEE.
[39] Fred, A. L., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. IEEE transactions on pattern analysis and machine intelligence, 27(6), 835-850.
[40] Azimi, J., & Fern, X. Z. (2009, July). Adaptive cluster ensemble selection. In Ijcai (Vol. 9, pp. 992-997).
[41] Bai, L., Liang, J., & Cao, F. (2020). A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters. *Information Fusion*, *61*, 36-47.