IDMC۲۰۰۹ - Paper ۱٤۲۹

ارائه یک مدل خوشه‌بندی جمعی وزن‌دار[i] با حفظ اختصاصی بودن[ii] داده-ها و قابلیت برخورد با نویز

نیلوفر مظفری[۱]؛ مینا علی بیگی[۲]؛ ستار هاشمی[۳]؛ اشکان سامی[۴]؛ علی حمزه[۵]

چکیده

خوشه‌بندی جمعی[iii] برای رویارویی با داده‌های با حجم بالا و تعداد ابعاد زیاد پیشنهاد شده‌است. این مدل با استفاده از خوشه-بندی‌های مختلف و ترکیب نتایج حاصل سعی در بهبود خوشه‌بندی سنتی دارد. امروزه، حفظ امنیت و اختصاصی بودن داده‌ها از مهمترین چالش‌های موجود در پردازش داده ها می‌باشد که بدین منظور می توان از خوشه‌بندی جمعی استفاده کرد.

این مقاله روشی وزن‌دار برای ترکیب چندین خوشه‌بندی مختلف از داده‌ها ارائه می‌کند که علاوه بر حفظ اختصاصی بودن داده-ها، نسبت به خرابکاری (نویز) در محیط‌های ناامن رفتار مناسبی نشان می‌دهد. مدل ارائه شده در این مقاله، می‌تواند اختصاصی بودن داده‌ها را حفظ نماید، بدین مفهوم که با کمترین اطلاعات ممکن از داده‌ها و در نتیجه حفظ اختصاصی بودن آنها عمل ترکیب را انجام دهد. علاوه بر آن نتایج به دست آمده نشان می‌دهد که در صورت دستکاری داده‌ها در محیط ناامن، نتیجه مدل چندان تحت تاثیر قرار نخواهد گرفت.

کلمات کلیدی

خوشه‌بندی[iv]، خوشه‌بندی کننده[v]، خوشه[vi]، ترکیب‌کننده[vii]، ویژگی[viii]، برچسب[ix]، خوشه‌بندی جمعی وزن‌دار

Toward Robust Privacy Preserving Clustering: A Weighted Ensemble Approach

Niloofar Mozafari; Mina Alibeigi; Sattar Hashemi; Ashkan Sami; Ali Hamzeh

ABSTRACT

Ensemble clustering has been shown to perform well on high dimensional data with huge volume. It decomposes data into several partitions, run a clustering algorithm on every partition and then combines the results together, what in turn can improve performance of the overall model. Nowadays, security and the privacy of data become an important challenge in data processing; where ensemble approach can be considered as a possible solution.

This paper presents a weighted ensemble approach to clustering that preserves data privacy and is robust to noise. The proposed approach can preserve data privacy in the sense that the combiner just needs to have only partial information of data available, rather than, all information as a whole. From robustness point of view, our results are also promising, say, when the results of some clustering are manipulated either intentionally due to attackers' breach or unintentionally due to any problem in communication or storage media, the performance of model keeps intact.

KEYWORDS Clustering, Combiner, Cluster, Feature, Labels, Weighted ensemble clustering

---

۱. کارشناسی ارشد دانشگاه شیراز؛ mozafari@cse.shirazu.ac.ir
۲. کارشناسی ارشد دانشگاه شیراز؛ alibeigi@cse.shirazu.ac.ir
۳. استادیار دانشگاه شیراز؛ s_hashemi@cse.shirazu.ac.ir
٤. استادیار دانشگاه شیراز؛ asami@ieee.org
٥. استادیار دانشگاه شیراز؛ ali@cse.shirazu.ac.ir

۱. مقدمه

خوشه‌بندی مجموعه‌ای از داده‌ها به طوری که در هر خوشه عناصر معناداری قرار گیرد، امری مهم و پرکاربرد است [۱]. هدف اصلی در خوشه‌بندی این است که داده‌های درون یک خوشه بیشترین شباهت را به یکدیگر داشته باشند و یا به عبارتی می‌توان گفت که داده‌های درون یک خوشه به یکدیگر "نزدیکتر" و داده‌های درون خوشه‌های مختلف، از یکدیگر "دورتر" باشند. دوری و یا نزدیکی داده‌ها نسبت به یکدیگر بر اساس معیار فاصله-ای که در مساله تعیین می‌شود، مشخص می‌گردد. مثلا در خوشه‌بندی اسناد دوری و یا نزدیکی داده‌ها متناسب است با تعداد کلمه‌های مشترکی که در دو سند وجود دارد و یا در خوشه‌بندی سبد خرید مشتریان، دوری و یا نزدیکی داده‌ها بر اساس شباهت خریدها مشخص می‌شود.

در یک مساله خوشه‌بندی هدف پیدا کردن بهترین خوشه‌بندی است. در حالت کلی پیدا کردن بهترین خوشه بندی از داده‌ها یک مساله NP complete است [۲]. ولی چندین الگوریتم heuristic وجود دارند که می‌توانند جواب‌های قابل قبولی را در بسیاری از کاربردها تولید کنند. از جمله این الگوریتم‌ها می‌توان به الگوریتم‌های K-means ،METIS و DBSCAN اشاره کرد [۳] ,[٤], [٥].

مشکل اساسی در تمامی این الگوریتم‌ها این است که در این الگوریتم‌ها فرض می‌شود، همه داده‌ها با تمامی ابعادشان در یک محل قرار دارند و با این فرض عمل خوشه‌بندی انجام می‌شود. علاوه بر این هنگامی که ابعاد داده‌ها زیاد شود، کارایی این الگوریتم‌ها کاهش می‌یابد. عملا در بسیاری از کاربردها به دلایلی از جمله نبودن فضای کافی برای ذخیره کردن تمام داده‌ها با کلیه‌ی ابعادشان و یا به دلایل امنیتی، امکان اینکه تمام داده‌ها با همه ابعادشان در یک محل قرار گیرند و عمل خوشه‌بندی انجام شود، وجود ندارد.

به منظور حل مشکلات فوق در این مقاله به جای ذخیره‌سازی تمام داده‌ها با تمامی ابعادشان در یک محل و انجام عمل خوش‌بندی، از خوشه‌بندی ترکیبی استفاده می‌شود. دو دیدگاه متداول برای حل مسائل خوشه‌بندی ترکیبی عبارتند از: روش‌های مبتنی بر گراف[x] و روش بهینه‌سازی حریصانه[xi] [٦] .

در روش بهینه‌سازی حریصانه روش کار بدین صورت است که از بین خوشه‌بندی‌های مختلف، خوشه‌بندی که بیشترین میزان شباهت را با دیگر خوشه‌بندی‌های موجود دارد انتخاب می‌شود. سپس داده‌ها به صورت تصادفی بین خوشه‌های مختلف جابجا می‌شوند و در صورتی یک داده در خوشه جدیدش می‌ماند که شباهت خوشه‌بندی کنونی را به خوشه‌بندی‌های دیگر بیشتر شود. این کار برای تمام داده‌ها تکرار می‌شود [٦] ,[٧].

روش‌های مبتنی بر گراف اینگونه عمل می‌کنند که هر خوشه را به عنوان یک ابریال[xii] در ابرگراف[xiii] نمایش می‌دهند. سه الگوریتم متداول مبتنی بر گراف عبارتند از:

- الگوریتم خوشه‌بندی مبتنی بر شباهت[xiv] : این الگوریتم بر اساس شباهت بین داده‌ها در خوشه‌بندی‌های مختلف، نتیجه خوشه‌بندی را تولید می‌کند [٦] ,[٧].
- الگوریتم خوشه‌بندی مبتنی بر ابرگراف[xv]: این الگوریتم با استفاده از ابرگراف در خوشه‌بندی‌های مختلف، نتیجه خوشه‌بندی را تولید می‌کند [٦] ,[٧] ,[٨].
- الگوریتم‌های ماوراء خوشه‌بندی[xvi]: این الگوریتم خوشه‌بندی نهایی را با قرار دادن چندین ابریال مرتبط در یک ماوراء خوشه[xvii] تولید می‌کند [٦] ,[٧].

مشکل روش بهینه‌سازی حریصانه این است که در حالت کلی جواب بهینه اصلی[xviii] را نمی‌تواند بیابد و علاوه بر این، مشکل دیگر این روش حجم بسیار زیاد محاسبات آن است. در سال ۲۰۰۸، K. Tumer و دیگران روشی برای حل مشکلات این روش مبتنی بر روش‌های پاداش و جزا[xix]، ارائه نمودند [۲]. یکی از مشکلات روش‌های مبتنی بر گراف نیز این است که خراب شدن هر کدام از خوشه‌بندی‌ها (سهوا و یا عمدا)، در نتیجه نهایی آن تاثیر بسزایی دارد. در این مقاله برای رفع این مشکل از یک روش وزن‌دهی استفاده شده است که در مقایسه با روش‌های بهینه‌سازی حریصانه نیز حجم محاسباتی کمتری دارد. نتایج به دست آمده نشان می‌دهد که روش ارائه شده در مقایسه با روش‌های پیشین، نسبت به نویز یا خرابی هر خوشه‌بندی مقاوم‌تر است.

در این مقاله روشی برای ترکیب چندین خوشه‌بندی مختلف از داده‌ها، که هرکدام می‌تواند نتیجه اجرای یک الگوریتم خوشه‌بندی متفاوت باشد، ارائه می‌شود. ترکیب‌کننده ارائه شده در این مقاله، این ویژگی را دارد که علاوه بر حفظ اختصاصی بودن داده‌ها نسبت به خراب‌کاری یا نویز در محیط‌های ناامن رفتار مناسبی را انجام می‌دهد و با کمترین اطلاعات ممکن از داده‌ها، عمل ترکیب را انجام می‌دهد. تنها اطلاعات مورد نیاز این نوع ترکیب‌کننده، این است که بداند چه داده‌هایی به یکدیگر شباهت دارند؛ که این شباهت با دادن برچسب یکسان به داده‌های مشابه در نظر گرفته می‌شود. بنابراین هر الگوریتم خوشه‌بندی می‌تواند تنها به زیر مجموعه‌ای از ویژگی داده‌ها دسترسی داشته باشد و بر اساس همان ویژگی‌ها و معیار خوشه‌بندی مورد نظر خود، داده‌ها را دسته‌بندی می‌کند و به داده‌های مشابه‌ای که در یک خوشه قرار می‌گیرند، برچسب یکسانی می‌دهد. هر الگوریتم خوشه‌بندی هیچ اطلاعی در مورد نحوه انجام عمل خوشه‌بندی و یا ویژگی‌های در دسترس دیگر الگوریتم‌های خوشه‌بندی، ندارد و بدین صورت اختصاصی بودن داده‌ها رعایت می‌شود. پس از تولید نتایج هر یک از الگوریتم‌های خوشه‌بندی، نتایج حاصل از تمامی الگوریتم‌های خوشه-بندی به ترکیب‌کننده فرستاده می‌شود و نهایتا ترکیب‌کننده این برچسب‌ها را می‌گیرد و یک نتیجه را به عنوان خوشه‌بندی نهایی از داده‌ها برمی-گرداند. حال اگر در این بین یک یا چند الگوریتم خوشه‌بندی به هر دلیلی (سهوا یا عمدا)، نتایج غیر معتبری در مورد شباهت بین داده‌ها به ترکیب‌کننده ارائه دهند، برخلاف روشهای پیشین، در روش ارائه شده، این عمل تاثیر چندانی در خوشه‌بندی نهایی نخواهد داشت.

در ادامه در بخش ۲ روش پیشنهادی مورد بررسی قرار می‌گیرد. نتایج به‌دست‌آمده در بخش ۳ نشان داده شده‌اند و در بخش چهارم نیز نتیجه-گیری و کارهای آینده ارائه می‌شود.

۲. روش پیشنهادی

قبل از توضیح روش پیشنهادی، لازم است چندین مفهوم پایه که در ادامه مورد استفاده قرار می‌گیرند، تعریف شوند. این مفاهیم عبارتند از:

- خوشه‌بندی کننده: الگوریتم خوشه‌بندی که روی داده‌ها اعمال می‌شود.
- خوشه‌بندی : نتایج حاصل از خوشه‌بندی کننده.
- خوشه : هر یک از دسته‌های موجود در خوشه‌بندی.

روی مجموعه داده‌های یکسان، با توجه به اینکه چه خوشه‌بندی کننده‌ای و در آن خوشه‌بندی کننده چه پارامترهایی استفاده می‌شود و یا اینکه چه ویژگی‌هایی از داده برای خوشه‌بندی به کار می‌روند، خوشه‌بندی‌های مختلفی از داده‌ها خواهیم داشت. از آنجا که هر داده فقط می‌تواند در یک خوشه قرار بگیرد، هر خوشه‌بندی معادل یک برچسب‌دهی به داده‌ها است [۳] ,[۹]؛ به طوری که برچسب هر داده متناسب با خوشه‌ای است که داده در آن قرار گرفته است و داده‌هایی که در یک خوشه قرار دارند، برچسب یکسانی می‌گیرند. پس هر داده می‌تواند در خوشه‌بندی‌های مختلف، برچسب‌های مختلفی داشته باشد. خوشه‌بندی ترکیبی، نتایج خوشه‌بندی‌های مختلف را ترکیب می‌کند و بدون نیاز به دسترسی به کل ویژگی داده‌ها، خوشه‌بندی تولید می‌کند که بتواند به خوبی، خوشه‌بندی‌های مختلف را توصیف کند. از آنجا که هر خوشه‌بندی کننده تنها با فرستادن برچسب‌دهی داده‌ها، نتایج خوشه‌بندی خود را برای ترکیب‌کننده می‌فرستد، بنابراین نه ترکیب‌کننده و نه دیگر خوشه‌بندی‌ها اطلاعی از ویژگی‌ها و یا پارامترهای خوشه‌بندی‌های دیگر ندارند و بدین ترتیب بحث اختصاصی بودن داده‌ها در نظر گرفته می‌شود.

خوشه‌بندی ترکیبی ارائه شده در این مقاله بدین صورت عمل می‌نماید که ابتدا هر خوشه‌بندی را به یک ابرگراف تبدیل می‌کند. ابرگراف، گرافی است که شامل چندین ابریال است. ابریال برخلاف یال که تنها دو گره را به هم متصل می‌نماید، می‌تواند چندین گره را به هم متصل کند [۶]. برای تبدیل هر خوشه‌بندی به یک ابرگراف، داده‌های هر خوشه با یک ابریال به هم متصل می‌شوند و برای هر خوشه‌بندی یک ماتریس مجاورت ابرگراف تشکیل می‌شود.

	$\lambda^{(1)}$	$\lambda^{(2)}$	$\lambda^{(3)}$
X_1	۱	۲	۱
X_2	۳	۲	۲
X_3	۱	۲	۱
X_4	۲	۳	۱
X_5	۲	۳	۲
X_6	۲	۳	۱
X_7	۳	۱	۲

↔

	$H^{(1)}$			$H^{(2)}$			$H^{(3)}$	
	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8
V_1	۱	۰	۰	۰	۱	۰	۱	۰
V_2	۰	۰	۱	۰	۱	۰	۰	۱
V_3	۱	۰	۰	۰	۱	۰	۱	۰
V_4	۰	۱	۰	۰	۰	۱	۱	۰
V_5	۰	۱	۰	۰	۰	۱	۰	۱
V_6	۰	۱	۰	۰	۰	۱	۱	۰
V_7	۰	۰	۱	۱	۰	۰	۰	۱

شکل ۱ : مساله خوشه‌بندی ترکیبی با ۳ خوشه‌بندی، خوشه‌بندی اول و دوم سه کلاسه و خوشه‌بندی سوم دو کلاسه است. جدول سمت چپ برچسب دهی‌های هر خوشه‌بندی و جدول سمت راست ماتریس مجاورت ابرگراف معادل با ۸ ابریال. هر خوشه به یک ابریال تبدیل شده است.

شکل ۱ نمونه‌ای از مساله خوشه‌بندی جمعی را نشان می‌دهد. جدول سمت چپ در شکل ۱ نتایج حاصل از ۳ خوشه‌بندی با در نظر گرفتن زیرمجموعه‌های مختلفی از ویژگی داده‌ها بر روی ۷ داده را نشان می‌دهد. تعداد ردیف‌های این ماتریس برابر با تعداد کل داده‌ها و تعداد ستون‌های آن به تعداد خوشه‌بندی‌هاست. به عنوان مثال اولین ستون (λ^1) نتایج یک خوشه‌بندی با ۳ خوشه است. در این خوشه‌بندی داده‌های X_1 و X_3 در یک خوشه قرار دارند و یا به عبارتی به هم شبیه هستند و به همین دلیل هر دو برچسب یکسان گرفته‌اند. جدول سمت راست نشان‌دهنده ماتریس مجاورت ابرگراف مربوط به هر خوشه‌بندی است. در این جدول $H^{(1)}$ ماتریس مجاورت مربوط به خوشه‌بندی اول می‌باشد. همان‌طور که گفته شد، هر خوشه به یک ابریال نگاشت می‌شود و تعداد ابریال‌های ابرگراف متناظر با هر خوشه‌بندی برابر است با تعداد خوشه‌های آن خوشه‌بندی. به عنوان مثال ۳ خوشه در خوشه‌بندی λ^1 با h_1، h_2 و h_3 نشان داده شده‌اند. اولین خوشه در خوشه‌بندی λ^1شامل داده‌های X_1 و X_3 است به همین علت ابریال متناظر با این خوشه به ازای این داده‌ها مقدار ۱ و به ازای دیگر داده‌ها مقدار صفر دارد.

ماتریس شباهت برای هر خوشه‌بندی با ضرب ماتریس مجاورت ابرگراف آن خوشه‌بندی در ترانهاده‌اش به دست می‌آید. حال ماتریس شباهت نهایی متناسب است با میانگین ماتریس‌های شباهتی که در مرحله قبل به دست آمده است. با توجه به ماتریس شباهت به دست آمده در این مرحله، خوشه‌بندی نهایی به دست می‌آید. برای این کار از هر الگوریتم خوشه‌بندی که بر اساس ماتریس شباهت کار می کند، می توان استفاده کرد. در این روش تمام خوشه‌بندی‌ها وزن مشابهی دارند و این امر در محیط‌های ناامن که امکان دستکاری برچسب داده‌ها وجود دارد، باعث خراب شدن نتیجه نهایی می‌گردد. به عنوان مثال فرض کنید شرکتی می‌خواهد بر اساس اطلاعاتی که از انبارهای مختلفش دریافت می‌کند، عمل خوشه‌بندی را انجام دهد. هر انبار به چندین ویژگی از داده‌ها دسترسی دارد که به دلایلی نمی‌خواهد این اطلاعات را با بقیه به اشتراک بگذارد. در این صورت هر انبار بر اساس ویژگی‌های در دسترس خود عمل خوشه‌بندی را انجام می‌دهد و نتیجه خوشه‌بندی را به شرکت می‌فرستد. تنها اطلاعاتی که هر یک از انبارها با بقیه به اشتراک می‌گذارند، این است که چه داده‌هایی به هم شباهت دارند که این شباهت، با دادن برچسب یکسان به داده‌هایی که در یک خوشه قرار دارند و یا به عبارتی به هم شبیه هستند، در نظر گرفته می‌شود. در نتیجه هر انبار یک برچسب‌دهی از داده‌ها را به شرکت می-فرستد. حال اگر به دلیل فرستادن بر چسب‌دهی اشتباه از انبار به شرکت و یا خراب شدن برچسب‌ها در طول انتقال آنها از طریق شبکه به شرکت، نتیجه نهایی هر یک از انبارها خراب شود، نتیجه نهایی شرکت نیز می‌تواند خراب شود.

بنابراین برای کاهش تاثیر خوشه‌بندی‌های خراب یا نویزی در محیط‌های ناامن، ماتریس شباهتی که از هر خوشه‌بندی به دست می‌آید در عددی که میزان شباهت آن خوشه‌بندی با بقیه خوشه‌بندی‌ها را نشان می‌دهد، ضرب می‌شود تا بدین وسیله اثر خوشه‌بندی‌های خراب در محیط‌های ناامن کم شود. برای این منظور چند خوشه‌بندی باید با یکدیگر مقایسه شوند. در این مقاله برای مقایسه چند خوشه‌بندی از بین معیارهای مختلف ارائه شده [۱۰] ,[۱۱]، از معیار شباهت NMI^{xx} [۶] ,[۱۲] استفاده می‌شود. فرمول (۱) نحوه محاسبه Mutual Information بین دو خوشه‌بندی Y_i و Y_k را نشان می‌دهد.

$$I(Y_i, Y_k) = \sum_{X_c^i \in Y_i} \sum_{X_{c'}^k \in Y_k} \frac{\left|X_c^i \cap X_{c'}^k\right|}{N} \log\left(\frac{N\left|X_c^i \cap X_{c'}^k\right|}{\left|X_c^i\right|\left|X_{c'}^k\right|}\right) \qquad (۱)$$

در این فرمول X^i_c نشان‌دهنده مجموعه داده‌هایی است که در مجموعه X و در خوشه بندی Y_iام در خوشه c قرار گرفته‌اند و N تعداد کل داده‌ها است. برای محاسبه NMI بین دو خوشه‌بندی آنتروپی هر یک از خوشه‌بندی‌ها مورد نیاز است. آنتروپی هر خوشه‌بندی با توجه به فرمول (۲) محاسبه می‌شود.

$$H(Y_i) = -\sum_{X^i_c \in Y_i} \frac{|X^i_c|}{N} \log(\frac{|X^i_c|}{N}) \qquad (۲)$$

با توجه به فرمول‌های (۱) و (۲)، NMI بین دو خوشه‌بندی Y_i و Y_j، طبق فرمول (۳) به دست می‌آید. از این معیار می توان برای مقایسه دو خوشه‌بندی استفاده کرد.

$$NMI(Y_i, Y_k) = \frac{I(Y_i, Y_k)}{\sqrt{H(Y_i)H(Y_k)}} \qquad (۳)$$

فرمول (۳) میزان شباهت دو خوشه‌بندی را نشان می‌دهد و فرمول (۴) میزان شباهت یک خوشه‌بندی با بقیه خوشه‌بندی‌ها را نشان می‌دهد. برای مقایسه یک خوشه‌بندی با دیگر خوشه‌بندی‌ها فرمول (۴) استفاده می‌شود. این فرمول میانگین NMI بین یک خوشه‌بندی با تمام خوشه‌بندی‌های دیگر است. همانگونه که قبلا نیز بیان شد برای مقاوم سازی خوشه‌بندی نهایی در مقابل نویز، برای ماتریس شباهت هر خوشه‌بندی یک وزن در نظر گرفته می‌شود؛ این وزن همان $ANMI^{xxi}$ مربوط به آن خوشه‌بندی است.

$$ANMI(\bar{Y}, Y_i) = \frac{1}{|\bar{Y}|} \sum_{Y_k \in \bar{Y}} NMI(Y_i, Y_k) \qquad (۴)$$

ماتریس شباهت هر خوشه‌بندی در ANMI یا به عبارتی در میزان شباهتش با بقیه خوشه‌بندی‌ها ضرب می‌شود. این عمل باعث می‌شود که اگر یک خوشه‌بندی به هر دلیلی نتایج غلطی را به ترکیب‌کننده بفرستد، این عمل در نتیجه نهایی تاثیرچندانی نداشته باشد. ماتریس شباهت نهایی برابر است با میانگین وزنی ماتریس‌های شباهت خوشه‌بندی های مختلف.

حال با داشتن ماتریس شباهت می‌توان خوشه‌بندی نهایی را تعیین کرد. برای این کار می‌توان از الگوریتم‌های graph partitioning استفاده کرد. در حالت کلی روش‌های مختلفی برای حل مسئله graph partitioning به وجود آمده است که نتایج قابل قبولی نیز می‌دهند [۱۳] ,[۱۴], [۱۵]. در این مقاله برای graph partitioning یا تولید خوشه‌بندی نهایی از روش METIS که یک روشmulti-level graph partitioning می‌باشد، استفاده می‌شود [۳].

۳. نتایج

آزمایش‌ها بر روی دو مجموعه داده با نام‌های ۸d۵k و Iris انجام شده است. Iris مجموعه داده‌ای سه کلاسه با ۱۵۰ نمونه و ۴ ویژگی می‌باشد. ۸d۵k از آدرس http://strehl.com قابل دانلود است که مجموعه داده‌ای شامل ۱۰۰۰ نمونه می‌باشد که از ۵ تابع توزیع گوسی در فضای ۸ بعدی به دست آمده‌اند؛ همه خوشه‌ها واریانس مشابهی دارند (۰.۱) اما با میانگین‌های متفاوت. میانگین خوشه‌ها با توزیع یکنواخت از یک ابرمکعب واحد انتخاب می‌شوند.

به منظور ارزیابی روش ارائه شده، لازم است که اثر سطح نویزهای مختلف روی مجموعه داده‌ها بررسی شود. در آزمایش‌ها نویز در دو سطح اعمال می‌شود. در سطح اول، نویز روی خوشه‌بندی‌ها اعمال می‌شود. در این سطح نویزهای۲۰٪ و ۴۰٪ اعمال شده است. اگر در خوشه‌بندی جمعی ۵ خوشه‌بندی داشته باشیم، نویز ۲۰٪ به این معناست که تنها برچسب داده‌های یک خوشه‌بندی که از بین ۵ خوشه‌بندی به صورت تصادفی انتخاب می‌شود، با درصدهای مختلفی خراب شده است. به عنوان مثال با ۲۰٪ نویز در خوشه‌بندی‌های نویزی و یا به عبارتی خراب شدن تنها یکی از ۵ خوشه‌بندی‌های موجود، اگر درصد نویز ۴۰٪ را اعمال کنیم یعنی از بین ۱۰۰۰ داده برچسب ۴۰۰ داده آن خوشه‌بندی با احتمال یکنواخت به صورت تصادفی خراب می‌شوند.

آزمایش‌ها بر روی مجموعه داده‌های Iris و ۸d۵k با ۵ خوشه‌بندی و سطح نویزهای مختلف انجام شده است. جدول۱ و جدول ۲ به ترتیب نشان-دهنده نتایج به دست آمده بر روی ۸d۵k و Iris می‌باشد. هر ردیف نشان‌دهنده درصد خوشه‌بندی‌های خرابی است که نویز با درصدهای ۲۰٪، ۴۰٪، ۶۰٪، ۸۰٪ و ۱۰۰٪ روی آن‌ها اعمال شده است. ستون اول نشان‌دهنده درصد خوشه‌بندی‌های خراب و ستون دوم درصد نویز در هر یک از خوشه‌بندی‌های خراب را نشان می‌دهد. بر روی خوشه‌بندی‌ها با سطوح نویز مختلف الگوریتم CSPA [۶] و الگوریتم ارائه شده در این مقاله، اعمال شده اند. ستون سوم دقت خوشه‌بندی نهایی حاصل از اعمال الگوریتم CSPA است و ستون چهارم نشان‌دهنده‌ی دقت خوشه‌بندی نهایی پس از اعمال روش ارائه شده می‌باشد. دقت هر روش با استفاده از معیار NMI بین خوشه‌بندی حاصله و برچسب واقعی داده‌ها به دست آمده است؛ که در واقع نشان‌دهنده میزان شباهت خوشه‌بندی با برچسب واقعی داده‌ها می‌باشد که هرچه شباهت خوشه‌بندی به دست‌آمده با برچسب واقعی داده‌ها بیشتر باشد، معیار NMI بیشتر خواهد بود و یا به عبارتی دقت افزایش می‌یابد. بدین ترتیب در روش ارائه شده با وجود نداشتن برچسب واقعی داده‌ها، خوشه‌بندی نهایی به برچسب واقعی داده‌ها نزدیک‌تر می‌گردد و یا به عبارتی در روش ارائه شده داده‌های درون یک خوشه با احتمال بیشتری در یک کلاس قرار می‌گیرند.

از نتایج به دست آمده مشاهده می‌شود که روش ارائه شده در مقایسه با روش CSPA در مقابل نویز رفتار مناسب‌تری دارد و دقت خوشه‌بندی که با این روش به دست می‌آید، نسبت به دقت خوشه‌بندی به دست آمده از روش CSPA در تمام سطوح مختلف نویز بالاتر است علاوه بر این به این دلیل که ترکیب‌کننده تنها با برچسب داده‌ها عمل ترکیب را انجام می‌دهد، اختصاصی بودن داده‌ها نیز حفظ می‌شود و اینگونه ترکیب‌کننده ارائه شده نیازی به داشتن ویژگی‌های در دسترس هر یک از خوشه‌بندی‌ها ندارد.

جدول ۱ : میانگین دقت خوشه بندي بر روي ۲۰ بار اجراي الگوريتم به درصد بر روي مجموعه داده ۸d۵k

درصد خوشه‌بندی‌های نویزی به کل خوشه‌بندی‌ها	درصد نویز اعمال شده در هر خوشه‌بندی	دقت CSPA	دقت Proposed
۲۰٪	۲۰٪	۹۲.۱۲	۹۳.۷۱
	۴۰٪	۹۲.۳۱	۹۴.۷۰
	۶۰٪	۹۱.۰۷	۹۲.۶۲
	۸۰٪	۹۰.۵۴	۹۱.۲۶
	۱۰۰٪	۹۲.۴۲	۹۲.۸۵
۴۰٪	۲۰٪	۹۰.۴۳	۹۱.۹۴
	۴۰٪	۹۱.۳۶	۹۳.۴۱
	۶۰٪	۸۶.۰۵	۸۸.۳۸
	۸۰٪	۹۰.۰۹	۹۱.۲۷
	۱۰۰٪	۹۲.۲۱	۹۲.۲۸

جدول ۲ : میانگین دقت خوشه‌بندي بر روي ۲۰ بار اجراي الگوريتم به درصد بر روي مجموعه داده Iris

درصد خوشه‌بندی‌های نویزی به کل خوشه‌بندی‌ها	درصد نویز اعمال شده در هر خوشه‌بندی	دقت CSPA	دقت Proposed
۲۰٪	۲۰٪	۸۱.۴۵	۸۳.۳۲
	۴۰٪	۸۱.۰۸	۸۳.۱۷
	۶۰٪	۸۰.۰۵	۸۱.۹۱
	۸۰٪	۸۰.۵۸	۸۱.۶۰
	۱۰۰٪	۸۱.۸۱	۸۲.۹۶
۴۰٪	۲۰٪	۸۰.۳۸	۸۱.۲۵
	۴۰٪	۷۷.۵۲	۷۸.۹۳
	۶۰٪	۷۷.۸۲	۷۸.۵۳
	۸۰٪	۷۷.۲۶	۷۷.۶۷
	۱۰۰٪	۷۶.۰۱	۷۶.۵۵

۴. نتیجه گیری و کارهای آینده

این مقاله به بررسی خوشه‌بندی جمعی می‌پردازد. در مواجهه با داده‌های با حجم زیاد و ابعاد بالا، خوشه‌بندی جمعی روشی مناسب می‌باشد. یکی از مسائلی که در دنیای امروز وجود دارد، حفظ امنیت و اختصاصی بودن داده‌هاست. برای این منظور خوشه‌بندی جمعی گزینه مناسبی برای برخورد با این مساله است. مدل ارائه شده در این مقاله علاوه بر حفظ اختصاصی بودن داده‌ها، در محیط‌های ناامن نسبت به خرابکاری داده‌ها (نویز) نیز رفتار مناسبی نشان می‌دهد.

مدل خوشه‌بندی وزنی ارائه شده در این مقاله با کمترین اطلاعات ممکن از داده‌ها عمل ترکیب را انجام می‌دهد و یا به عبارتی برای خوشه‌بندی نیازی به داشتن ویژگی‌های در دسترس هر خوشه‌بندی کننده ندارد؛ بنابراین می‌تواند اختصاصی بودن داده‌ها را حفظ کند. علاوه بر این با در نظر گرفتن یک وزن در ترکیب‌کننده که میزان شباهت خوشه‌بندی با دیگر خوشه‌بندی‌ها را نشان می‌دهد، می‌تواند در محیط‌های ناامن نسبت به خرابکاری (نویز) در مقایسه با روش‌های قبل رفتار مناسب‌تری را نشان دهد.

در روش ارائه شده برای وزن‌دهی تنها شباهت هر خوشه‌بندی با دیگر خوشه‌بندی‌ها در نظر گرفته شده است؛ در صورتی که بتوان روشی ارائه داد که اهمیت ویژگی‌های در دسترس هر خوشه‌بندی کننده را نیز در نظر بگیرد و به هر خوشه‌بندی کننده با توجه به ویژگی‌هایی که برای خوشه-بندی داده‌ها استفاده می‌کند، وزن دهد؛ می‌توان خوشه‌بندی جمعی با دقت بالاتری را ارائه داد. به این منظور از ایده ماوراء ویژگی[xxii] [۱۶] و ایده ارائه شده در [۱۷] می‌توان استفاده کرد.

۵. مراجع

[۱] A. K. Jain, M. N. Murty, and P. J. Flynn; *"Data clustering: a review"*, ACM Computing Surveys, ۳۱ (۳):۲۶۴–۳۲۳, September ۱۹۹۹.

[۲] K. Tumer, A. K. Agogino; *"Ensemble Clustering with voting active Clusters"*, Pattern Recognition Letters ۲۹, ۱۹۴۷-۱۹۵۳, ۲۰۰۸.

[۳] G. Karypis, V. Kumar; *"A fast and high quality multi-level scheme for partitioning irregular graphs"*, SIAM ۲۰ (۱), ۳۵۹-۳۹۲, ۱۹۹۸.

[۴] M. Ester, H. Kriegel, J. Sander, X. Xu; *"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise"*, International Conference on Knowledge Discovery and Data Mining (KDD-۹۶).

[۵] J. MacQueen; *"Some methods for classification and analysis of multivariate observations"*, Proc. ۵th Berkeley Symposium on Mathematical Statistics and Probability, ۱۹۶۷.

[۶] A. Strehl, J. Ghosh; *"Cluster ensembles-a knowledge reuse framework for combining multiple partitions"*, Journal of Machine Learning Research ۳, ۵۸۳-۶۱۷, ۲۰۰۲.

[۷] A. Strehl, J. Ghosh; *"Cluster ensembles-a knowledge reuse framework for combining multiple partitions"*, Edmonton, Alberta, Canada, Proc. Of ۱۱-th national conference on Artificial Intelligence, ۲۰۰۲.

[۸] G. Karypis, R. Aggarwal, R. Kumar, S. Shekhar; *"Multi-level HyperGraph partitioning: Applications in VLSI domain"*, Proc. Design and Automation Conference, ۱۹۹۷.

[۹] I.S. Dhillon, D.S. Modha; *"Concept decomposition for large sparse text data using clustering"*, Mach. Learning ۴۲ (۱), ۱۴۳-۱۷۵, ۲۰۰۱.

[۱۰] E.B Fowlkes, C.L Mallows; *"A method for comparing two hierarchical clusterings"*, J.Amer.Statist.Assoc.۷۸ (۳۸۳), ۵۵۳-۵۶۹, ۱۹۸۳.

[۱۱] L. Hubert, P. Arabie; *"Comparing partitions"*, J. Classification ۲(۱), ۱۹۳-۲۱۸, ۱۹۸۵.

[۱۲] X.Z Fren, C.E Brodley; *"Solving cluster ensemble problem by bipartite graph partitioning"*, NY, USA, ICML'۰۴:Proc. ۲۱st international Conf. on Machine Learning ACM press, ۲۰۰۴.

[۱۳] B. Hendrickson, R. Leland; *"An Improved Spectral Graph Partitioning Algorithm for Mapping Parallel Computations"*, Sandia National Laboratories, Albuquerque, NM, Tech. rep. SAND۹۲-۱۴۶۰, ۱۹۹۲.

[۱۴] A. Pothen, H. D. Simon, K.-P. Liou; *"Partitioning sparse matrices with eigenvectors of graphs"*, SIAM J. Matrix Anal. Appl., ۱۱, ۱۹۹۰.

[۱۵] A. Pothen, H. D. Simon, L. Wang, and S. T. Bernard; *"Towards a fast implementation of spectral nested dissection"*, Washington, DC, Supercomputing '۹۲ Proceedings, IEEE Computer Society Press, ۱۹۹۲.

[۱۶] E. Krupka, A. Navot, and N. Tishby, *"Learning to select features using their properties"*, Journal of

Machine Learning Research ۹, ۲۳۴۹-۲۳۷۶, ۲۰۰۸.

[۱۷] G. Chechik, G. Heitz, G. Elidan, P. Abbeel, D. Koller*, "Max-margin Classification of Data with Absent Features"*, Journal of Machine Learning Research ۹, ۱-۲۱, ۲۰۰۸.

زیرنویس

[i] Weighted Ensemble Clustering
[ii] Privacy
[iii] Ensemble Clustering
[iv] Clustering
[v] Clusterer
[vi] Cluster
[vii] Combiner
[viii] Feature
[ix] Label
[x] Graph-based methods
[xi] Greedy optimizing methods
[xii] Hyper-edge
[xiii] Hyper-graph
[xiv] Cluster-based Similarity Partitioning Algorithm (CSPA)
[xv] Hyper-Graph Partitioning Algorithm (HPGA)
[xvi] Meta-Clustering Algorithm (MCLA)
[xvii] Meta-cluster
[xviii] Global Optima
[xix] Reinforcement Learning Methods
[xx] Normalized Mutual Information
[xxi] Average Normalized Mutual Information
[xxii] Meta_feature