مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره سوم شماره سوم. ص ۲۱۳-۲۰۵، ۱۳۹۵

روش نوین خوشه‌بندی داده‌های بیان‌ژنی

داود شاهسونی[۱]*، زهره فرهادی[۲]

• **دریافت مقاله:** ۹۵/۸/۲۱ • **پذیرش مقاله:** ۹۵/۹/۲۲

مقدمه: یکی از تحولات مهم علم ژنتیک، ظهور فناوری ریزآرایه و تولید داده‌های بیان‌ژنی است که امکان مطالعه رفتار هزاران ژن را به طور همزمان فراهم می‌کند. خوشه‌بندی یکی از روش‌های داده‌کاوی است که در تحلیل داده‌های بیان‌ژنی مورد استفاده قرار می‌گیرد. از آنجا که عملکرد روش‌های خوشه‌بندی به شدت تحت تأثیر داده‌ها است، نتیجه خوشه‌بندی همواره با عدم قطعیت روبه‌رو بوده و الگوریتمی وجود ندارد که بتوان آن را برای تمام داده‌ها، کارا قلمداد نمود. در این تحقیق، در تحلیل داده‌های بیان‌ژنی از خوشه‌بندی اجماعی (ترکیب نتایج چندین الگوریتم خوشه‌بندی) به جای اجرای یک الگوریتم منفرد استفاده شده است.

روش: این مقاله عملکرد خوشه‌بندی اجماعی را بر روی سه مجموعه داده بیان‌ژنی Nutt-v3، Alizadeh-v2 وSU، توسط شاخص رند تعدیل یافته مورد ارزیابی قرار می‌دهد. برای پیاده‌سازی خوشه‌بندی اجماعی، دوازده خوشه‌بندی متفاوت حاصل از ترکیب چهار الگوریتم خوشه‌بندی با سه معیار عدم تشابه، به طور همزمان روی داده‌ها اجرا شده‌اند. پس از ادغام نتایج، میزان تطابق خوشه‌های تخمینی با گروه‌های واقعی توسط شاخص رند تعدیل یافته سنجیده شده است.

نتایج: مقدار شاخص رند تعدیل یافته برای سه مجموعه داده Nutt-v3، Alizadeh-v2 و SU، به ترتیب برابر ۱، ۰/۹ و ۰/۵۸به دست آمد که حاکی از دقت بالای روش پیشنهادی در کشف ساختارهای نهفته در داده‌ها است. همچنین الگوریتم طراحی شده، توانست تعداد واقعی خوشه‌ها را بدون خطا تشخیص دهد.

نتیجه‌گیری: خوشه‌بندی اجماعی روشی توانمند برای خوشه‌بندی داده‌های بیان‌ژنی است. با توجه به دقت این روش در کشف ساختارهای واقعی، می‌توان آن را با اطمینان جایگزین الگوریتم‌های خوشه‌بندی منفرد نمود.

کلیدواژه‌ها: داده‌کاوی، خوشه‌بندی اجماعی، خوشه‌بندی سلسله مراتبی، خوشه‌بندی افراز حول مدوید، مقیاس‌گذاری چند بعدی کلاسیک

• **ارجاع:** شاهسونی داود، فرهادی زهره. روش نوین خوشه‌بندی داده‌های بیان‌ژنی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۵؛ ۳(۳): ۲۰۵-۲۱۳.

۱. دکترای تخصصی آمار کاربردی، دانشیار گروه آمار، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود، شاهرود، ایران.

۲. کارشناس ارشد آمار ریاضی، دانشکده علوم ریاضی، دانشگاه صنعتی شاهرود، شاهرود، ایران.

* **نویسنده مسئول:** شاهرود، میدان هفت تیر، دانشگاه صنعتی شاهرود. کد پستی ۳۶۱۹۹۹۵۱۶۱

• **شماره تماس:** ۳۲۳۰۰۳۳۵-۰۲۳

• **Email:** dshahsavani@shahroodut.ac.ir

مقدمه

فناوری نوین ریزآرایه (Microarray) یکی از پیشرفت‌های مهم علم ژنتیک است که امکان مطالعه و تحلیل رفتار هزاران ژن را به طور هم‌زمان فراهم می‌کند. داده‌های حاصل از فناوری ریز آرایه، داده‌های بیان‌ژنی (Gene Expression Data) نامیده می‌شوند که تحلیل آن‌ها می‌تواند کمک شایانی را در زمینه تشخیص و درمان بیماری سرطان به جامعه پزشکی عرضه کند. داده‌های بیان‌ژنی در سال‌های اخیر بارها توسط دانشمندان علوم مختلف به ویژه محققان بیولوژی مولکولی مورد مطالعه و تحقیق قرار گرفته‌اند. استخراج اطلاعات معنی‌دار از بطن این داده‌ها، مسئله‌ای است که موجب تشریک مساعی علم آمار، داده‌کاوی و یادگیری ماشین شده است و هر یک با روش‌های خاص خود و نهایتاً مشترک، به حل این مسئله پرداخته‌اند. یکی از گونه‌های استخراج اطلاعات به‌منظور درک بهتر و قابل فهم‌تر این نوع داده‌ها، تقسیم‌بندی آن‌ها به دسته‌های کوچک‌تر با حداکثر تشابه و سپس تحلیل آن‌ها است که تحت عنوان "خوشه‌بندی"(Clustering) شناخته می‌شود. خوشه‌بندی به‌طور گسترده در تجزیه‌وتحلیل داده‌های بیان‌ژنی مورداستفاده قرار می‌گیرد به گونه‌ای که می‌توان علم پزشکی را یکی از دلایل پیشرفت و گسترش روش‌های خوشه‌بندی در دهه‌های اخیر دانست [۱].

هر یک از الگوریتم‌های خوشه‌بندی با بهینه‌سازی پارامترهایی از پیش تعریف شده، داده‌های مشابه را در گروه‌های مجزا از هم سازمان‌دهی می‌کنند. با وجود پیدایش الگوریتم‌های خوشه‌بندی نوین و کارآمد، روش فراگیری وجود ندارد که بتوان عملکرد آن را در همه داده‌ها یا حتی مجموعه‌ای از داده‌ها با ویژگی‌های یکسان، رضایت‌بخش تلقی نمود. از طرفی عملکرد الگوریتم‌های خوشه‌بندی به شدت تحت تأثیر ساختار پنهان داده‌ها قرار می‌گیرد و اجرای الگوریتم‌های مختلف روی مجموعه داده یکسان، می‌تواند منجر به نتایج کاملاً متفاوت گردد. حتی ممکن است اجرای یک الگوریتم خوشه‌بندی واحد اما با پارامترهای ورودی مختلف، نتایج متفاوتی را بر روی یک مجموعه داده‌ها ایجاد کند. با وجود مشکلات و محدودیت‌های ذکر شده، ارزیابی عملکرد الگوریتم‌های خوشه‌بندی همواره با عدم قطعیت همراه بوده و انتخاب یک الگوریتم قابل اطمینان و همچنین برآورد صحیح پارامترهای ورودی موردنیاز در یک الگوریتم خاص، همواره یک چالش برای محققین است. D'haeseleer در مقاله خود گفت شناسایی بهترین الگوریتم خوشه‌بندی به صورت کلی امکان‌پذیر نیست. وی اجرای چندین روش مختلف به ویژه روش‌هایی با نتایج متناقض را برای رسیدن به خوشه‌بندی بهینه پیشنهاد نمود [۲]. de Souto و همکاران به مقایسه پنج الگوریتم خوشه‌بندی (چهار الگوریتم سلسله مراتبی و یک الگوریتم افرازی) بر روی سی و پنج مجموعه داده بیان‌ژنی پرداختند [۳]، نتایج حاکی از آن بود که الگوریتم‌های مورد بررسی، همانند دیگر الگوریتم‌های خوشه‌بندی به شدت به داده‌ها وابسته هستند. همچنین نتایج پژوهش Jaskowiak و همکاران، با مقایسه عملکرد پانزده معیار، عدم تشابه را در خوشه‌بندی مجموعه داده‌های مذکور، نشان دادند که پارامترهای ورودی روش‌های خوشه‌بندی مانند ماتریس عدم تشابه به شدت بر نتیجه آن‌ها تأثیرگذار است [۴]. نتایج تحقیقات مذکور و پژوهش‌های مشابه، صحت این ادعا را تأیید می‌کنند که انتخاب روش خوشه‌بندی مناسب و معیار عدم تشابه یکی از چالش‌های مهم خوشه‌بندی است.

برای اولین بار در سال ۲۰۰۲ ایده ترکیب نتایج خوشه‌بندی‌های متفاوت و پذیرفتن اجماع آن‌ها به عنوان نتیجه نهایی، تحت عنوان کلی خوشه‌بندی اجماعی (Clustering Ensemble) توسط Strehl و Ghosh مطرح شد [۵]. خوشه‌بندی اجماعی راه‌حلی کارا برای مواجهه با چالش‌ها و محدودیت‌های ذکر شده است که با هدف افزایش کیفیت و دقت خوشه‌بندی اجرا می‌شود و نتیجه پایدارتری را نسبت به یک خوشه‌بندی منفرد پدید می‌آورد. از آنجا که در این روش، خطای یک خوشه‌بندی می‌تواند توسط دیگر خوشه‌بندی‌ها جبران شود، نتیجه‌ای قابل اطمینان‌تر در مقایسه با خوشه‌بندی منفرد حاصل می‌شود. به طور کلی خوشه‌بندی اجماعی در مواقعی که کاربر در انتخاب روش خوشه‌بندی یا پارامترهای ورودی دچار تردید است، بهترین گزینه است [۶،۷].

تحقیقات متعددی وجود دارند که به معرفی خوشه‌بندی اجماعی و انواع روش‌های پیاده‌سازی آن، پرداخته‌اند. به عنوان نمونه می‌توان به پژوهش‌های قائمی و همکاران، Li و همکاران و Vega-Pons و Shulcloper Ruiz- اشاره کرد [۸-۶]. نمونه‌هایی از کاربرد خوشه‌بندی اجماعی در تحلیل داده‌های بیان‌ژنی را می‌توان در پژوهش‌های Deodhar و Ghosh، Hu و همکاران، Yu و Wong، Souto و همکاران، Yu و همکاران همچنین Kim و همکاران مشاهده کرد [۹-۱۴]. نتایج موفقیت‌آمیز

پژوهش‌های مذکور، تصدیقی بر عملکرد مؤثر و مفید خوشه‌بندی اجماعی است.

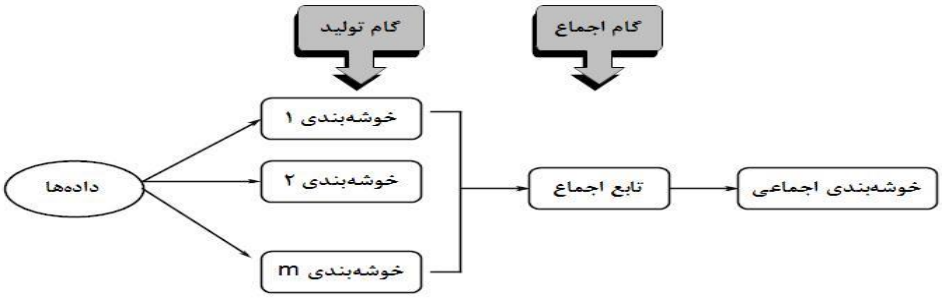
هدف از این پژوهش، بهبود کارآیی و رفع نقایص روش‌های خوشه‌بندی منفرد توسط خوشه‌بندی اجماعی در تحلیل داده‌های بیان ژنی است. بدین منظور، چهار روش خوشه‌بندی متفاوت بر روی سه مجموعه داده بیان ژنی به طور همزمان اعمال شده‌اند. روش‌های مورد مطالعه، به طور کلی مبتنی بر عدم تشابه بین داده‌ها هستند که با تغییر معیار عدم تشابه، هر یک از آن‌ها نتیجه‌ای متفاوت خواهند داشت. از این رو با هدف افزایش کیفیت خوشه‌بندی در هریک از چهار الگوریتم مذکور، سه نوع عدم تشابه متفاوت اعمال گردیده است. بدین ترتیب خوشه‌بندی اجماعی شامل دوازده ترکیب متفاوت خواهد بود. علاوه بر ارزیابی عملکرد خوشه‌بندی اجماعی در کشف گروه‌های واقعی، در این تحقیق قصد داریم توانایی خوشه‌بندی اجماعی را در برآورد تعداد خوشه‌ها مورد بررسی قرار دهیم.

ساختار این مقاله این گونه تنظیم شده است که ابتدا به معرفی خوشه‌بندی اجماعی می‌پردازیم. در ادامه پس از معرفی اجمالی داده‌ها، جزییات الگوریتم مورد مطالعه را در بخش یافته‌ها بیان می‌کنیم و در پایان، ضمن ذکر نتایج پیاده‌سازی الگوریتم پیشنهادی، عملکرد آن را مورد ارزیابی قرار داده و به بحث و نتیجه‌گیری می‌پردازیم. لازم به ذکر است که تمامی محاسبات انجام شده در این مقاله در نرم‌افزار R صورت گرفته است.

روش

خوشه‌بندی اجماعی

خوشه‌بندی اجماعی شامل دو گام اصلی موسوم به تولید (Generation) و اجماع (Consensus) است. در شکل ۱ ساختار کلی خوشه‌بندی اجماعی نمایش داده شده است. در مرحله تولید، با هدف بهبود کیفیت و دقت نتایج نهایی، چندین خوشه‌بندی که آن‌ها را خوشه‌بندی‌های مبنا (Base clusterings) نام‌گذاری می‌کنند، روی داده‌ها اجرا می‌شوند. این خوشه‌بندی‌های منفرد می‌توانند از یک نوع واحد ولی با مقادیر مختلفی برای پارامترها، یا کلاً خوشه‌بندی‌های متفاوتی باشند. از آنجا که نتیجه نهایی وابسته به نتایج خوشه‌بندی‌های مبنا است، نقش مرحله تولید بسیار حائز اهمیت است، به طوری که هر چه تنوع بین خوشه‌بندی‌های مبنا بیشتر باشد، نتیجه خوشه‌بندی اجماعی از کیفیت بالاتری برخوردار خواهد بود [۱۵].



شکل ۱: ساختار خوشه‌بندی اجماعی

در چگونگی تولید خوشه‌بندی‌های مبنا و ایجاد تنوع بین آن‌ها، هیچ محدودیتی وجود ندارد و روش‌های زیادی برای این منظور وجود دارند که در ذیل به تعدادی از آن‌ها اشاره شده است [۷]:

الف–استفاده از الگوریتم‌های متفاوت خوشه‌بندی

ب– استفاده از یک الگوریتم خوشه‌بندی واحد با پارامترهای ورودی متفاوت

ج– انتخاب تصادفی زیر مجموعه‌ای از مجموعه متغیرها برای هر خوشه‌بندی مبنا

د–انتخاب تصادفی مجموعه‌ای از داده‌ها برای هر خوشه‌بندی مبنا

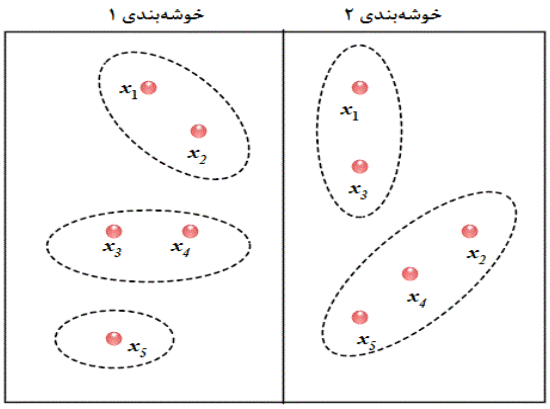
در مورد روش‌های (الف)، (ب) و (ج) حجم نمونه در خوشه‌بندی‌های مبنا برابر با حجم واقعی داده‌ها در نظر گرفته می‌شود. در حالت (د) در هر یک از خوشه‌بندی‌های مبنا، نمونه‌ای تصادفی از داده‌ها که حجم آن توسط کاربر تعیین می‌شود، انتخاب می‌شود تا بدین ترتیب بتوان، بین خوشه‌بندی‌های مبنا تنوع ایجاد کرد، در این مقاله، رهیافت الف برای اجرای گام تولید مورد توجه قرار گرفته است. بدین منظور از چهار الگوریتم خوشه‌بندی مرسوم شامل یک الگوریتم

خوشه‌بندی افرازی به نام "افراز حول مدوید" (Partition around medoids) و سه الگوریتم خوشه‌بندی سلسله linkage) استفاده شده است؛ که به ترتیب آن‌ها را با نمادهای PAM، SL، CL و AL نمایش می‌دهیم [۱۶].

در مرحله اجماع، خروجی‌های مرحله تولید که در واقع همان نتایج خوشه‌بندی‌های مبنا هستند، با هم ادغام شده و نتیجه نهایی را رقم می‌زنند. جهت ادغام نتایج مرحله تولید، توابعی موسوم به توابع اجماع (Consensus Functions) طراحی شده‌اند. انتخاب تابع اجماع مناسب یکی از موارد مهم در فرآیند خوشه‌بندی اجماعی است، زیرا حتی در صورتی که خوشه‌بندی‌های مبنا، عملکرد ضعیفی را از خود نشان داده باشند، با انتخاب تابع اجماع مناسب می‌توان نتیجه نهایی را تا حد قابل ملاحظه‌ای بهبود بخشید [۱۷]. در این مقاله، تابع اجماع مبتنی بر تشابه دو به دو (Pairwise similarity approach) مورد ارزیابی قرار گرفته است.

در تابع اجماع مبتنی بر تشابه دو به دو ابتدا نتایج خوشه‌بندی های مبنا، در قالب ماتریسی موسوم به ماتریس اطلاعات (Information Matrix) خلاصه می‌شوند و سپس در اختیار یک روش خوشه‌بندی متناسب قرار می‌گیرند [۱۵]. در این مقاله از ماتریس تشابه دو به دو یا هم پیوندی (co-association matrix) با نماد اختصاری CO به عنوان مراتبی به نام‌های تک اتصالی (Single linkage)، اتصال کامل (Complete linkage) و اتصال میانگین (Average) ورودی الگوریتم خوشه‌بندی مبتنی بر تشابهات PAM استفاده شده است.

نحوه تشکیل ماتریس مربعی CO با بعد n به این صورت است که با فرض داشتن تعداد m خوشه‌بندی مبنا در مرحله تولید و داشتن n مشاهده، درآیه ij ام آن برابر با نسبت تعداد دفعاتی در بین m دفعه است که دو مشاهده i و j در هر یک از خوشه‌بندی‌های مبنا، در یک خوشه واحد قرار گرفته‌اند و درآیه‌های قطر اصلی آن برابر با عدد ۱ است. در شکل ۲ نحوه تشکیل ماتریس CO برای مثالی ساده شامل دو خوشه‌بندی مبنا، نمایش داده شده است. لازم به ذکر است که x_i $i = 1,2, \dots, p$ معرف i-امین سطر ماتریس داده‌های بیان‌ژنی $X_{n*p} = [x_{ij}]$ است که در آن p تعداد ژن‌ها و n تعداد داده‌ها (نمونه‌ها) است. همچنین x_{ij} بیانگر سطح بیان ژن j از داده i می‌باشد؛ بنابراین $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ با تفاضل ماتریس CO از ماتریس ۱(ماتریسی که کلیه درآیه‌های آن برابر با عدد ۱ است) یعنی محاسبه ماتریس CO-۱، این ماتریس تشابه به یک ماتریس "عدم تشابه" تبدیل شده و می‌تواند برای انجام خوشه‌بندی نهایی در روش‌های خوشه‌بندی مبتنی بر عدم تشابه اعمال گردد [۶, ۷, ۱۵].



ماتریس CO

	x_1	x_2	x_3	x_4	x_5
x_1	1	0.5	0.5	0	0
x_2		1	0	0.5	0.5
x_3			1	0.5	0
x_4				1	0.5
x_5					1

شکل ۲: مثالی ساده از تشکیل ماتریس تشابه دو به دو CO برای ۵ مشاهده توسط دو خوشه‌بندی مبنا [۱۵]

تخمین تعداد خوشه‌ها

یکی از روش‌های پیشنهادی برای تخمین تعداد خوشه‌ها استفاده از میانگین شاخص نیمرخ (Silhouette index) است که میزان تراکم درون خوشه‌ای را می‌سنجد و مقدار آن در بازه [۱-،۱] متغیر است. برابری مقدار این شاخص با عدد ۱ نشان می‌دهد که خوشه‌ها متراکم هستند و به خوبی از هم جدا شده‌اند، لذا می‌توان نتیجه گرفت که مدل خوشه‌بندی ایجاد شده مناسب است. همچنین مقادیر صفر و منفی در مورد این

شاخص، عدم توانایی مدل خوشه‌بندی را در یافتن داده‌های مشابه نشان می‌دهد [۱۸].

داده‌های تحقیق

در این تحقیق از سه مجموعه داده بیان‌ژنی مرسوم در تحقیقات پزشکی استفاده شده است. معلوم بودن گروه‌های واقعی در این داده‌ها موجب سهولت ارزیابی نتیجه خوشه‌بندی خواهد شد. در ذیل توضیح مختصری در مورد هر یک از مجموعه داده‌های مورد استفاده در این تحقیق ارائه شده است:

Nutt و همکاران مقاله‌ای منتشر کردند که در آن با استفاده از تکنیک ریزآرایه ۱۱۵۲ ژن مورد تحلیل قرار گرفت. تعداد بیماران مورد مطالعه در این تحقیق ۲۲ نفر بوده است و نتایج منجر به تعیین دو زیر رده برای تومور مغزی شد [۱۹].

Alizade و همکاران، تعداد ۲۰۹۳ ژن از ۶۲ بیمار را برای شناسایی زیرگروه‌های سرطان خون مورد مطالعه قرار دادند. نتایج نشان دهنده سه زیرگروه برای این سرطان بود [۲۰].

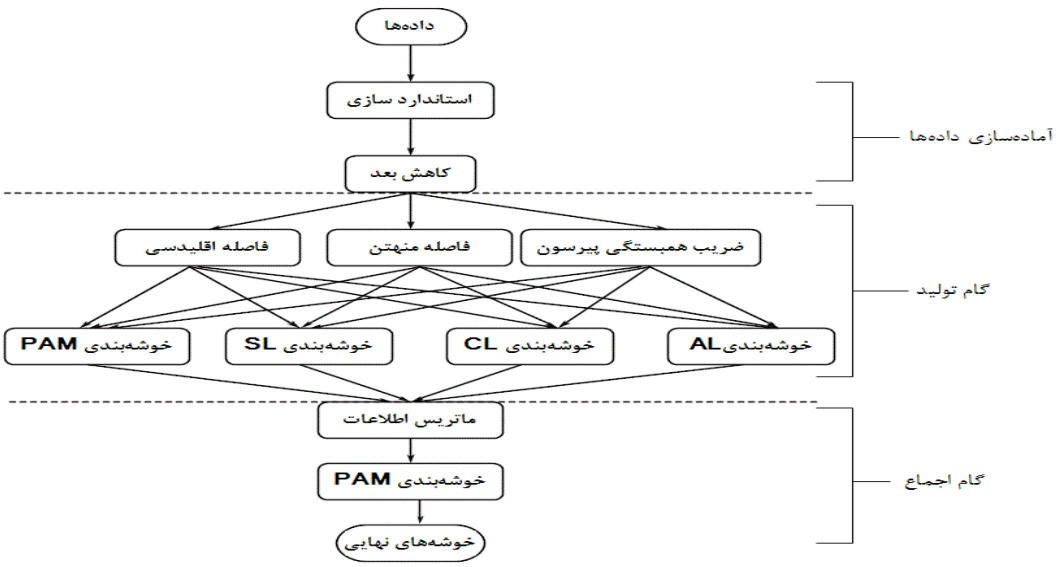
Su و همکاران مطالعه‌ای را به منظور طبقه‌بندی ۱۰ نوع از تومورهای انسانی انجام دادند. نویسندگان در مقاله مذکور ۱۵۷۱ ژن مربوط به یک نمونه ۱۷۴ تایی را مورد مطالعه قرار دادند [۲۱].

خلاصه مشخصات داده‌های تحقیق در جدول ۱ درج شده است.

جدول ۱: مشخصات داده‌های تحقیق

نام داده	تعداد ژن	تعداد نمونه	تعدادخوشه
Nutt-v3	۱۱۵۲	۲۲	۲
Alizadeh-v2	۲۰۹۳	۶۲	۳
Su	۱۵۷۱	۱۷۴	۱۰

در شکل ۳، روند اجرای الگوریتم تحقیق در قالب یک فلوچارت نمایش داده شده است. در اولین مرحله از این الگوریتم، پیش‌پردازش داده‌ها صورت می‌گیرد. بدین منظور پس از استانداردسازی داده‌ها با هدف کاهش هزینه محاسباتی و سهولت تحلیل، بُعد داده‌ها کاهش داده می‌شود. روش مقیاس گذاری چند بعدی (Multidimensional Scaling) کلاسیک یکی از روش‌های کاهش بعد است که با دریافت ماتریس عدم تشابه داده‌ها و بعد دلخواه، به ساخت پیکره‌ای از آن‌ها در ابعاد پایین‌تر می‌پردازد به طوری که تشابه بین آن‌ها حفظ شود [۲۲]. ذکر این نکته ضروری است که انتخاب نامناسب بعد، منجر به از دست رفتن اطلاعات و نتایج گمراه کننده خواهد شد. در نرم‌افزار R با دستوراتی ساده، امکان تعیین بعد مناسب برای روش مقیاس‌گذاری چندبعدی کلاسیک از طریق محاسبه معیارهای نیکویی برازش که بدین منظور تعبیه شده‌اند، فراهم شده است [۲۳].



شکل ۳: الگوریتم تحقیق

در اینجا بعد مناسب جهت کاهش تعداد ژن‌ها با شرط حفظ حداکثر اطلاعات، برای سه مجموعه داده Nutt-v3، Alizadeh-v2 و SU به ترتیب برابر با ۲۱، ۶۰ و ۱۷۲ به دست آمد و سپس با اعمال روش مقیاس‌گذاری چند بعدی کلاسیک، مختصات جدید داده‌ها در فضای تقلیل یافته محاسبه شده است.

در گام تولید خوشه‌بندی اجماعی، چهار الگوریتم SL، CL، AL و PAM به عنوان خوشه‌بندی‌های مبنا بر روی داده‌ها اجرا شده‌اند. هر یک از این الگوریتم‌ها برای انجام خوشه‌بندی نیازمند دریافت عدم تشابه داده‌ها هستند. در این تحقیق، عدم تشابه داده‌ها از طریق سه معیار فاصله اقلیدسی، فاصله منهتن و ضریب همبستگی پیرسون محاسبه شده است. در واقع هر کدام از چهار الگوریتم مذکور، با دریافت سه معیار عدم تشابه مختلف، سه خوشه‌بندی متفاوت انجام داده‌اند؛ بنابراین در مرحله تولید، دوازده خوشه‌بندی متفاوت، روی داده‌ها اجرا گردیده است. برای استفاده از روش‌های سلسله مراتبی و افرازی بایستی تعداد خوشه‌ها مشخص باشد. در هر یک از خوشه‌بندی‌های مبنا، میانگین شاخص نیمرخ برای تعداد مختلف خوشه‌ها $(k=2,3,\ldots,\sqrt{n})$ محاسبه شده و در نهایت تعداد خوشه‌ای که منجر به بیشینه شدن مقدار این شاخص گردیده، به عنوان مقدار بهینه انتخاب شده است.

در مرحله اجماع، نتایج خوشه‌بندی‌های اجرا شده در مرحله تولید، در قالب ماتریس تشابه CO خلاصه شده و در نهایت پس از تبدیل آن به ماتریس عدم تشابه، در الگوریتم خوشه‌بندی PAM اعمال گردیده است. در این مرحله نیز همانند گام تولید، از میانگین شاخص نیمرخ برای برآورد تعداد خوشه‌های مناسب استفاده شده است. نتیجه اجرای الگوریتم PAM که در واقع نتیجه نهایی محسوب می‌شود، توسط شاخص رند تعدیل یافته (Adjusted rand index) مورد ارزیابی قرار گرفته است. این شاخص تعیین کننده میزان تطابق گروه‌های واقعی با خوشه‌های تخمینی است و مقداری در بازه [۱،۱−] اختیار می‌کند که عدد ۱ نشان‌دهنده هم‌پوشانی کامل گروه‌های واقعی با خوشه‌های تخمینی است. همچنین مقادیر صفر و منفی عملکرد بسیار ضعیف روش خوشه‌بندی را نشان می‌دهند [۲۴]. جدول ۲ مقادیر این شاخص و تعداد خوشه‌های برآورد شده در مرحله نهایی را برای هر یک از سه داده، نشان می‌دهد.

جدول ۲: نتایج خوشه‌بندی

نام داده	تعداد خوشه‌های تخمینی	مقدار شاخص رند تعدیل یافته
Nutt-v3	۲	۱
Alizadeh-v2	۳	۰/۹
Su	۱۰	۰/۵۸

بحث و نتیجه‌گیری

عملکرد الگوریتم‌های متفاوت خوشه‌بندی بارها در تحلیل داده‌های بیان‌ژنی مورد ارزیابی قرار گرفته‌اند و تلاش برای رسیدن به خوشه‌بندی بهینه توسط محققان کماکان ادامه دارد. در میان انواع مختلف روش‌های خوشه‌بندی، الگوریتم‌های سلسله مراتبی و افرازی به دلیل عملکرد قابل قبولی که نسبت به سهولت سازوکار و اجرا دارند، بیش از دیگر الگوریتم‌ها موردتوجه قرار گرفته‌اند. شاکری و همکاران، دو الگوریتم خوشه‌بندی سلسله مراتبی و یک الگوریتم خوشه‌بندی افرازی را به صورت منفرد برای خوشه‌بندی پنج مجموعه داده بیان ژنی به کار گرفتند. آن‌ها در هریک از این الگوریتم‌ها سه ماتریس عدم تشابه فاصله اقلیدسی، فاصله منهتن و ضریب همبستگی پیرسون را اعمال کردند. نتایج این تحقیق نشان داد که بهترین روش خوشه‌بندی به صورت مطلق وجود ندارد و یک الگوریتم خوشه‌بندی واحد بر اساس شرایط موجود در داده‌ها ممکن است عملکردهای کاملاً متفاوتی را از خود بروز دهد [۲۵].

Iam-On و همکاران، خوشه‌بندی اجماعی را بر روی ده مجموعه داده بیان ژنی اجرا کردند. آن‌ها در مرحله تولید (شکل ۱)، از الگوریتم خوشه‌بندی k– میانگین بر مبنای فاصله اقلیدسی استفاده کردند و پارامترهای ورودی متفاوت را عامل تنوع میان خوشه‌بندی‌های مبنا قرار دادند. Iam-On و همکاران، چندین ماتریس اطلاعات را برای خلاصه‌سازی نتایج مرحله تولید معرفی نموده و در گام اجماع، رویکرد مبتنی بر تشابه دو به دو را مورد توجه قرار دادند. استفاده از ماتریس‌های اطلاعات متفاوت، منجر به کشف خوشه‌های متفاوت در خوشه‌بندی نهایی گردید که در بهترین خوشه‌بندی صورت گرفته، میانگین شاخص رند تعدیل یافته برای ده مجموعه داده

موردبررسی، حدود ۰/۴ گزارش شد که حاکی از عملکرد نامطلوب مدل خوشه‌بندی مذکور در تشخیص گروه‌های واقعی است [۲۶].

یکی از مهم‌ترین چالش‌هایی که در تحلیل داده‌های بیان ژنی با آن مواجه هستیم، تعداد بسیار زیاد ژن‌ها نسبت به حجم نمونه است که این موضوع کارآیی معیارهای عدم تشابه و همچنین روش‌های خوشه‌بندی را کاهش می‌دهد. از طرفی اهمیت فراوان تحلیل این داده‌ها در تشخیص و درمان سرطان، حساسیت در انتخاب معیار عدم تشابه و روش خوشه‌بندی را دو چندان می‌کند؛ از این رو در تحلیل داده‌های بیان ژنی نیاز به مدل خوشه‌بندی که نتیجه‌ای قابل اعتماد و پایدار داشته باشد و بتواند از قابلیت‌های چندین معیار عدم تشابه و روش خوشه‌بندی بهره ببرد، بیش از سایر کاربردها احساس می‌شود. برآورد تعداد خوشه‌ها، یکی دیگر از چالش‌های پیش رو در خوشه‌بندی داده‌های بیان ژنی است که با توجه به ساختار پیچیده این نوع داده‌ها، کاری دشوار بوده و احتمال خطا، بسیار زیاد است. الگوریتم طراحی شده در این مقاله با هدف ارائه راه‌حلی برای چالش‌های فوق بر روی سه مجموعه داده بیان‌ژنی مورد ارزیابی قرار گرفت. برای تحقق اهداف تحقیق و آزمودن الگوریتم پیشنهادی، داده‌های تحقیق با ساختارها و تعداد گروه‌های متفاوت انتخاب شده‌اند.

در این تحقیق پس از استانداردسازی و کاهش بعد داده‌ها، توسط رویکرد خوشه‌بندی اجماعی، دوازده خوشه‌بندی متفاوت به طور هم‌زمان روی سه مجموعه داده بیان‌ژنی اجرا گردید که از ترکیب چهار الگوریتم خوشه‌بندی (سه الگوریتم خوشه‌بندی سلسله مراتبی و یک الگوریتم خوشه‌بندی افرازی) با سه معیار عدم تشابه (فاصله اقلیدسی، فاصله منهتن و ضریب همبستگی پیرسون) حاصل شدند. نتایج تحقیق مندرج در جدول ۲، حاوی تعداد خوشه‌های برآورد شده در مرحله نهایی و شاخص رند تعدیل یافته برای هریک از مجموعه داده‌های جدول ۱ است. این نتایج نشان می‌دهد که الگوریتم پیشنهادی توانسته است در مورد مجموعه داده Nutt-v3 بدون هیچ خطایی خوشه‌های واقعی را تشخیص دهد. همچنین مقدار ۰/۹ برای شاخص رند تعدیل یافته در مورد داده Alizade-v2 بیانگر خوشه‌بندی مطلوب و تطابق مناسب خوشه‌های تخمینی با گروه‌های واقعی است. بدیهی است هر چه تعداد گروه‌های واقعی در مجموعه داده‌ها بیشتر باشد، کشف ساختارهای موجود دشوارتر بوده و خوشه‌بندی با پیچیدگی مواجه خواهد شد. لذا با توجه به تعداد زیاد گروه‌های واقعی در مورد مجموعه داده Su، پایین بودن مقدار شاخص رند تعدیل یافته در مورد این داده نسبت به دو مجموعه داده دیگر، قابل توجیه است و می‌توان خوشه‌بندی صورت گرفته را رضایت‌بخش تلقی کرد. شایان ذکر است که الگوریتم پیشنهادی توانسته است در مورد هر سه مجموعه داده، تعداد خوشه‌ها را به طور صحیح و دقیقاً برابر با تعداد واقعی خوشه‌ها تخمین بزند. با توجه به بالاتر بودن مقادیر شاخص رند تعدیل یافته مربوط به این تحقیق در مقایسه با پژوهش Iam-On و همکاران [۲۶]، می‌توان نتیجه گرفت که الگوریتم پیشنهادی توانسته است خوشه‌بندی را با دقت بیشتری انجام دهد. مزیت دیگر الگوریتم طراحی شده در این مقاله، سادگی ساختار و سهولت اجرا است.

نتایج حاصل از این تحقیق نشان داد که با استفاده از خوشه‌بندی اجماعی و به کارگیری همزمان الگوریتم‌های متفاوت خوشه‌بندی می‌توان کیفیت و دقت خوشه‌بندی را تا حد قابل ملاحظه‌ای افزایش داد. با توجه به نتایج این مطالعه، خوشه‌بندی اجماعی قادر است حتی با ترکیب الگوریتم‌های ساده، به کشف خوشه‌های بهینه منجر شود. همچنین از آنجا که در این رویکرد، چندین الگوریتم خوشه‌بندی در برآورد تعداد خوشه‌ها سهیم خواهند شد، خطای یک الگوریتم توسط دیگر الگوریتم‌ها قابل جبران است و احتمال خطا تا حد زیادی کاهش پیدا می‌کند؛ بنابراین خوشه‌بندی اجماعی می‌تواند راه‌حلی توانمند برای حل چالش‌های پیش رو در خوشه‌بندی باشد و کاستی‌های خوشه‌بندی منفرد را رفع کند. یکی از مزیت‌های مهم خوشه‌بندی اجماعی این است که در انتخاب و تعداد الگوریتم‌های خوشه‌بندی که قرار است با هم ترکیب شوند هیچ محدودیتی وجود ندارد و هرچه تنوع خوشه‌بندی‌های مبنا بیشتر باشد، کیفیت نهایی بالاتر است. در کنار قابلیت‌های خوشه‌بندی اجماعی این نکته را نیز باید در نظر داشت که اجرای این روش به زمان و محاسبات بیشتری نیاز دارد. به عنوان پیشنهادی برای آینده تحقیق می‌توان در مرحله تولید خوشه‌بندی اجماعی از ترکیب الگوریتم‌های پیشرفته خوشه‌بندی نظیر الگوریتم‌های فازی، مبتنی بر شبکه و مبتنی بر چگالی استفاده کرد. همچنین در مرحله اجماع به جای استفاده از رویکرد مبتنی بر تشابه دو به دو، از روش‌های دیگری همچون مبتنی بر گراف (Graph-based approach) و مبتنی بر ویژگی (Feature-based approach) برای ادغام نتایج خوشه‌بندی‌های مبنا بهره گرفت.

References

1. Jiang DC, Tang, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Transactions on Knowledge and Data Engineering; 2004 Oct 4; IEEE; 2004. p. 1370 – 86.

2. D'haeseleer P. How does gene expression clustering work? Nat Biotechnol. 2005;23(12):1499-501.

3. de Souto MC, Costa IG, de Araujo DS, Ludermir TB, Schliep A. Clustering cancer gene expression data: a comparative study. BMC Bioinformatics 2008;9:497.

4. Jaskowiak PA, Campello RJ, Costa IG. On the selection of appropriate distances for gene expression data clustering. BMC Bioinformatics 2014;15 Suppl 2:S2.

5. Strehl A, Ghosh J. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. J Mach Learn Res 2003;3:583-617.

6. Ghaemi R, Sulaiman N, Ibrahim H, Mustapha N. A survey: clustering ensembles techniques. Proceedings of World Academy of Science, Engineering and Technology 2009; 38: 644-53.

7. Li T, Ogihara M, Ma S. On combining multiple clusterings: an overview and a new perspective. Applied Intelligence 2010;33(2):207-19.

8. Vega-Pons S, Ruiz-Shulcloper J. A Survey of Clustering Ensemble Algorithms. International Journal of Pattern Recognition 2011; 25(3): 337-72.

9. Deodhar M, Ghosh J. Consensus Clustering for Detection of Overlapping Clusters in Microarray Data. Proceedings of the Sixth IEEE International Conference on Data Mining-Workshops; 2006 Dec 18-22; Washington, DC, USA: IEEE; 2006.

10. Hu X, Park EK, Zhang X. Microarray gene cluster identification and annotation through cluster ensemble and EM-based informative textual summarization. IEEE Trans Inf Technol Biomed 2009;13(5):832-40.

11. Yu Z, Wong HS. Class discovery from gene expression data based on perturbation and cluster ensemble. IEEE Trans Nanobioscience 2009;8(2):147-60.

12. Souto MC, Araujo DS, Silva BL. Cluster Ensemble for Gene Expression Microarray Data: Accuracy and Diversity. International Joint Conference on Neural Network Proceedings; 2006 Jul 16-21; Canada: IEEE; 2006

13. Yu Z, Wong HS, Wang H. Graph-based consensus clustering for class discovery from gene expression data. Bioinformatics 2007;23(21):2888-96.

14. Kim EY, Kim SY, Ashlock D, Nam D. MULTI-K: accurate classification of microarray subtypes using ensemble k-means clustering. BMC Bioinformatics 2009;10:260.

15. Iam-On N, Boongoen T. Comparative study of matrix refinement approaches for ensemble clustering. Machine Learning 2015;98(1):269-300.

16. Gan G, Ma C, Wu J. Data Clustering: Theory, Algorithms, and Applications. Philadelphia: SIAM, Society for Industrial and Applied Mathematics; 2007.

17. Topchy A, Jain AK, Punch W. Clustering ensembles: models of consensus and weak partitions. IEEE Trans Pattern Anal Mach Intell 2005;27(12):1866-81.

18. Yang C, Wan B, Gao X. Effectivity of Internal Validation Techniques for Gene Clustering. In: Maglaveras N, Chouvarda I, Koutkias V, Brause R, editors. Biological and Medical Data Analysis: 7th International Symposium, ISBMDA 2006, Thessaloniki, Greece; 2006 Dec 7-8; Berlin, Heidelberg: Springer Berlin Heidelberg; 2006. p. 49-59.

19. Nutt CL, Mani DR, Betensky RA, Tamayo P, Cairncross JG, Ladd C, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res 2003;63(7):1602-7.

20. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403(6769):503-11.

21. Su AI, Welsh JB, Sapinoso LM, Kern SG, Dimitrov P, Lapp H, et al. Molecular classification of human carcinomas by use of gene expression signatures. Cancer Res 2001;61(20):7388-93.

22. Rencher AC, Christensen WF. Methods of Multivariate Analysis. 3th ed. NewJersey: Wiley; 2012.

23. Everitt B, Hothorn T. An Introduction to Applied Multivariate Analysis with R. 1th ed. New York: Springer; 2011.

24. Warrens MJ. On the Equivalence of Cohen's Kappa and the Hubert-Arabie Adjusted Rand Index. Journal of Classification 2008;25(2): 177-83.

25. Shakeri M, Sabaghian E, Esmaeili H. CCK (Clustering-Classification-Kappa) a new validation index to assessing clustering results of gene expression data. Journal of North Khorasan University of Medical Sciences 2012;3(5) :67-78. Persian

26. Iam-on N, Boongoen T, Garrett S. LCE: a link-based cluster ensemble method for improved gene expression data analysis. Bioinformatic 2010;26(12):1513-9.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2016; 3(2): 205-213

A Novel Method of Gene Expression Data Clustering

Davood Shahsavani[1]*, Zohreh Farhadi[2]

• **Received:** 11 Nov, 2016 • **Accepted:** 12 Dec, 2016

Introduction: The microarray technology and production of gene expression data are among the important developments in genetic science that provide ability to study the behavior of thousands of genes, simultaneously. Clustering is one of the most important data mining techniques used in gene expression data analysis. As, the performance of clustering methods is strongly affected by the structure of data, the result of clustering is always uncertain and there is no algorithm that can be used for all kinds of data. In this study, ensemble clustering (combined results of multiple clustering algorithms) was used for gene expression data analysis rather than using a single algorithm.

Methods: The performance of ensemble clustering in three gene expression data sets, Nutt-v3, Alizadeh-v2 and SU, were evaluated by adjusted Rand index. Twelve different clusterings resulted from the combination of four clustering algorithms with three dissimilarity matrices were simultaneously applied on data. After merging the results, and running the final clustering, the estimated clusters were compared with actual groups by the adjusted Rand index.

Results: The adjusted Rand index for the three data sets of Nutt-v3, Alizadeh-v2 and SU, were respectively 1, 0.9 and 0.58 which shows the remarkable accuracy of the proposed method in detecting patterns in data sets. Moreover, the designed algorithm could detect the actual number of clusters without errors.

Conclusion: Ensemble clustering is a powerful and reliable method for gene expression data analysis. Due to the accuracy and quality of this method in detection of real data structures, it can be replaced the individual clustering algorithms.

Keywords: Data mining, Ensemble clustering, Hierarchical clustering, Partition around medoids, Classic multidimensional scaling

• **Citation:** Shahsavani D, Farhadi Z. A Novel Method of Gene Expression Data Clustering. Journal of Health and Biomedical Informatics 2016; 3(2): 205- 213.

1. PhD, Associate Professor of Statistics, School of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran
2. MSc of Statistics, School of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran

***Correspondence**: Shahrood, Hafte-Tir Square, Shahrood University of Technology, Postal code 3619995161
• **Tel:** 023-32300335 • **Email:** dshahsavani@shahroodut.ac.ir