مجله دانشگاه علوم پزشکی خراسان شمالی
۱۳۹۰؛ ۳ (ویژه نامه آمار زیستی و اپیدمیولوژی):۷۸- ۶۷

Journal of North Khorasan University of Medical Sciences
2011;3(Biostatistics and Epidemiology Supplement):67-87

مقایسه روش‌های خوشه‌بندی در داده‌های بیان ژنی

محمد تقی شاکری[۱]، احسان صباغیان[۲]، حبیب الله اسماعیلی[*۱]*

[۱]دانشیار گروه آمار زیستی و اپیدمیولوژی، دانشکده بهداشت، دانشگاه علوم پزشکی مشهد
[۲]کارشناس ارشد آمار زیستی، دانشگاه علوم پزشکی مشهد
***نویسنده مسئول:** دانشکده بهداشت، دانشگاه علوم پزشکی مشهد
پست الکترونیکُ EsmailyH@mums.ac.ir

چکیده

زمینه و هدف: *با گسترش روش‌های استخراج داده‌های ژنتیکی، روش‌های تجزیه و تحلیل این نوع داده‌ها نیز در حال توسعه می‌باشند. این مطالعه با هدف مقایسه یکی از پرکاربردترین روش‌های تجزیه و تحلیل این نوع داده ها، یعنی خوشه‌بندی انجام شده است.*

مواد و روش کار: *در این پژوهش با استفاده از ۵ مجموعه داده مایکروآرایه، نه ترکیب روش خوشه‌بندی سلسله مراتبی تجمعی، سلسله مراتبی تقسیم‌شونده و K- میانگین با متریک‌های فاصله اقلیدسی، منهاتان و ضریب همبستگی پیرسون پیوند و با استفاده از شاخص پهنای نیمرخ با استفاده از روش نمونه‌گیری بوت استرپ مقایسه شده است.*

یافته‌ها: *نتایج نشان داد روش خوشه‌بندی سلسله مراتبی تجمعی با پیوند متوسط دارای بهترین عملکرد بود. همچنین این روش در مقایسه با دیگر روش‌ها از پایایی بیشتری برخوردار بود. درعین حال روش خوشه‌بندی سلسله مراتبی تقسیم شونده عملکرد نسبتاً مشابهی با روش خوشه‌بندی K- میانگین داشته است.*

نتیجه‌گیری: *با توجه به نتایج می توان گفت که مبتنی بر شرایط موجود در داده ها بهترین روش خوشه‌بندی انتخاب می شود.*

واژه‌های کلیدی: *خوشه‌بندی، ریزآرایه، بوت استرپ، شاخص‌های ارزیابی نتایج روش‌های خوشه‌بندی*

مقدمه

سلول را می‌توان به عنوان یک ماشین بسیار پیچیده در نظر گرفت. تا کنون تنها بخش‌هایی از این ماشین پیچیده به صورت تک بعدی مورد بررسی قرار گرفته است که مهمترین آن‌ها بررسی بخش‌های کوچکی از ژن‌ها است، اما امروزه می‌توان با استفاده از فناوری‌های جدید، عملکرد همزمان تمام ژنوم یک سلول را مورد بررسی قرار داد. فناوری ریزآرایه‌ها این امکان را به محققین داده است که بتوانند به طور هم زمان فعالیت هزاران ژن و بر هم کنش آن‌ها را مورد ارزیابی قرار دهند. مطالعات ریزآرایه‌ها حجم عظیمی از داده‌ها را تولید می‌کنند که روش‌های قدیمی تجزیه و تحلیل داده‌ها ممکن است توانایی کافی برای استخراج اطلاعات از این مجموعه داده‌های جدید را نداشته باشند. یکی از مهمترین مشخصات این نوع داده‌ها کم بودن تعداد نمونه (افراد) و بالا بودن تعداد متغیرهای اندازه‌گیری شده برای هر یک از نمونه‌ها است که مسئله نوینی به نام تجزیه و تحلیل داده‌ها با بعد بالا را به وجود آورده‌اند. روش‌های بسیار متعددی برای تحلیل این‌گونه داده‌ها مورد استفاده قرار می‌گیرند، اما یکی از مهمترین و پرکاربردترین روش‌های تحلیل داده‌ها که در زمره روش‌های اکتشافی است استفاده از تحلیل خوشه‌ای می‌باشد. هدفی که در این روش به دنبال آن هستیم یافتن گروه‌هایی از ژن‌ها است که دارای الگوی بیان مشابهی هستند، که این امر می‌تواند باعث کشف و شناسایی زیرگروه‌های جدیدی از بیماری‌ها و سرطان‌ها شود. این

نوع تحلیل‌ها برای اولین بار توسط گلوب (۱) و علیزاده (۲) مورد استفاده قرار گرفت و پس از آن به یکی از پرکاربردترین روش‌های تجزیه و تحلیل در زمینه ریزآرایه‌ها تبدیل شد (۳).

تعداد روش‌های خوشه‌بندی که در حال حاضر برای تجزیه و تحلیل داده‌ها، به خصوص داده‌های بیان ژنی مورد استفاده قرار می‌گیرند بسیار زیاد است و دقیقاً همین نکته، پاشنه آشیل استفاده از روش‌های خوشه‌بندی است، چرا که هیچ‌گاه نمی‌توان به صورت کلی در میان روش‌های خوشه‌بندی بهترین روش را شناسایی نمود، همچنین هیچ‌گونه معیاری وجود ندارد که به تنهایی بتواند بهترین روش خوشه‌بندی را معرفی نماید، علت این امر به دلیل عدم وجود یک تعریف دقیق و عملی از خوشه است. خوشه‌ها می‌توانند دارای هر نوع شکل و اندازه دلخواهی در یک فضای چند بعدی از الگوها باشند. هر روش خوشه‌بندی معیاری برای ساخت و ایجاد خوشه‌ها دارد، حال اگر هر کدام از داده‌ها با معیار مد نظر هم‌خوانی داشته باشد خوشه‌های حقیقی شناسایی و ساخته می‌شوند(۳).

برای حل این مشکل تاکنون معیارهای و شاخص‌های بسیاری جهت ارزیابی نتایج روش‌های خوشه‌بندی ارائه شده است. در این مقاله با استفاده از شاخص پهنای نیمرخ به بررسی عملکرد سه روش خوشه‌بندی سلسله مراتبی تجمعی، سلسله مراتبی تقسیم‌شونده و K- میانگین بر روی پنج مجموعه داده حقیقی زیرآرایه‌ها پرداخته شده است.

این مطالعه قصد دارد در ابتدا سه روش مرسوم خوشه‌بندی سلسله مراتبی تجمعی، سلسله مراتبی تقسیم‌شونده و K- میانگین، که به طور گسترده‌ای به خصوص در داده‌های بیان ژنی مورد استفاده قرار می‌گیرند را توضیح داده و هر یک از آن‌ها را در پنج مجموعه داده مورد استفاده قرار داده و در نهایت با استفاده از شاخص پهنای نیمرخ هر یک از این روش‌ها را مورد ارزیابی و مقایسه قرار دهد. همچنین با توجه به این‌که شاخص پهنای نیمرخ تنها یک آماره توصیفی بوده و انحراف معیار آن‌ها فرمول‌بندی نشده است، لذا از روش بازنمونه‌گیری بوت‌استرپ برای یافتن مقادیر انحراف معیار این شاخص استفاده شده است.

روش کار

مجموعه داده‌های مورد استفاده: در این مطالعه تعداد ۵ مجموعه داده بیان ژنی که همگی در اتباط با انواع سرطان‌ها در انسان می‌باشد مورد استفاده قرار گرفته‌اند. در ذیل به طور مختصر به خصوصیات هر یک از این مجموعه داده‌ها پرداخته شده است:

۱. در سال ۲۰۰۲ آرمسترانگ[1] مقاله‌ای در ژورنال Nat Genet در ارتباط با شناسایی زیر مجموعه‌های سرطان خون به چاپ رساند. در این مطالعه تعداد ۷۲ بیمار مورد بررسی قرار گرفتند که نتایج نشان‌دهنده سه زیرگروه برای این سرطان بود. گفتنی است تعداد ژن‌های مورد استفاده در تجزیه و تحلیل داده‌ها ۱۰۸۱ عدد بوده است (۹).

۲. بی‌هاتاچارچی[2] و همکارانش در سال ۲۰۰۱ مقاله‌ای در ارتباط با شناخت زیر کلاس‌های سرطان ریه با استفاده از تکنیک ریزآرایه‌ها انجام دادند. تعداد ۱۵۴۳ ژن در میان ۲۰۳ فرد مورد بررسی قرار گرفته شده است. داده‌های مورد بررسی آن‌ها دارای ۵ زیر گروه بوده است(۱۰).

۳. چادوری[3] و همکارانش در سال ۲۰۰۶ مقاله‌ای در ارتباط با شناخت ساختار ژن‌های مسئول سرطان سینه و روده بزرگ را بر روی ۶۲ بیمار با سرطان سینه و ۳۲ بیمار با سرطان روده بزرگ انجام دادند. در این مطالعه تعداد ۱۸۲ ژن مورد بررسی قرار گرفتند (۱۱).

۴. در سال ۲۰۰۳ نتایج مقاله درسکجویت[4] و همکارانش در مجله Nat Genet به چاپ رسید. در این مقاله با استفاده از تکنیک ریزآرایه‌ها به بررسی کشف زیر گروه‌های جدیدی از سرطان مثانه پرداخته شده است. محققین از ۳ گروه بیمار که سرطان مثانه داشتند اما به لحاظ کلینیکی در زیر گروه‌های جدا از یکدیگر قرار می‌گرفتند به تعداد ۴۰ نفر نمونه تهیه کردند. در مجموع تعداد ۱۲۰۳ ژن مورد بررسی قرار گرفته‌اند(۱۲).

1 -Armstrong
2 -Bhattacharjee
3- Chowdary
4 -Dyrskjot

۵.گلوب[1] و همکارانش در سال ۱۹۹۹ نتایج مطالعه‌ای را در نشریه ساینس به چاپ رساندند که در آن از تکنولوژی ریزآرایه به منظور کشف زیر کلاس‌های سرطان خون استفاده شده است. در این مجموعه داده تعداد ۱۸۶۸ ژن از ۷۲ بیمار مبتلا به این سرطان مورد بررسی قرار گرفته‌اند (۱).

خصوصیات هر یک از مجموعه داده‌های فوق به صورت خلاصه در جدول ۱ زیر آورده شده است:

هر یک از مجموعه داده‌های فوق با استفاده از سه روش خوشه‌بندی سلسله مراتبی تجمعی، سلسله مراتبی تقسیم‌شونده و K-میانگین خوشه‌بندی شده‌اند. تعداد خوشه‌ها برابر با تعداد گروه‌های بیماران که از آن‌ها نمونه گرفته شده است در نظر گرفته شده است.

از هر یک از ۵ مجموعه داده فوق تعداد ۱۰۰۰ نمونه با جایگذاری به صورت تصادفی گرفته شده و برای هر نمونه با استفاده از روش‌های خوشه‌بندی مورد بررسی خوشه‌ها ساخته شده است. سپس برای هر نمونه گرفته شده شاخص پهنای نیمرخ نیز محاسبه گردیده است. لذا برای هر مجموعه داده و برای هر روش خوشه‌بندی تعداد ۱۰۰۰ مقدار شاخص پهنای نیمرخ محاسبه شده است. حال با استفاده از این تعداد عدد، انحراف معیار بوت‌استرپ این شاخص در شرایط مختلف محاسبه شده است. تمامی محاسبات انجام شده با استفاده از نرم‌افزار R صورت گرفته است. از بسته نرم‌افزاری clValid برای ساخت خوشه‌ها و محاسبه شاخص پهنای نیمرخ استفاده شده است(۱۳).

همچنین به منظور انجام بازنمونه‌گیری به روش بوت‌استرپ و محاسبه میانگین و انحراف معیار شاخص پهنای نیمرخ برنامه‌ای در نرم‌افزار R نوشته شده است. در ادامه روشهای خوشه بندی در این مطالعه تبیین می شود.

شاخص پهنای نیمرخ: این شاخص یک شاخص ترکیبی است بدان معنا که هم نشان‌دهنده تراکم درون خوشه‌ای و هم نشان‌دهنده میزان تفکیک میان خوشه‌ها است. این شاخص به صورت زیر محاسبه می‌شود:

$$s(i)=\frac{b(i)-a(i)}{\max\{a(i),b(i)\}}$$

که در آن $a(i)$ میانگین فاصله میان مشاهده iام با دیگر مشاهدات در همان خوشه‌ای است که قرار دارد و $b(i)$ مینیمم میانگین فاصله میان مشاهده iام با تمامی مشاهدات در دیگر خوشه‌ها است. لذا شاخص پهنای نیمرخ همواره مقادیر بین ۱ تا ۱- را کسب می‌نماید. هر چه مقدار این شاخص بزرگتر باشد نشان‌دهنده عملکرد بهتر روش خوشه‌بندی و خوشه‌های ساخته شده در داده‌ها می‌باشد(۴). بازنمونه‌گیری به روش بوت‌استرپ: استنباط آماری مبتنی بر توزیع یک آماره یا توزیع نمونه‌گیری آن می‌باشد. مهمترین نقش روش بوت‌استرپ در یافتن توزیع نمونه‌ای یک آماره با استفاده از اطلاعات یک نمونه می باشد. روش کار بوت استرپ به صورت زیر است:

۱. باز نمونه‌گیری: توزیع نمونه‌گیری یک آماره بر اساس تعداد بسیار زیادی نمونه که به صورت تصادفی از یک

۲. جامعه گرفته شده‌اند ساخته می‌شود. در روش بوت‌استرپ به جای آن‌که تعداد زیادی نمونه از جامعه گرفته شود، تعداد بسیار زیادی نمونه با جایگذاری از یک نمونه واحد گرفته می‌شود به طوری که حجم هر نمونه برابر با نمونه اصلی گرفته شده از جامعه می‌باشد. نمونه با جایگذاری به این معنا است که پس از انتخاب هر مشاهده از نمونه اصلی، مشاهده دوباره به نمونه اصلی برگردانده

جدول ۱: خصوصیات هر یک از مجموعه داده‌های مورد استفاده

مجموعه داده	بافت	تعداد مشاهدات	تعداد کلاس‌ها	تعداد ژن‌های مورد بررسی
آرمسترانگ	خون	۷۲	۲	۱۰۸۱
بی‌هاتاچارچی	ریه	۲۰۳	۵	۱۵۴۳
چادوری	سینه، روده بزرگ	۱۰۴	۲	۱۸۲
درسکجویت	مثانه	۴۰	۳	۱۲۰۳
گلوب	مغز استخوان	۷۲	۲	۱۸۶۸

1 -Golub

شده تا شانس انتخاب مجدد را داشته باشد. اگر نمونه‌گیری انجام شده به صورت بدون جایگذاری باشد، آن‌گاه نمونه‌های گرفته شده همگی یکسان خواهند بود.

۳. توزیع بوت‌استرپ: توزیع نمونه‌ای یک آماره از گردآوری مقادیر آن آماره از تعداد زیادی نمونه ساخته می‌شود. این توزیع بوت‌استرپ آماره تمامی خصوصیات توزیع نمونه‌ای آماره را داراست. (۵)

در واقع روش بوت‌استرپ این امکان را در اختیار محقق قرار می‌دهد که بتواند واریانس و توزیع یک آماره مانند $T_n = g\left(X_1,...,X_n\right)$ را محاسبه نماید بدون آن‌که فرمول محاسباتی نظری آن را در اختیار داشته باشد. بدین منظور فرض کنید $V_F\left(T_n\right)$ نشان‌دهنده واریانس T_n باشد. در این‌جا کلمه F به منظور تاکید بر این موضوع آورده شده است که واریانس تابعی از F است. اگر این تابع F شناخته شده باشد می‌توان واریانس آماره را محاسبه نمود. به عنوان مثال اگر $T_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ باشد آن‌گاه خواهیم داشت:

$$V_F\left(T_n\right) = \frac{\sigma^2}{n} = \frac{\int x^2 dF\left(x\right) - \left(\int x dF\left(x\right)\right)^2}{n}$$

که تابعی از F است.

حال با استفاده از روش بوت‌استرپ ما به دنبال برآوردی برای $V_F\left(T_n\right)$ هستیم که آن را با $V_{\hat{F}_n}\left(T_n\right)$ نشان می‌دهیم. از آن‌جایی که محاسبه $V_{\hat{F}_n}\left(T_n\right)$ ممکن است دشوار و یا حتی امکان‌پذیر نباشد، لذا با استفاده از یک برآورد شبیه‌سازی شده آن را تقریب می‌زنیم و با v_{boot} آن را نمایش می‌دهیم. گام‌های محاسباتی این واریانس که به نام واریانس بوت‌استرپ مشهور است به صورت زیر می‌باشد:

۱. تعداد n مشاهده به صورت تصادفی ساده با جایگذاری از نمونه اصلی انتخاب می‌کنیم $\left(X_1^*,...,X_n^*\right)$

۲. آماره مد نظر را محاسبه می‌نماییم $T_n^* = g\left(X_1^*,...,X_n^*\right)$

۳. گام‌های ۱ و ۲ را B بار انجام می‌دهیم تا $T_{n,1}^*,...,T_{n,B}^*$ را داشته باشیم.

حال خواهیم داشت:

$$v_{boot} = \frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^* - \frac{1}{B}\sum_{r=1}^{B}T_{n,r}^*\right)^2$$

بنا به قانون اعداد بزرگ، در صورتی که B به بینهایت میل نماید، آن‌گاه v_{boot} به $V_{\hat{F}_n}\left(T_n\right)$ میل می‌کند (۶).

گرچه بحث نظری آماری، توانایی بسیار بالایی در شناساندن توزیع نمونه‌گیری و واریانس آماره‌ها دارد، اما توزیع نمونه‌گیری و واریانس بوت‌استرپ که با استفاده از تعداد بسیار زیادی باز نمونه‌ها ساخته شده است، خواص مشابهی با توزیع نمونه‌ای اصلی داده‌ها دارد. در واقع محاسبات زیاد و سنگین جایگزین بحث تئوری بسیار سنگین در ارتباط با مسائلی همچون قضیه حد مرکزی، میانگین و انحراف معیار $\overline{x}$ شده است. به همین دلیل بزرگترین مزیت روش بوت‌استرپ زمانی است که مباحث نظری توان پاسخ‌گویی به مشکلات و حل آنان را ندارند. در سال‌های اخیر به دلیل پیچیده شدن مسائل کاربردی و همچنین تولید داده‌هایی با بعد بالا، توان یافتن راه‌حل‌های کلاسیک آماری برای پی بردن به ساختار توزیع نمونه‌گیری آماره‌ها و همچنین واریانس آن‌ها بسیار کاهش یافته است. همچنین بسیاری از آماره‌ها وجود دارند که محققین هنوز راه‌حل هایی نظری برای یافتن واریانس یا انحراف معیار و همچنین توزیع نمونه‌گیری آن‌ها ارائه نکرده‌اند. لذا در این شرایط با توجه به خصوصیات ذکر شده برای روش نمونه‌گیری بوت‌استرپ، می‌توان از آن به عنوان جایگزینی مناسب برای یافتن توزیع نمونه‌ای آماره‌ها استفاده نمود.

شاخص‌های ارزیابی روش‌های خوشه‌بندی جزو آن دسته از آماره‌ها هستند که واریانس و توزیع نمونه‌گیری آنان به صورت نظری فرمول‌بندی نشده‌اند. بدین منظور برای یافتن انحراف معیار شاخص پهنای نیمرخ از روش نمونه‌گیری بوت‌استرپ استفاده شده است. هدف از محاسبه انحراف معیار این شاخص بررسی پایایی خوشه‌های ساخته شده توسط روش‌های خوشه‌بندی است. در صورتی که انحراف معیار این شاخص کمتر باشد نشان از آن دارد که خوشه‌های ساخته شده در هر مرحله از نمونه‌گیری در مراحل بعد تغییراتی کوچکتری داشته و لذا خوشه‌های ساخته شده در برابر تغییرات نمونه‌ای مقاوم‌تر می‌باشند (۷) و (۸).

شاخص RPT[1]: میانگین شاخص پهنای نیمرخ نشان از کیفیت خوشه‌های ساخته شده دارد و هر چه مقدار آن نزدیکتر به ۱ باشد گویای کیفیت بالاتر خوشه‌ها است. همچنین پایین بودن مقدار انحراف معیار بوت‌استرپ این شاخص نشان از آن دارد که خوشه‌های ساخته شده وابستگی کمتری به تغییرات نمونه‌ای کوچک در داده‌ها دارند و این عدم وابستگی یکی از مهمترین فاکتورها در تحلیل داده‌هایی است که حجم نمونه در آن نسبت به تعداد متغیرها بسیار کمتر است، چرا که در صورتی که با افزوده شدن یا کاسته شدن یک یا چند نمونه ساختار خوشه‌های ساخته شده دست خوش تغییرات زیادی شود، آن‌گاه محقق نمی‌تواند به نتایج حاصله اعتماد لازم را داشته و نتایج از اعتبار پذیری کمی برخوردار خواهد بود.

برای آن‌که بتوان هم میزان کیفیت خوشه‌های ساخته شده و هم پایداری آن‌ها را به طور همزمان مورد بررسی قرارداد از شاخص ترکیبی RPT که توسط سیز ارائه شده است استفاده شده است. این شاخص به منظور ارزیابی عملکرد و پایایی روش‌های انتخاب خصیصه در داده‌های بیان ژنی مورد استفاده قرار گرفته است. شاخص فوق به صورت زیرتعریف شده است:

$$RPT_{\beta} = \frac{\left(\beta^2+1\right)\text{robustness}\ \ \text{performance}}{\beta^2\ \text{robustness} + \text{performance}}$$

که در آن robustness شاخص پایداری و performance شاخص عملکرد روش‌های مربوطه و β وزن اهمیت پایداری در مقابل عملکرد است و زمانی که مقدار آن برابر با یک باشد وزن یکسانی به هر دو پارامتر پایایی و عملکرد می‌دهد(۷). با توجه به این‌که در مطالعه ما پایداری با استفاده از انحراف معیار شاخص پهنای نیمرخ و عملکرد با استفاده از میانگین شاخص پهنای نیمرخ اندازه‌گیری می‌شود و همچنین با توجه به این‌که جهت مطلوبیت این دو شاخص معکوس یکدیگر است (کمتر بودن انحراف معیار و بیشتر بودن میانگین)، لذا شاخص مورد استفاده در این مطالعه به صورت زیر می‌باشد:

$$RPT_{\beta} = \frac{\left(\beta^2+1\right)\ \text{performance}}{\text{robustness}\left(\frac{\beta^2}{\text{robustness}} + \text{performance}\right)}$$

که در نهایت به شکل زیر خواهد بود:

$$RPT_{\beta} = \frac{\left(\beta^2+1\right)\ \text{performance}}{\beta^2 + \text{robustness performance}}$$

که با جایگذاری‌های لازمه خواهیم داشت:

$$RPT_{1} = \frac{2\times\ \text{Mean (Si)}}{1 + \text{SD(Si) Mean(Si)}}$$

که در آن Si گویای شاخص پهنای نیمرخ است. هر چه مقدار این شاخص بیشتر باشد، کیفیت و پایایی خوشه‌های ساخته شده بیشتر خواهد بود.

یافته‌ها

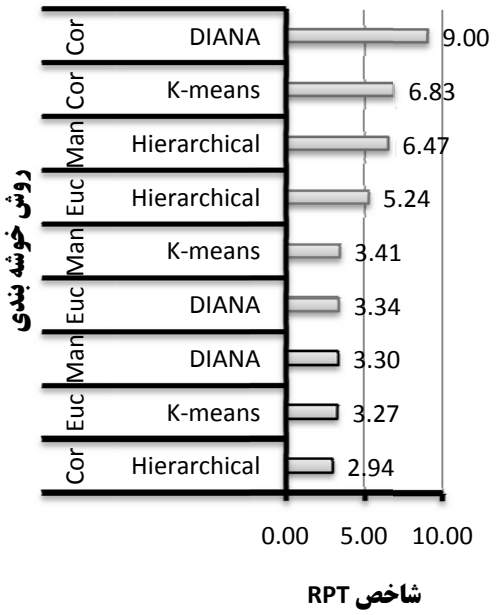
نتایج حاصل از به کارگیری روش بوت‌استرپ در ۹ ترکیب مختلف مد نظر در جدول ۲ زیر آورده شده است.

مجموعه داده آرمسترانگ: با استفاده از شاخص RPT مشخص می‌گردد ترکیب روش خوشه‌بندی DIANA و تابع اندازه ضریب همبستگی بهترین عملکرد را در میان دیگر ترکیبات نه تنها به لحاظ یافتن خوشه‌های متراکم، بلکه به لحاظ پایدار بودن خوشه‌ها دارا می‌باشد. همچنین با توجه به این‌که در دو ترکیب اول که بهترین نتایج را ارائه نموده‌اند ضریب همبستگی به عنوان تابع اندازه در نظر گرفته شده است، لذا می‌توان گفت در مجموعه داده آرمسترانگ تابع اندازه ضریب همبستگی توانایی شناخت

1- Robustness-Performance Trade-off

جدول ۲: نتایج حاصل از ارزیابی روش‌های خوشه‌بندی با استفاده از شاخص پهنای نیمرخ

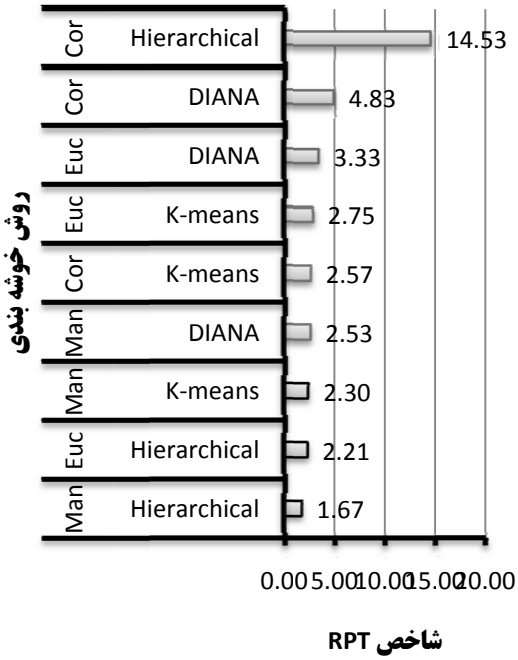
مجموعه داده		آرمسترانگ		باهاتاچارپی		چادوری		درسکجویت	
تابع اندازه	روش خوشه‌بندی	میانگین	انحراف معیار	میانگین	انحراف معیار	میانگین	انحراف معیار	میانگین	انحراف معیار
Euclidean	Hierarchical	۰/۲۷۵۲	۰/۰۲۸۷	۰/۱۷۶۶	۰/۰۲۸۱	۰/۸۴۰۷	۰/۰۴۳۴	۰/۴۰۸۵	۰/۰۹۶۴
	K-means	۰/۲۲۶۲	۰/۰۳۱۱	۰/۱۸۴۰	۰/۰۲۴۵	۰/۸۳۲۹	۰/۰۴۵۳	۰/۳۹۹۸	۰/۰۹۶۸
	DIANA	۸۰/۲۴۵۳	۰/۰۳۵۷	۰/۱۶۸۱	۰/۰۱۶۹	۰/۸۴۱۳	۰/۰۴۷۵	۰/۴۰۷۱	۰/۰۹۸۲
Correlation	Hierarchical	۰/۳۰۸۸	۰/۰۶۳۶	۰/۴۹۴۴	۰/۰۳۳۱	۰/۴۳۲۱	۰/۰۲۶۳	۰/۳۳۲۴	۰/۰۴۱۸
	K-means	۰/۳۳۵۶	۰/۰۳۲۶	۰/۲۵۱۱	۰/۰۴۸۵	۰/۰۲۹۲	۰/۰۶۸۲	۰/۲۸۶۴	۰/۰۶۴۱
	DIANA	۰/۳۴۰۸	۰/۰۲۵۶	۰/۴۱۰۳	۰/۰۶۷۸	۰/۴۳۵۷	۰/۰۲۶۶	۰/۳۱۵۴	۰/۰۴۶۱
Manhattan	Hierarchical	۰/۲۹۶۱	۰/۰۲۶۹	۰/۱۳۴۹	۰/۰۲۱۷	۰/۸۵۵۱	۰/۰۳۲۷	۰/۳۴۹۴	۰/۰۹۱۴
	K-means	۰/۲۲۸۴	۰/۰۳۰۴	۰/۱۴۵۸	۰/۰۱۸۴	۰/۸۴۷۹	۰/۰۳۲۳	۰/۳۲۵۶	۰/۱۰۹۱
	DIANA	۰/۲۵۹۰	۰/۰۴۰۲	۰/۱۴۲۵	۰/۰۱۶۰	۰/۸۵۵۴	۰/۰۳۲۱	۰/۳۳۱۳	۰/۱۰۵۳



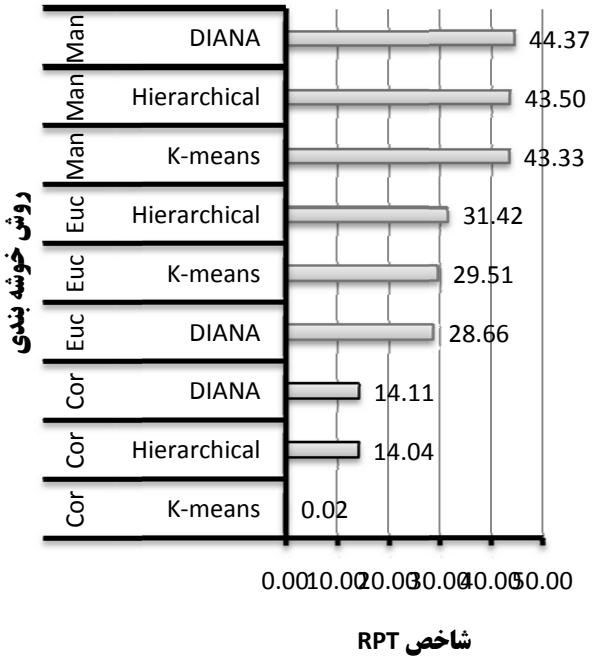
نمودار ۱: مقادیر شاخص RPT برای مجموعه داده آرمسترانگ

خوشه های مناسب را در دو روش خوشه‌بندی DIANA و K- میانگین بالا برده است.

مجموعه داده بی‌هاتاچارچی: به مانند مجموعه داده آرمسترنگ، دو ترکیب اول که بهترین نتایج را ارائه کرده‌اند از تابع اندازه ضریب همبستگی برای ساخت خوشه‌ها استفاده کرده‌اند. بهترین عملکرد را روش خوشه‌بندی سلسله مراتبی تجمعی و در رده بعد روش خوشه‌بندی DIANA داشته‌اند. ترکیب روش‌های خوشه‌بندی با تابع اندازه فاصله منهاتان نتایج به نسبه ضعیفتری را ارائه کرده است. با این‌که روش خوشه‌بندی سلسله مراتبی با تابع اندازه ضریب همبستگی بهترین عملکرد را در میان ترکیبات دیگر داشته است، اما همین روش با ترکیب تابع اندازه‌های فاصله اقلیدسی و منهاتان ضعیفترین نتایج را در بر داشته است.



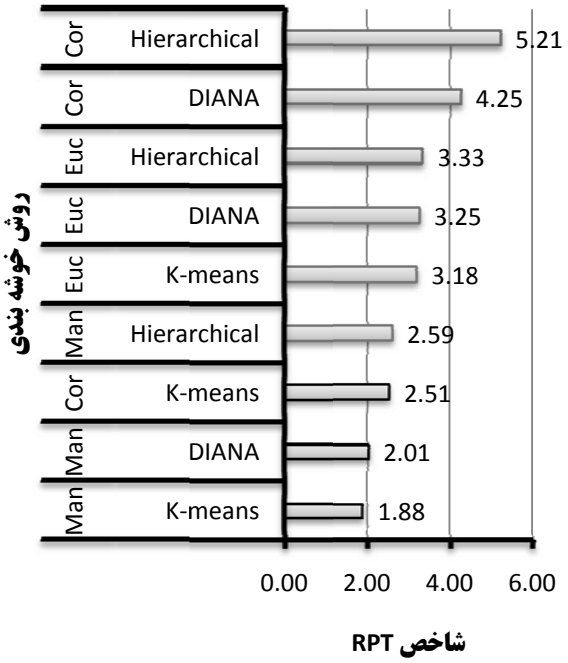
نمودار ۲: مقادیر شاخص RPT برای مجموعه داده بی‌هاتاچارچی



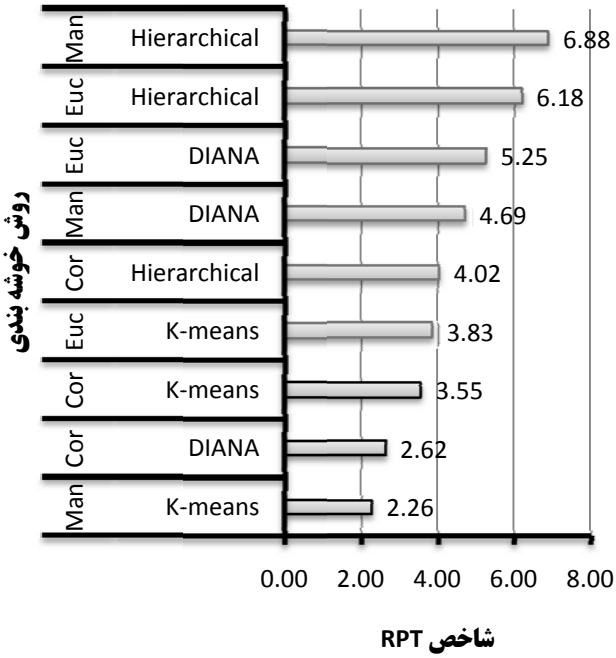
نمودار ۳: مقادیر شاخص RPT برای مجموعه داده چادوری

مجموعه داده چادوری: تابع اندازه فاصله منهاتان بهترین عملکرد را در ترکیب با سه روش خوشه‌بندی مورد بررسی در مجموعه داده چادوری داشته است. ترکیب این تابع اندازه فاصله با روش خوشه‌بندی DIANA بهترین ترکیب را به وجود آورده است. با اختلاف اندکی روش خوشه‌بندی سلسله مراتبی تجمعی در رده دوم جای دارد و روش K- میانگین در رده سوم جای گرفته است. ضعیفترین عملکرد مربوط به ترکیب روش‌های خوشه‌بندی با تابع اندازه ضریب همبستگی است، به طوری که ترکیب روش خوشه‌بندی K- میانگین با این تابع اندازه عدد بسیار کوچکی که نشان از مطلوب نبودن خوشه‌های ساخته شده دارد، ارائه کرده است. در مجموع روش خوشه‌بندی DIANA با هر سه تابع اندازه عملکرد مناسبتری داشته است.

مجموعه داده درسکجویت: نمودار۴ گویای این مطلب است که ترکیب روش‌های خوشه‌بندی سلسله مراتبی با تابع اندازه همبستگی در داده‌های درسکجویت نتایج مطلوبتری را نسبت به دیگر ترکیبات در بر داشته است، به طوری که بهترین عملکرد مربوط به روش خوشه‌بندی سلسله مراتبی تجمعی است که از تابع اندازه ضریب همبستگی برای ساخت خوشه‌ها استفاده کرده است. پس از آن روش خوشه‌بندی DIANA با استفاده از تابع اندازه مذکور در رده دوم جای دارد. در این مجموعه داده ترکیب روش‌های خوشه‌بندی با تابع اندازه فاصله منهاتان نتایج به نسبه ضعیفتری را داشته است. روش خوشه‌بندی سلسله مراتبی تجمعی با استفاده از هر سه تابع اندازه فاصله درمیان دیگر روش‌ها از عملکرد بهتری برخوردار بوده است و این موضوع در مورد روش خوشه‌بندی K- میانگین معکوس است، بدین معنا که این روش در ترکیب با سه تابع اندازه مورد بررسی ضعیفترین نتایج را داشته است.



نمودار ۴: مقادیر شاخص RPT
برای مجموعه داده درسکجویت



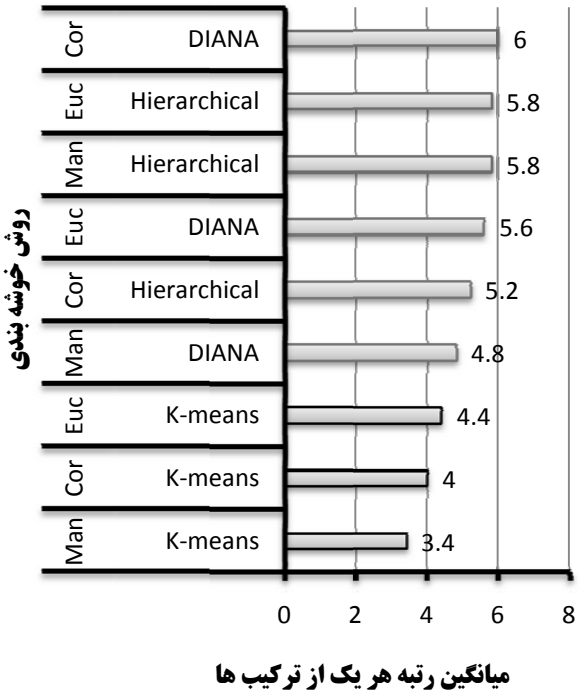
تمودار ۵: مقادیر شاخص RPT برای مجموعه داده گلوب

مجموعه داده گلوب: روش خوشه‌بندی سلسله مراتبی تجمعی در ترکیب با تابع اندازه فاصله منهاتان و اقلیدسی بهترین عملکرد را در ساخت خوشه‌ها در مجموعه داده گلوب داشته است. پس از این دو ترکیب، روش خوشه‌بندی DIANA با استفاده از تابع اندازه فاصله منهاتان و اقلیدسی عملکرد مناسبی در ساخت خوشه‌ها داشته است. ضعیفترین عملکرد مربوط به روش خوشه‌بندی K- میانگین است در ترکیب با تابع اندازه منهاتان ایجاد شده است. همچنین روش خوشه‌بندی K- میانگین در ترکیب با دو تابع اندازه دیگر نیز نتایج به نسبه ضعیفتری را ارائه نموده است. در میان تابع اندازه‌های مورد استفاده، تابع اندازه ضریب همبستگی نتایج ضعیفتری را در ساخت خوشه‌ها داشته است. بر خلاف آن‌چه مترک فاصله اقلیدسی و منهاتان ارائه نموده‌اند.

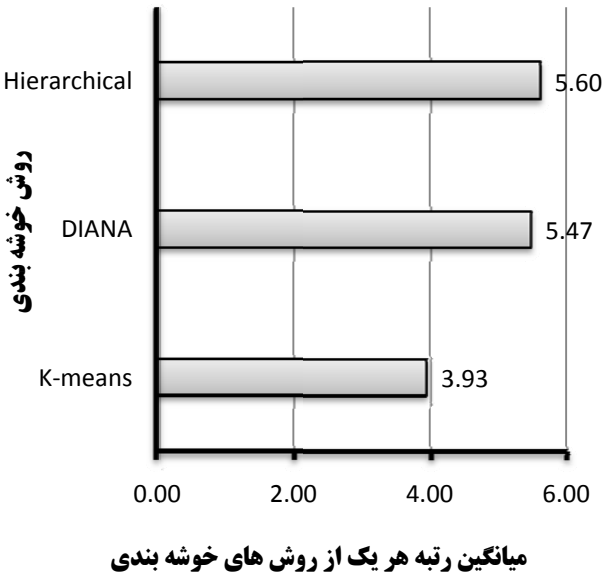
رتبه‌بندی روش‌های خوشه‌بندی: با میانگین‌گیری از رتبه ترکیب‌های روش خوشه‌بندی و تابع اندازه فاصله در تمامی ۵ مجموعه داده مشخص گردید که ترکیب روش خوشه‌بندی DIANA با تابع اندازه ضریب همبستگی بهترین عملکرد را در مجموع داشته است. پس از آن، ترکیب روش خوشه‌بندی سلسله مراتبی تجمعی با تابع اندازه فاصله اقلیدسی و همچنین با تابع اندازه فاصله منهاتان در رده دوم جای گرفته‌اند. یکی از نکات قابل توجه در ارتباط با روش خوشه‌بندی K- میانگین است، چرا که سه ترکیب انتهایی این رده‌بندی هر سه متعلق به ترکیب تابع اندازه فاصله‌های مختلف با این روش خوشه‌بندی می‌باشند که نشان‌دهنده نتایج ضعیفی است که از ساخت خوشه‌ها توسط این روش در کل ۵ مجموعه داده به دست آمده است. همچنین ضعیفترین رتبه مربوط به ترکیب روش خوشه‌بندی K- میانگین با تابع اندازه فاصله منهاتان بوده است. همچنین در نمودار ۵ با میانگین گیری از رتبه هر یک از روش‌های خوشه‌بندی

جدول ۳: رتبه هر یک از ترکیب‌های میان روش‌های خوشه‌بندی و تابع اندازه فاصله
(بهترین رتبه =۹، بدترین رتبه=۱)

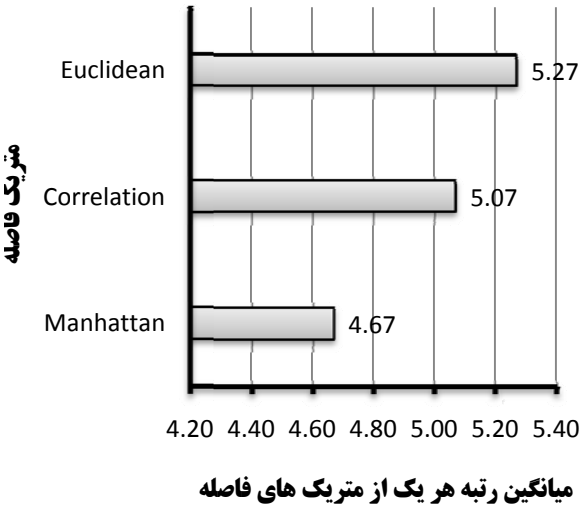
تابع اندازه فاصله	روش خوشه‌بندی	آرمسترانگ	باهاتاچارپی	چادوری	درسکجویت	گلوب	میانگین
Manhattan	K-means	۵	۳	۷	۱	۱	۳/۴
Correlation	K-means	۸	۵	۱	۳	۳	۴
Euclidean	K-means	۲	۶	۵	۵	۴	۴/۴
Manhattan	DIANA	۳	۴	۹	۲	۶	۴/۸
Correlation	Hierarchical	۱	۹	۲	۹	۵	۵/۲
Euclidean	DIANA	۴	۷	۴	۶	۷	۵/۶
Manhattan	Hierarchical	۷	۱	۸	۴	۹	۵/۸
Euclidean	Hierarchical	۶	۲	۶	۷	۸	۵/۸
Correlation	DIANA	۹	۸	۳	۸	۲	۶



نمودار ۶: میانگین رتبه ترکیب روش‌های خوشه‌بندی و تابع اندازه‌های فاصله در کل مجموعه داده‌ها



نمودار ۷: میانگین رتبه هر یک از روش‌های خوشه‌بندی در کل مجموعه داده‌ها



نمودار ۸: میانگین رتبه هر یک از متریک‌های فاصله در کل مجموعه داده‌ها

مشخص می‌گردد که صرف نظر از نوع تابع اندازه فاصله مورد استفاده، روش خوشه‌بندی سلسله مراتبی تجمعی عملکرد مناسبتری نسبت به دو روش دیگر داشته است. همچنین به طور کاملاً واضح مشخص می‌گردد که روش‌های خوشه‌بندی سلسله مراتبی عملکرد بهتری نسبت به روش محبوب و پرکاربرد K- میانگین در این مجموعه داده‌ها داشته‌اند.

نمودار ۷ نشان‌دهنده عملکرد کلی تابع اندازه‌های مورد استفاده صرف نظر از نوع روش خوشه‌بندی است. همان‌طور که نمودار نشان می‌دهد، میانگین رتبه روش‌هایی که از تابع اندازه فاصله اقلیدسی استفاده کرده‌ند بالاتر از دو تابع اندازه ضریب همبستگی و فاصله منهاتان بوده است.

بحث

مطالعات پزشکی در سال‌های اخیر و به خصوص پس از انتشار اطلاعات مربوط به پروژه ردیف‌یابی ژنوم انسان حیطه جدیدی را تجربه نموده است که راه‌کارهای بسیار امید بخشی را در جهت شناسایی و درمان بیماری‌های ناعلاجی همچون سرطان‌ها به روی محققین گشوده است. هم اکنون در جوامع تحقیقاتی پزشکی بخش مهمی در کنار پزشکی بالینی به نام پزشکی ژنتیکی مولکولی در حال ایجاد شدن است که اطلاعات تکمیلی بسیار مفیدی را در اختیار محققین قرار می‌دهد. اما بر خلاف داده‌های گردآوری شده در مطالعات پیشین پزشکی که محققین با تعداد متغیرهای کمی (در حدود ۱۰ الی ۱۵ متغیر) سر و کار داشته‌اند، حیطه جدید پزشکی سلولی مولکولی با حجم بسیار عظیمی از اطلاعات که به وسیله روش‌ها و تکنینک‌های نوین آزمایشگاهی تولید می‌شوند رو به رو است. بسیاری از روش‌های کلاسیک محاسباتی، اعم از آماری، ریاضی و کامپیوتری توسط این مجموعه داده‌های جدید به چالش کشیده شده‌اند. گرچه با توجه به سرعت بسیار بالاتر توسعه روش‌های استخراج داده‌ها نسبت به توسعه روش‌های محاسباتی، نقاط ضعف بسیار بزرگی در استفاده از روش‌های کلاسیک محاسباتی وجود دارد، اما بسیاری از محققین در کنار توسعه روش‌های موجود به دنبال بررسی کارایی روش‌های کلاسیک در برخورد با این حجم اطلاعات می‌باشند.

یکی از کارآمدترین و رایجترین روش‌های استخراج اطلاعات ژنتیکی استفاده از روش ریزآرایه‌ها است که توانایی بسیار زیادی را به محققین در جهت استخراج حجم بسیار عظیمی از اطلاعات را در آن واحد و بر حسب شرایط دلخواه ایجاد کرده است. هم اکنون نحوه مدیریت و تجزیه و تحلیل این نوع از داده‌ها یکی از بحث برانگیزترین مباحث علمی موجود در زمینه روش‌های محاسباتی است.

با توجه به هدف محققین از به کارگیری تکنیک ریزآرایه‌ها در استخراج داده‌ها، که همانا شناخت عملکرد گروهی ژن‌ها در فعالیت‌های سلول است، روش‌های محاسباتی مرتبط با این حیطه که با نام روش‌های خوشه‌بندی مشهوراند نیز از اهمیت دوچندانی برخوردار شده‌اند.

با توجه به وجود روش‌های بسیار متنوع خوشه‌بندی داده‌ها، اعم از آماری، ریاضی و محاسباتی، یکی از مهمترین سوالاتی که همواره در استفاده از این روش‌ها مطرح بوده است، شناسایی روش یا روش‌هایی بوده است که از عملکرد مناسبی در یافتن خوشه‌ها برخوردار باشند.

باز هم به دلیل شکل بسیار پیچیده داده‌های تولید شده از روش‌های ریزآرایه، شناخت روش‌های خوشه‌بندی مناسب به آسانی امکان‌پذیر نیست.

در این مطالعه با استفاده از ۵ مجموعه داده ژنتیکی که از تکنیک ریزآرایه‌ها در آن‌ها استفاده شده است، به ارزیابی و مقایسه سه روش خوشه‌بندی مشهور و پر کاربرد سلسله مراتبی تجمعی، سلسله مراتبی تقسیم‌شونده و K- میانگین که با سه تابع اندازه فاصله ضریب همبستگی، فاصله اقلیدسی و منهاتان ترکیب شده بودند، پرداخته شده است.

در این مطالعه هر یک از ترکیب‌های ایجاد شده میان روش‌های خوشه‌بندی و تابع اندازه فاصله در هر ۵ مجموعه داده مورد استفاده قرار گرفتند و نتایج آن‌ها با استفاده از شاخص پهنای نیمرخ مورد ارزیابی قرار گرفته است. با توجه به این‌که این شاخص دارای انحراف معیار نمی‌باشد، به منظور ارزیابی پایایی خوشه‌های ساخته شده توسط هر یک از ترکیبات فوق، از روش بوت استرپ برای محاسبه انحراف معیار این شاخص استفاده شده است.

نتایج نشان می‌دهد که هیچ یک از روش‌های خوشه‌بندی و یا ترکیبات آن‌ها برتری مطلق در تمامی ۵ مجموعه داده ندارند. این امر می‌تواند ناشی از متغیرهایی باشند که با تغییر مجموعه داده‌ها، آن‌ها نیز تغییر می‌کنند. از این قسمت می‌توان این نتیجه‌گیری را نمود که ممکن است متغیر یا متغیرهایی که وابسته به خواص مجموعه داده‌ها می‌باشند بر روی عملکرد روش‌های خوشه‌بندی تاثیرگذار باشند، متغیرهایی همچون حجم نمونه، تعداد ژن‌های مورد بررسی، بافت مورد بررسی و ... .

یکی از بارزترین نتایج حاصله در این مطالعه، روند تغییرات شاخص پهنای نیمرخ در زمانی است که روش‌ها از تابع اندازه‌های مختلف استفاده می‌نمایند. با بررسی دقیق نمودارهای موجود مشخص می‌گردد که مقادیر شاخص پهنای نیمرخ در روش‌های خوشه‌بندی که از دو تابع اندازه فاصله منهاتان و اقلیدسی استفاده می‌کنند بسیار به یکدیگر شبیه‌اند. به عنوان مثال در هر ۵ مجموعه داده‌ها مورد استفاده اختلاف میان شاخص پهنای نیمرخ در حالتی که از تابع اندازه فاصله اقلیدسی استفاده شده است در مقایسه با حالتی که از تابع اندازه منهاتان استفاده شده است بسیار اندک است، در حالی که اختلاف شاخص پهنای نیمرخ در دو تابع اندازه ذکر شده با تابع اندازه ضریب همبستگی به مراتب بیشتر است. همچنین با بررسی دقیق‌تر می‌توان متوجه این موضوع گردید که یک روند معکوس میان مقادیر شاخص پهنای نیمرخ در زمانی که از دو تابع اندازه فاصله منهاتان و اقلیدسی استفاده شده است در مقایسه با تابع اندازه ضریب همبستگی وجود دارد، بدان معنا که اگر در یک مجموعه داده مقدار شاخص پهنای نیمرخ در تابع اندازه‌های فاصله منهاتان و اقلیدسی کم باشد، در مورد تابع اندازه ضریب همبستگی بیشتر است و بالعکس.

در هیچ یک از ۵ مجموعه داده مورد بررسی، روشی که بالاترین میانگین را داشته است، بالاترین مقدار انحراف معیار را نداشته است، این بدان معنا است که می‌توان با بالا بودن مقدار شاخص پهنای نیمرخ تا حدودی به پایا بودن خوشه‌های ساخته شده نیز اطمینان داشت.

در نهایت شاخص RPT نشان داد که بهترین ترکیب برای ساخت خوشه‌ها در مجموعه داده آرمسترانگ، روش خوشه‌بندی DIANA با تابع اندازه ضریب همبستگی، در مجموعه داده بی‌هاتاچارچی ترکیب روش خوشه‌بندی سلسله مراتبی تجمعی با تابع اندازه ضریب همبستگی، در مجموعه داده چادوری ترکیب روش خوشه‌بندی DIANA با تابع اندازه فاصله منهاتان، در مجموعه داده درسکجویت ترکیب روش خوشه‌بندی سلسله مراتبی تجمعی با تابع اندازه ضریب همبستگی و در مجموعه داده گلوب ترکیب روش سلسله مراتبی تجمعی با تابع اندازه منهاتان بوده‌اند. با استفاده از رتبه‌گذاری برای ترکیب‌های مختلف مشخص گردید که در مجموع ۵ مجموعه داده بهترین ترکیب شامل روش خوشه‌بندی DIANA با استفاده از تابع اندازه ضریب همبستگی بوده است، پس از آن روش خوشه‌بندی سلسله مراتبی تجمعی با تابع اندازه‌های فاصله اقلیدسی و منهاتان در رده‌های بعدی قرار دارند. روش خوشه‌بندی K- میانگین نیز ضعیف‌ترین نتایج را در مجموع ۵ مجموعه داده داشته است. همچنین با صرف نظر از نوع تابع اندازه انتخاب شده، روش خوشه‌بندی سلسله مراتبی تجمعی بهترین عملکرد را در میان سه روش داشته است. پس از آن روش DIANA و در نهایت روش خوشه‌بندی K- میانگین می‌باشد. این موضوع در ارتباط با تابع اندازه‌های مورد استفاده به این صورت است که تابع اندازه فاصله اقلیدسی بهترین عملکرد را در مجموع داشته است. پس از آن تابع اندازه ضریب همبستگی و فاصله منهاتان جای دارند.

نتیجه گیری

این مطالعه نشان‌دهنده این موضوع است که در حالت کلی بهترین روش خوشه‌بندی به صورت مطلق وجود ندارد و بر اساس شرایط موجود در داده‌ها ممکن است انواع روش‌های خوشه‌بندی برتری نسبی بر دیگر روش‌ها داشته باشند.به عنوان پیشنهاداتی برای مطالعات آینده می‌توان موارد زیر را در نظر داشت:

- استفاده از انحراف معیار محاسبه شده به روش بوت‌استرپ به منظور ساخت فواصل اطمینان و مقایسه
- روش‌های خوشه‌های بندی بر اساس این فواصل اطمینان
- انجام آزمون‌های آماری با استفاده از شناخت توزیع نمونه‌گیری شاخص‌های ارزیابی نتایج خوشه‌بندی
- بررسی تاثیر عوامل مختلف همچون خواص مختلف مجموعه داده‌ها بر روی توزیع نمونه‌گیری شاخص‌های ارزیابی روش‌های خوشه‌بندی
- استفاده از دامنه گسترده‌تری از مجموعه داده‌های بیان ژنی با خصوصیات مختلف به منظور اختصاصی سازی روش‌های خوشه‌بندی در تحلیل این‌گونه داده‌ها
- به کارگیری بعد وسیع‌تری از روش‌های خوشه‌بندی اعم از آماری و غیر آماری به منظور یافتن روش‌های مناسبتر

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999, 286(5439):531-537.
2. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000, 403:503-511.
3. Souto CP, Costa IG, Arauja SA,. Clustering Cancer Gene Expression Data: a comparison study. BMC Bioinformatics 2008, 9:494
4. Rousseeuw J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 1987, 20:53–65
5. Tibshirani R, and Efron B. An Introduction to the Bootstrap (Chapman & Hall/CRC Monographs on Statistics & Applied Probability), 1993
6. Wasserman L. All of nonparametric statistics. NY: Springer; 2006
7. Saeys Y, Abeel T., and Peer Y.V. Robust feature selection using ensemble feature selection techniques. In Proceedings of the ECML Conference, 2008; 313-325
8. Loscalzo S, Yu L, and Chris Ding. "Consensus Group Based Stable Feature Selection". In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-09), 2009; 567-576, Paris, France, June,
9. Armstrong SA, Staunton JE, Silverman LB, Pieters R, den Boer ML, Minden MD, Sallan SE, Lander ES, Golub TR, Korsmeyer SJ: MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. Nat Genet 2002, 30:41-47.
10. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ, Meyerson M: Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci USA 2001, 98(24):13790-13795.
11. Chowdary D, Lathrop J, Skelton J, Curtin K, Briggs T, Zhang Y, Yu J, Wang Y, Mazumder A: Prognostic gene expression signatures can be measured in tissues collected in RNAlater preservative. J Mol Diagn 2006, 8:31-39.
12. Dyrskjot L, Thykjaer T, Kruh0ffer M, Jensen JL, Marcussen N, Hamilton- Dutoit S, Wolf H, Orntoft TF: Identifying distinct classes of bladder carcinoma using microarrays. Nat Genet 2003, 33:90-96.
13. Brock G, Pihur V, Datta Su, Datta So. clValid , an R package for cluster validation. July 27, 2008. Available at: http://pandawa.ipb.ac.id/cran/web/packages/clValid/vignettes/clValid.pdf

CCK (Clustering-Classification-Kappa); a new validation index to assessing clustering results of gene expression data

Shakeri MT[1], Sabaghian E[2], Esmaeili H[1]*

[1] Associate Professor of Biostatistics, Biostatistics Dept., Mashhad University of medical Science, Mashhad, Iran
[2] M.Sc. of Biostatistics, Mashhad university of Medical Sciences, Mashhad, IRAN

***Corresponding Author:** School of Health, Mashhad university of Medical Sciences (MUMS), Mashhad, Iran
Email: EsmailyH@ mums.ac.ir*

Abstract

Background& Objective: *The use of clustering methods for the discovery of cancer subtypes has drawn a great deal of attention in the scientific community. While bioinformaticians have proposed new clustering methods that take advantage of characteristics of the gene expression data, the medical community has a preference for using "classic" clustering methods. There have been no studies thus far performing a large-scale evaluation of different clustering methods in this context.*

Method & Material: *We present CCK index for assessing clustering result of gene expression data. This index was made by combining two arbitrary classification and clustering algorithms result and finally.*

the first large-scale analysis of nine different clustering methods, Hierarchical clustering with Single, Average, Complete and Ward linkages, UPGMA, Diana, K-means, PAM and CLARA methods for the analysis of 5 cancer gene expression data sets. Afterward we use Margin Trees method for assessing quality of result of clustering methods. Ultimately we calculate quality of result of clustering methods via Kappa coefficient between result of clustering methods and result of Margin Tree method for each clustering methods.

Results*: Our results reveal that the PAM, followed closely by CLARA, exhibited the best performance in terms of recovering the true structure of the data sets. Also we found that Partitioning clustering methods (PAM, CLARA and K-means) have better performance than Hierarchical clustering methods (Hierarchical clustering with Single, Average, Complete and Ward linkages, UPGMA and Diana).*

Conclusion*: The validation technique was used in this paper (Margin Trees) can aid in the selection of an optimal algorithm, for a given data set, from a collection of available clustering algorithms.*

Keyword: Clustering, Microarray,Bootstrap, Indicator to assess the of clustering methods