مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره اول. ص ۷۲-۶۰، ۱۳۹۹

ارائه یک روش خوشه‌بندی گراف–محور جهت شناسایی جمعیت‌های سلولی در داده‌های توالی‌یابی RNA سلول–منفرد

امین عینی پور[۱]، محمد مصلح[۲]*، کریم انصاری اصل[۳]

• دریافت مقاله: ۱۳۹۸/۴/۲۵ • پذیرش مقاله: ۱۳۹۸/۹/۲۳

مقدمه: استفاده از فناوری «توالی‌یابی RNA سلول–منفرد» باعث شناخت بهتر ساختارهای سلولی شده و داده‌های با وضوح بسیار بالایی از بیان ژن‌های مختلف هر سلول را در یک زمان واحد ارائه می‌دهد. یکی از زمینه‌های پرکاربرد در این حوزه، خوشه‌بندی داده‌ها بر اساس ژن‌های بیان شده است که بعضاً منتج به شناسایی جمعیت‌های سلولی جدید می‌گردد. عملکرد روش‌های پیشنهادی عمدتاً به شکل جمعیت‌ها و ابعاد داده‌ها بستگی دارد؛ لذا توسعه یک روش که بتواند فارغ از این موانع به شناسایی جمعیت‌های سلولی بپردازد، بسیار مهم است.

روش: در روش پیشنهادی که یک روش کتابخانه‌ای بود، ابتدا تعداد جمعیت‌های سلولی تخمین زده شد. این تخمین از آن جهت اهمیت دارد که در دنیای واقعی، اطلاعات اولیه مثل تعداد و نوع جمعیت‌های سلولی در دسترس نیست. سپس با استفاده از یک کرنل گاوسی مبتنی بر گراف، ضمن کاهش ابعاد مسئله، اقدام به شناسایی جمعیت‌های سلولی با روش خوشه‌بندی kmeans++ شد.

نتایج: نتایج پیاده‌سازی نشان داد که روش پیشنهادی می‌تواند نسبت به سایر روش‌های یادگیری ماشین ارائه شده در این زمینه، بهبود قابل قبولی را حاصل کند. به عنوان مثال برای معیار ARI، مقادیر ۱۰۰، ۹۳/۴۷ و ۸۴/۶۹ به ترتیب برای مجموعه داده‌های سلول–منفرد Kolod، Buettner و Usoskin حاصل شد.

نتیجه‌گیری: روش پیشنهادی بدون هیچ اطلاعات اولیه در مورد تعداد و نوع جمعیت‌های سلولی و فارغ از ابعاد بالای مسئله، می‌تواند اقدام به خوشه‌بندی و در نتیجه شناسایی جمعیت‌های سلولی با دقت و کیفیت بالایی نماید.

کلید واژه‌ها: توالی‌یابی RNA سلول–منفرد، خوشه‌بندی، شناسایی جمعیت‌های سلولی، کرنل گاوسی مبتنی بر گراف

• **ارجاع:** عینی‌پور امین، مصلح محمد، انصاری اصل کریم. ارائه یک روش خوشه‌بندی گراف –محور جهت شناسایی جمعیت‌های سلولی در داده‌های توالی یابی RNA سلول–منفرد. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۹؛ ۷(۱): ۶۰-۷۲.

۱. گروه مهندسی کامپیوتر، واحد دزفول، دانشگاه آزاد اسلامی، دزفول، ایران
۲. گروه مهندسی کامپیوتر، واحد دزفول، دانشگاه آزاد اسلامی، دزفول، ایران
۳. گروه مهندسی برق، دانشکده مهندسی، دانشگاه شهید چمران اهواز، اهواز، ایران

* **نویسنده مسئول:** محمد مصلح
آدرس: دزفول، کوی آزادگان، بلوار دانشگاه، دانشگاه آزاد اسلامی واحد دزفول، دانشکده فنی، گروه مهندسی کامپیوتر، صندوق پستی ۳۱۳
• **شماره تماس:** ۰۶۱-۴۲۴۲۰۶۰۱ • **Email:** mosleh@iaud.ac.ir

مقدمه

میزان بیان هر ژن، مقدار رونوشت‌های (transcriptome) آن ژن را در نمونه‌های آزمایشگاهی نشان می‌دهد؛ بنابراین دانستن سطوح بیان ژن‌های یک نمونه، به توصیف حالات مولکولی آن نمونه و نحوه عملکرد سلول‌ها و بافت‌ها کمک می‌کند. داده‌های بیان ژن اطلاعات ارزشمندی در مورد شبکه‌های بیولوژیک، حالات سلولی و فهم عملکرد ژن‌ها ارائه می‌دهد. از آنجا که تمام سلول‌های بدن از یک سلول مشتق شده‌اند یکی از دلایل تفاوت‌ها و تمایزات بین سلول‌ها، حاصل بیان شدن یا نشدن قسمت‌هایی از ژنوم می‌باشد. بیان ژن همچنین می‌تواند به عنوان یکی از زیر لایه‌های تکامل در نظر گرفته شود؛ زیرا کنترل زمان‌بندی، مکان و مقدار ژن می‌تواند تأثیرات مهمی در عملکرد ژن‌ها درون سلول یا کل ارگانیسم داشته باشد [۱].

در سال‌های گذشته، روش‌های مختلفی جهت بررسی بیان ژن از قبیل لکه‌گذاری نورترن (northern blotting)، هیبریداسیون درجا (in situ hybridization)، رونوشت بردار معکوس (reverse transcription)، شناسایی توالی رونویسی شده (expressed sequence tag)، میکروآرایه (microarray)، آنالیز سریالی بیان ژن (serial analysis of gene expression) و توالی‌یابی (RNA sequencing) RNA معرفی شده است [۲].

روش‌های توالی‌یابی RNA، یکی از تکنیک‌های توالی‌یابی نسل جدید (Next Generation Sequencing) محسوب می‌شوند که خود به دو دسته توالی‌یابی توده‌ای (Bulk) و سلول-منفرد (Single-Cell) تقسیم می‌شوند [۱،۲]. در روش توده‌ای، داده‌های ژنومی تولید شده به‌طورمعمول از تجمیع کل جمعیت هزاران تا میلیون‌ها سلول درون بافت حاصل می‌شود که در بسیاری از موارد، این داده‌ها، ناهمگونی‌های ژنتیکی را مشخص نمی‌کنند. در این نوع داده‌ها، مطالعه تأثیر سلول‌ها به صورت انفرادی بر یکدیگر و همچنین شناسایی انواع سلولی کمیاب که غالباً نقش اصلی را در شروع بیماری‌های مهلکی مثل سرطان بازی می‌کند، مشکل است [۳].

روش توالی‌یابی RNA سلول-منفرد به‌عنوان یک فناوری جدید شناخته می‌شود که برای اولین بار در سال ۲۰۰۹ ارائه شد [۴]. این روش تا سال ۲۰۱۴ یعنی زمانی‌که پروتکل‌های جدید تدوین شد و هزینه‌های توالی‌یابی آن کاهش پیدا کرد، محبوبیت چندانی به دست نیاورد. این فناوری قادر است داده‌های سلولی بسیار دقیق و با ارزشی را از هر سلول به صورت جداگانه در اختیار ما قرار دهد. تجزیه‌وتحلیل داده‌ها در سطح سلول-منفرد یک بینش جدید در سیستم‌های بیولوژیکی پیچیده ایجاد کرده است که باعث ایجاد هویت سلولی در موجودات زنده می‌شود. این موضوع اجازه می‌دهد که به سؤالات علمی بی‌پاسخ در سال‌های گذشته، از قبیل شناسایی عوامل ناهمگونی در جمعیت‌های سلولی، مطالعه فرآیندهای پویا مانند تبدیل حالت سلولی و ساخت شبکه‌های تنظیم‌کننده ژنی (gene regulatory networks) پاسخ داده شود. با این حال، این روش جدید هنوز در مراحل اولیه خود به سر می‌برد به‌طوری که همراه با مزایای مذکور، چالش‌های محاسباتی جدیدی را نیز مطرح کرده است. این چالش‌ها شامل حجم بسیار بالای مجموعه داده‌های حاصل از این آزمایش، ابعاد بالای آن‌ها، ناهمگونی بیولوژیکی ناشی از احتمالی بودن ذاتی بیان ژن در سلول‌های منفرد، نویز فنی موجود در پردازش سلول، فساد سلولی (cell lysis)، آماده‌سازی کتابخانه بر اساس مقادیر بسیار کم ورودی پیام‌رسان‌های RNA و عدم وجود اطلاعاتی در مورد نوع و تعداد جمعیت‌های سلولی هست [۵،۶]. در سال‌های اخیر و همزمان با پیشرفت فناوری توالی RNA سلول-منفرد، تحقیقات زیادی در زمینه روش‌های محاسباتی جهت تجزیه‌وتحلیل داده‌های حاصل از این فناوری انجام شده که بتوانند بر این چالش‌ها غلبه کنند.

یکی از کاربردهای مهم در حوزه تحلیل داده‌های حاصل از توالی‌یابی سلول-منفرد، خوشه‌بندی داده‌ها است که منجر به شناسایی جمعیت‌های سلولی می‌شود. در خوشه‌بندی، نقطه‌های داده‌ای یا همان سلول‌ها به گروه‌هایی تقسیم می‌شوند که بازتاب دهنده زیر مجموعه‌ای از سلول‌های بسیار مشابه است که به آن‌ها جمعیت‌های سلولی گفته می‌شود [۷].

امروزه میلیون‌ها نفر در سراسر دنیا از بیماری‌های سلولی از جمله سرطان رنج می‌برند. درمان بیماری سرطان مستلزم شناخت گونه‌های متنوع سلولی درون این گونه بافت‌ها می‌باشد. ارائه یک روش خوشه‌بندی مناسب می‌تواند منجر به شناسایی جمعیت‌های سلولی نادر شود. این موضوع می‌تواند در شناسایی دقیق توده‌های سلولی خاص که منجر به بیماری‌هایی مثل سرطان می‌شوند کاربرد داشته باشد [۹-۷].

در طول سال‌های اخیر ترکیب روش‌های سنتی مرتبط با یادگیری ماشین در زمینه شناسایی جمعیت‌های سلولی داده‌های توالی‌یابی RNA سلول-منفرد توسط افراد مختلف و به شکل‌های مختلف استفاده شده است. الگوریتم خوشه‌بندی kmeans و سایر الگوریتم‌های خوشه‌بندی مبتنی بر فاصله

مثل خوشه‌بندی سلسله مراتبی به‌عنوان الگوریتم‌های سنتی یادگیری ماشین، به طور گسترده‌ای در این زمینه مورد استفاده قرار گرفته‌اند. در مطالعه‌ای، Jaitin و همکاران، خوشه‌بندی سلسله مراتبی و مدل‌های ترکیبی احتمالی را برای دسته‌بندی سلول‌های منفرد از بافت‌های مختلف ترکیب کردند [۱۰].

با توجه به پیچیدگی‌های موجود در داده‌های توالی‌یابی RNA سلول-منفرد مثل احتمالی بودن ذاتی بیان ژن، وجود نویز، حجم و ابعاد بسیار بالای این نوع از داده‌ها، استفاده از روش‌های یادگیری ماشین سنتی کارایی چندانی ندارد.

علاوه بر روش‌های سنتی، در سال‌های اخیر روش‌های جدید خوشه‌بندی مختص داده‌های سلول-منفرد نیز توسعه داده شده است. به‌عنوان مثال، در مطالعه Zurauskiene و Yau [۱۱] یک روش خوشه بندی اصلاح شده به نام pcaReduce برای داده‌های توالی‌یابی RNA سلول منفرد طراحی شده است که به طور تکراری روش کاهش ابعاد (Principal Component Analysis)PCA را با kmeans ترکیب می‌کند تا یک درخت سلسله مراتبی از سلول‌ها را تولید کند. بسته SINCERA نمونه دیگری است که از خوشه‌بندی سلسله مراتبی استفاده می‌کند که در آن از همبستگی پیرسون مرکزی برای معیار شباهت و میانگین پیوند برای روش پیوند به‌عنوان تنظیمات پیش فرض استفاده می‌کند [۱۲]. SNN-cliq نیز از مفهوم نزدیک‌ترین همسایه مشترک برای تعریف شباهت بین نقاط داده‌ای (سلول‌ها) استفاده می‌کند و از طریق یک الگوریتم مبتنی بر نظریه گراف، خوشه‌بندی را انجام می‌دهد [۱۳].

نکته قابل توجه در مورد اکثر این روش‌ها این است که اغلب آن‌ها شامل مدل‌سازی‌های آماری هستند که نیازمند تخمین یکسری از پارامترها می‌باشند. همچنین این الگوریتم‌ها غالباً از یکسری حلقه‌های تکرار برای رسیدن به راه‌حل‌های بهینه محلی و سراسری استفاده می‌کنند؛ لذا هنگامی‌که مجموعه‌های داده‌ای بزرگ که غالباً شامل چند صد سلول-منفرد می‌باشد را پردازش می‌کنند، بسیار کُند عمل می‌کنند. علاوه بر این در اکثر روش‌های معرفی شده به اطلاعات اولیه در مورد تعداد جمعیت‌های سلولی نیاز است که عملاً در دنیای واقعی در دسترس نیست.

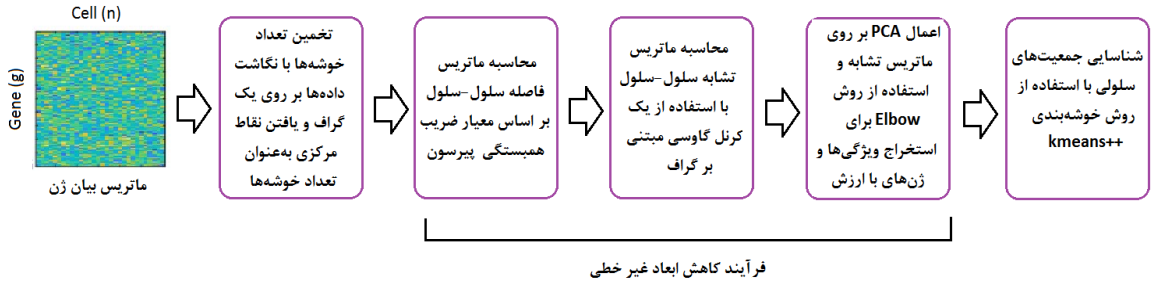
در این پژوهش یک روش جدید مبتنی بر گراف جهت خوشه‌بندی داده‌های توالی‌یابی RNA سلول-منفرد با هدف شناسایی هرچه دقیق‌تر جمعیت‌های سلولی ارائه شد. در روش پیشنهادی، ابتدا بر اساس یک تکنیک جدید مبتنی بر گراف، اقدام به تخمین تعداد خوشه‌ها در ابعاد بالا شد. تخمین تعداد خوشه‌ها یا همان جمعیت‌های سلولی از آن جهت حائز اهمیت است که در دنیای واقعی، اطلاعات اولیه مثل تعداد و نوع جمعیت‌های سلولی در دسترس نیست که این موضوع می‌تواند به یکی از چالش‌ها در این زمینه تبدیل شود. در مرحله بعد، با توجه به ابعاد بالا، متغیر و ماهیت غیرخطی این نوع از داده‌ها، با استفاده از یک رویکرد بدون نظارت و غیرخطی جدید و به کمک یک کرنل گاوسی مبتنی بر گراف، اقدام به استخراج ویژگی‌ها یا همان ژن‌های با اهمیت بالا شد. در مرحله آخر، از طریق یک روش خوشه‌بندی به‌نام kmeans++ که یک روش توسعه‌یافته از روش خوشه‌بندی معروف kmeans است، خوشه‌بندی و شناسایی جمعیت‌های سلولی بر اساس تعداد خوشه‌های تخمینی و ویژگی‌های استخراج شده در مراحل قبل انجام شد و کیفیت جمعیت‌های سلولی شناسایی شده بررسی شدند.

روش

روش پیشنهادی یک روش کتابخانه‌ای بود که اقدام به تجزیه‌وتحلیل مجموعه داده‌های توالی‌یابی RNA سلول-منفرد می‌کند. یکی از کاربردهای مهم فرآیند تجزیه‌وتحلیل داده‌های توالی‌یابی RNA سلول-منفرد، خوشه‌بندی یا شناسایی جمعیت‌های سلولی است که بعضاً منجر به شناسایی زیرمجموعه‌های سلولی نادر می‌شود؛ این امر به محققان کمک کند تا بینش جدیدی در سطح سلولی به دست آورند و می‌توانند در بیماری‌های صعب‌العلاج مثل سرطان و پاسخ‌های درمانی آن کاربرد داشته باشد. ورودی این فرآیند، ماتریسی به‌نام ماتریس بیان ژن به‌صورت $X_{g\times n}$ است که دارای g سطر و n ستون می‌باشد. در ماتریس مذکور، g و n به ترتیب نشان‌دهنده تعداد ژن‌ها و تعداد سلول‌ها هستند که تعداد ژن‌ها بالغ بر ده‌ها-هزار و تعداد سلول‌ها از صدها تا میلیون‌ها سلول در برخی مجموعه داده‌ها متغیر است.

یکی از چالش‌های اصلی در شناسایی جمعیت‌های سلولی این است که در دنیای واقعی، تعداد جمعیت‌های سلولی یا همان تعداد خوشه‌ها نامشخص است. همچنین، می‌توان به مسائل دیگری مثل وجود نویز و ابعاد بالای داده‌های توالی‌یابی RNA سلول-منفرد اشاره کرد که باعث پیچیدگی بیشتر تحلیل این نوع داده‌ها می‌شود؛ لذا در روش پیشنهادی، سعی شد که با ارائه رویکردهای مناسب بر این چالش‌ها غلبه کرد تا بتوان در مرحله خوشه‌بندی با دقت بالایی اقدام به شناسایی

جمعیت‌هـای سـلولی نمـود. شِـمای کلـی روش پیشـنهادی را می‌توان در شکل ۱ مشاهده کرد.



شکل ۱: شِمای کلی روش پیشنهادی برای شناسایی جمعیت‌های سلولی در داده‌های توالی‌یابی RNA سلول-منفرد

مراحل روش پیشنهادی

تخمین تعداد خوشه‌ها

در این مطالعه از یک روش جدید مبتنی بر گراف جهت تخمین تعداد خوشه‌ها استفاده شد. در روش پیشنهادی، ابتدا با استفاده از ماتریس مجاورتی که بر اساس معیار ضریب همبستگی پیرسون محاسبه شد، یک گراف از روی مجموعه داده توالی‌یابی RNA سلول-منفرد موجود ایجاد شد. برای این کار ابتدا با استفاده از ضریب همبستگی پیرسون ماتریس فاصله سلول-سلول محاسبه شد، سپس با استفاده از یک کرنل گاوسی مبتنی بر گراف پیشنهادی (رابطه ۶) این ماتریس فاصله به یک ماتریس تشابه تبدیل شد. استفاده از این کرنل از آن جهت اهمیت دارد که با توجه به ماهیت پیچیده و غیرخطی حاکم بر داده‌های سلول-منفرد نمی‌توان با استفاده از معیارهای همبستگی رایج به سادگی روابط بین سلولی را کشف کرد. به این ترتیب با محاسبه ماتریس مجاورت، مجموعه‌داده مورد نظر بر روی یک گراف نگاشت شد که هر گره آن معادل یک سلول است.

در ادامه بر اساس گراف حاصل، مجموعه پیشنهادی $\rho = \{\rho_1, \rho_2, \dots, \rho_n\}$ به‌صورت زیر محاسبه شد:

$$\rho_i = \text{degree}(\text{node}_i) \; : 1 \le i \le n \qquad (۱)$$

که در آن ρ_i به‌عنوان چگالی هر گره محسوب می‌شود و معادل درجه هر گره در گراف مذکور در نظر گرفته شد که به راحتی قابل محاسبه است.

بعد از محاسبه مجموعه ρ یک مجموعه دیگر به‌نام $\delta = \{\delta_1, \delta_2, \dots, \delta_n\}$ نیز به‌صورت زیر محاسبه شد:

$$\delta_i = \min_{j:\rho_j > \rho_i} \{\|c_i - c_j\|\} \; : 1 \le i,j \le n \qquad (۲)$$

که در آن δ_i به‌عنوان کم‌ترین فاصله هر گره c_i با گرهی چگال‌تر از آن (c_j) در نظر گرفته شد. برای به دست آوردن این مجموعه، از الگوریتم معروف دیکسترا (Dijkstra) استفاده شد [۱۴].

بعد از محاسبه مجموعه‌های ρ و δ می‌توان با نگاشت این مقادیر بر روی یک نقشه دو بُعدی و تفسیر مقادیر موجود اطلاعات مفیدی مثل نقاط مرکزی (Centroid)، نقاط معمولی و نقاط پرت (Outlier) را به صورت زیر شناسایی کرد:

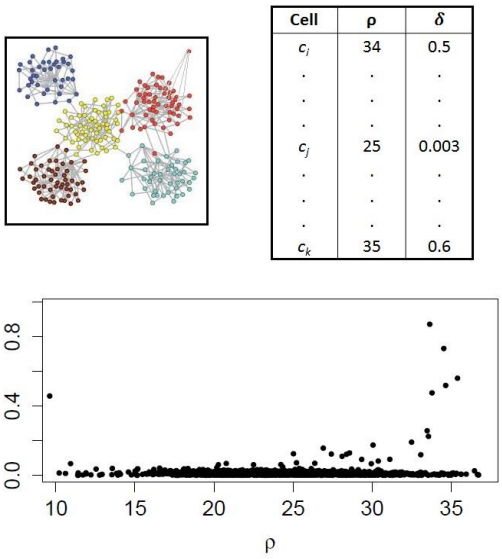
الف) نقاط مرکزی: نقاط و گره‌های دارای مقادیر ρ_i و δ_i بالا را می‌توان به‌عنوان داده‌های مرکزی در نظر گرفت؛ زیرا این گره‌ها دارای چگالی بالایی هستند و ضمناً فاصله آن از گره چگال‌تر از خود نیز زیاد است.

ب) نقاط معمولی: نقاط دارای مقادیر δ_i پائین را می‌توان به‌عنوان نقاط داده‌ای معمولی در نظر گرفت؛ زیرا این گره‌ها در فاصله کمی با هم و حول یک نقطه مرکزی در یک خوشه جمع شده‌اند.

ج) نقاط پرت: نقاط دارای مقادیر ρ_i پایین و δ_i بالا را نیز می‌توان به‌عنوان نقاط داده‌ای پرت در نظر گرفت، زیرا این نقاط دارای چگالی پایین و البته نسبت به سایر مراکز خوشه‌ای نیز از فاصله بالایی برخوردار هستند.

همچنین با تفسیر مقادیر ρ و δ علاوه بر اطلاعات مذکور می‌توان اطلاعات مفید دیگری نیز به دست آورد؛ مثل شناسایی و تخمین خوشه‌های بزرگ و کوچک موجود در داده‌ها به‌طوری‌که می‌توان نقطه‌ای با بیشترین مقدار ρ_i را به‌عنوان مرکز بزرگ‌ترین خوشه در نظر گرفت و بالعکس.

برای درک بهتر و ساده‌تر تفاسیر داده‌ای فوق می‌توان بعد از محاسبه مجموعه مقادیر ρ و δ، آن‌ها را بر روی یک محور مختصات و به صورت دو بُعدی نگاشت کرد که از طریق آن راحت‌تر بتوان اقدام به شناسایی نقاط مرکزی نمود. نمونه‌ای از این نگاشت به صورت شِماتیک در شکل ۲ نمایش داده شد.



شکل ۲: تخمین تعداد جمعیت‌های سلولی بر اساس گراف

برای این که بتوان از معیار دقیقی جهت تفکیک نقاط مرکزی از سایر نقاط (عادی و پرت) استفاده کرد از معیارهای فاصله (δ) و چگالی (ρ) که در [۱۵] ارائه شده به‌عنوان یک حد آستانه جهت خط بُرش نمودار حاصل استفاده شد، به این ترتیب که ابتدا نقاطی که دارای مقادیر δ بیشتر از حد مشخصی باشند به عنوان نقاط مرکزی کاندید از نقاط عادی جدا شد. نقاط باقی‌مانده شامل نقاط مرکزی و پرت هستند؛ زیرا نقاط پرت نیز مانند نقاط مرکزی دارای مقادیر δ بالایی هستند. از آنجایی‌که نقاط پرت معمولاً دارای مقدار چگالی پایینی هستند؛ لذا با اِعمال فیلتر مربوط به چگالی، نهایتاً، نقاطی که چگالی بیشتر از حد مشخصی دارند را می‌توان به‌عنوان نقاط مرکزی نهایی و تعداد خوشه‌ها در نظر گرفت. شرایط مربوط به تعریف این حد آستانه خط برش در روابط (۳) و (۴) نشان داده شده است [۱۵].

$$Ec_i = (\delta_i) \geq 2\sigma(\delta_i) \quad (۳)$$

$$c_i = Ec_i \geq \mu(\rho_i) \quad (۴)$$

در این روابط، Ec_i، نقاط مرکزی کاندید، σ و μ نیز به ترتیب پارامترهای مربوط به انحراف معیار و میانگین نقاط داده‌ای در نظر گرفته شدند که به‌راحتی قابل محاسبه هستند.

کاهش ابعاد

از جمله چالش‌های موجود در داده‌های توالی‌یابی RNA سلول–منفرد وجود نویزهای فنی و همچنین ابعاد بسیار بالای آن‌ها می‌باشد که باعث پیچیده شدن فرآیند تحلیل و در نتیجه کیفیت بسیار پایین خوشه‌بندی داده‌ها می‌شود؛ لذا در روش پیشنهادی یکی از بخش‌های اصلی کار تحت عنوان کاهش ابعاد جهت غلبه بر این مسائل اختصاص داده شد.

از آنجا که معمولاً در مجموعه داده‌های توالی‌یابی RNA سلول–منفرد نمی‌توان یک رابطه خطی بین نشانگرهای مهم سلولی تعریف کرد در روش پیشنهادی سعی شد با استفاده از تئوری گراف بر این مسئله نیز غلبه کرد. روش پیشنهادی بر اساس یک کرنل گاوسی مبتنی بر گراف و روش PCA، عمل می‌کند و بر اساس ضریب معیار همبستگی پیرسون و به کمک روش Elbow، یک مجموعه ویژگی نهایی استخراج می‌کند که در واقع یک مجموعه کاهش یافته از نشانگرها یا ژن‌های با اهمیت بالا هستند که می‌توانند به شناسایی جمعیت‌های سلولی در مرحله خوشه‌بندی کمک کنند.

فیلتر کردن ژن‌ها

گسترش نویزها در داده‌های توالی‌یابی RNA سلول–منفرد به صورت افزایش بیش‌ازحد مقادیر صفر و نزدیک صفر در مجموعه داده ظاهر می‌شود که مشکلاتی را در تحلیل این نوع داده‌ها ایجاد می‌کند؛ لذا معمولاً در اولین مرحله از فرآیند تجزیه‌وتحلیل این نوع از داده‌ها، یک فیلتر بر روی این داده‌ها اعمال می‌شود تا ویژگی‌ها و ژن‌هایی که به احتمال زیاد نویز هستند از مجموعه داده مورد نظر حذف شوند و ادامه عملیات بر روی ژن‌های با درجه اهمیت بالا انجام شود.

در اینجا از روش فیلتر کردن فرکانسی (Frequency filtering) استفاده شد که در آن تنها ژن‌هایی در نظر گرفته می‌شوند که در یک کسر مشخص از سلول‌ها بیان شده باشند. به طور مشخص در اینجا به صورت تجربی ژن‌هایی که کمتر از ۵٪ در تمام نمونه سلول‌ها بیان شده‌اند به عنوان نویز شناسایی

و حذف شد و باقی‌مانده ژن‌ها به عنوان ویژگی‌های با اهمیت برای استفاده در مراحل بعد حفظ شد [۱۶].

محاسبه ماتریس فاصله سلول-سلول (Cell-to-Cell Distance Matrix)

در این مرحله با استفاده از معیار ضریب همبستگی پیرسون، ماتریس فاصله سلول-سلول محاسبه شد. ضریب همبستگی پیرسون که به نام‌های ضریب همبستگی گشتاوری و یا ضریب همبستگی مرتبه صفر نیز نامیده می‌شود، به منظور تعیین میزان رابطه، نوع و جهت رابطه بین دو متغیر فاصله‌ای یا نسبی و یا یک متغیر فاصله‌ای و یک متغیر نسبی به کار برده شد که از طریق رابطه زیر محاسبه شد [۱۷]:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2} \qquad (۵)$$

که در آن x و y متغیرهای مورد نظر و $\bar{x}$ و $\bar{y}$ میانگین آن‌ها می‌باشد. هرچه مقدار قدر مطلق ضریب همبستگی به ۱ نزدیک‌تر باشد، نشان می‌دهد شدت رابطه بین دو متغیر قوی‌تر است. در مقابل، مقدار همبستگی نزدیک صفر نشان می‌دهد که رابطه بسیار ضعیفی بین متغیرهای x و y برقرار است. ضریب همبستگی پیرسون، روشی پارامتری است و معمولاً برای داده‌هایی با توزیع نرمال یا تعداد داده‌های زیاد استفاده می‌شود.

محاسبه ماتریس تشابه سلول-سلول (Cell-to-Cell Affinity Matrix)

با توجه به ماهیت داده‌های به دست آمده از آزمایش توالی‌یابی RNA سلول منفرد، نمی‌توان رابطه خطی بین نشانگرهای سلولی در این داده‌ها پیدا کرد. با توجه به ماهیت غیرخطی این نوع از داده‌ها، در روش پیشنهادی با انجام یک نگاشت مبتنی بر گراف، ماتریس فاصله به دست آمده در مرحله قبل به یک فضای غیرخطی منتقل شد. در این مرحله با استفاده از یک کرنل گاوسی مبتنی بر گراف، ماتریس تشابه سلول-سلول از روی ماتریس فاصله به دست آمده در مرحله قبل محاسبه شد. تابع کرنل گاوسی مبتنی بر گراف پیشنهادی بر اساس k-نزدیک‌ترین همسایه، ماتریس تشابه را با استفاده از رابطه (۶) به صورت زیر محاسبه می‌کند [۱۸]:

$$K(x_i, x_j) = \exp\left(-\frac{d^2(x_i, x_j)}{\mu \varepsilon_{i,j}}\right) \qquad (۶)$$

که در آن، $d(x_i, x_j)$ میزان فاصله بین دو نمونه است که در اینجا بر اساس معیار ضریب همبستگی پیرسون محاسبه شد، μ یک پارامتر است که معمولاً به صورت تجربی تنظیم می‌شود و نهایتاً $\varepsilon_{i,j}$ نیز یک عبارت است که بر اساس لوکالیتی k-همسایه هر نمونه داده، طبق رابطه (۷) به دست می‌آید [۱۸]:

$$\varepsilon_{i,j} = \frac{\text{mean}(d(x_i,k_i)) + \text{mean}(d(x_j,k_j)) + d(x_i,x_j)}{3} \qquad (۷)$$

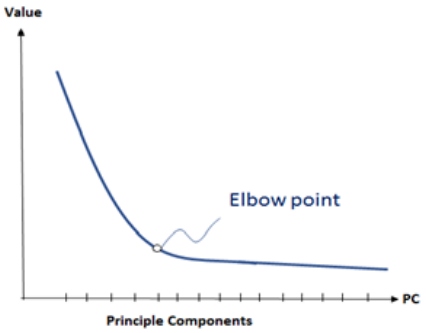
که در آن $\text{mean}(d(x_i, k_i))$ میانگین فاصله بین نمونه x_i و k-همسایه آن، $\text{mean}(d(x_j, k_j))$ میانگین فاصله بین نمونه x_j و k-همسایه آن و $d(x_i, x_j)$ میزان فاصله بین دو نمونه x_i و x_j می‌باشد.

به‌طور خلاصه می‌توان گفت که کرنل گاوسی مبتنی بر گراف معرفی شده، بر اساس لوکالیتی k-همسایه هر سلول، ماتریس تشابه را از روی ماتریس فاصله محاسبه می‌کند. با این کار در واقع فضای ویژگی‌های ورودی توسط یک نگاشت غیرخطی به فضای جدید منتقل می‌شود سپس با استفاده از روش PCA در این فضای جدید، بردارها و مقادیر ویژه که همان مؤلفه‌های اصلی هستند، استخراج می‌شوند. با این ترفند در واقع می‌توان بر مسئله خطی بودن PCA غلبه کرد.

ساخت مجموعه ویژگی نهایی

بعد از محاسبه ماتریس تشابه و در ادامه کار، تکنیک PCA را بر روی ماتریس تشابه حاصل اعمال کرده و پس از به دست آمدن مؤلفه‌های اصلی (PC)، از بین آن‌ها و با استفاده از روش Elbow [۱۷] بهترین مؤلفه‌ها به‌عنوان یک مجموعه ویژگی با اهمیت بالا (PC_i)، استخراج می‌شوند. روش Elbow به این صورت عمل می‌کند که مؤلفه‌های اصلی به دست آمده را بر اساس ارزش و به صورت نزولی بر روی محور مختصات رسم می‌کند و جایی‌که در نمودار شکست حاصل می‌شود را به عنوان نقطه شکست در نظر می‌گیرد. مؤلفه‌های اصلی واقع شده قبل از این نقطه را به‌عنوان مؤلفه‌های اصلی یا مجموعه ویژگی‌های مهم در نظر گرفته شد (شکل ۳). در صورتی‌که از روش Elbow برای انتخاب مؤلفه‌های اصلی مهم‌تر استفاده نشود، باید تعداد مؤلفه‌ها را یک عدد ثابت از قبل تعیین شده در نظر گرفت که این باعث می‌شود انعطاف‌پذیری روش پیشنهادی زیر سؤال برود. با این حال استفاده از روش Elbow باعث می‌شود که همواره با توجه به مجموعه داده مورد نظر، سیستم به طور خودکار بعد از اعمال PCA، مؤلفه‌های با اهمیت بالا را به دست آورده که باعث انعطاف‌پذیری و داده محور شدن روش

پیشنهادی می‌شود.



شکل ۳: استفاده از روش Elbow جهت استخراج مؤلفه‌های اصلی مهم‌تر و ساخت مجموعه ویژگی نهایی

خوشه‌بندی و شناسایی جمعیت‌های سلولی

بعد از فرآیند کاهش ابعاد و استخراج نشانگرهای مهم سلولی، در نهایت عملیات خوشه‌بندی سلول‌ها در فضای جدید و کاهش یافته انجام می‌شود تا به این ترتیب جمعیت‌های سلولی شناسایی شوند. یکی از روش‌های خوشه‌بندی معروف و پرکاربرد که در داده‌های توالی‌یابی RNA سلول–منفرد نیز به دفعات استفاده شده است روش kmeans است [۱۷]. در این مطالعه نیز به‌خاطر سادگی و سرعت بالا، از روش خوشه‌بندی kmeans استفاده شد. یکی از مشکلات روش خوشه‌بندی kmeans ناپایداری آن است که به خاطر انتخاب تصادفی مراکز اولیه رخ می‌دهد. برای غلبه بر این چالش، روشی به‌نام kmeans++ ارائه شد که تا حدودی این مشکل را برطرف کرده و یک الگوریتم خوشه‌بندی پایدارتر را ارائه می‌دهد [۱۹]. در روش kmeans++، ابتدا طی یک فرآیند تکراری و با انتخاب و آزمایش نقاط مرکزی متفاوت بهترین نقاط مرکزی اولیه شناسایی می‌شوند، سپس خوشه‌بندی استاندارد kmeans را با استفاده از این نقاط انجام می‌دهد. هرچند عملیات پیدا کردن این نقاط مرکزی زمان‌بر است؛ اما از آنجا که باعث می‌شود زمان همگرا شدن روش kmeans استاندارد کاهش یابد، به نحوی آن زمان اضافه نیز جبران می‌شود. یکی دیگر از چالش‌های موجود در روش‌های خوشه‌بندی مبتنی بر پارتیشن مثل kmeans، تعیین تعداد بهینه خوشه‌ها یا همان مقدار k است که در این مطالعه و همان‌طور که در مرحله اول روش پیشنهادی تشریح شد، یک رویکرد جدید مبتنی بر گراف برای غلبه بر این چالش ارائه شد.

نتایج

روش پیشنهادی با استفاده از زبان R و بر روی سه مجموعه داده توالی‌یابی RNA سلول–منفرد معروف Kolodziejczyk و همکاران [۲۰]، Buettner و همکاران [۲۱] و Usoskin و همکاران [۲۲] پیاده‌سازی شد (جدول ۱).

جدول ۱: مشخصات مجموعه داده‌های scRNA-seq مورد استفاده جهت ارزیابی روش ارائه شده

مجموعه داده	مرجع	تعداد خوشه‌ها	تعداد ژن‌ها (ابعاد)	تعداد سلول
Kolod	[۲۰]	۳	۱۳۴۷۳	۷۰۴
Buettner	[۲۱]	۳	۹۵۷۳	۱۸۲
Usoskin	[۲۲]	۴	۱۷۷۷۲	۶۲۲

پیچیدگی محاسباتی روش پیشنهادی

در روش پیشنهادی برای تخمین تعداد خوشه‌ها ابتدا سلول‌ها بر روی یک گراف نگاشت شدند که برای این کار نیاز به محاسبه ماتریس مجاورت است. این کار دارای پیچیدگی زمانی $O(n^2)$ است که در آن n تعداد سلول‌ها بود. همچنین در گراف مربوطه برای پیدا کردن نزدیک‌ترین گره چگال‌تر از هر گره، از الگوریتم معروف دیکسترا استفاده شد که این الگوریتم نیز دارای پیچیدگی $O(n^2)$ است. در ادامه جهت کاهش ابعاد مسئله نیز کار اصلی محاسبه ماتریس فاصله سلول–سلول است که این عملیات نیز دارای پیچیدگی $O(n^2)$ می‌باشد. با توجه به عملیات اصلی ذکر شده می‌توان گفت پیچیدگی محاسباتی روش پیشنهادی از مرتبه $O(n^2)$ است.

پارامترهای ارزیابی

در این پژوهش جهت ارزیابی کیفیت خوشه‌بندی در شناسایی جمعیت‌های سلولی از سه معیار معروف خوشه‌بندی یعنی ARI، NMI و Purity استفاده شد.

ARI(*Adjusted Rand Index*): فرض کنید که ما n سلول را به k خوشه تقسیم می‌کنیم و $\{u_i\}_{i=1}^{n}$ نشان‌دهنده‌ی برچسب‌های نهایی تولید شده توسط یک روش خوشه‌بندی باشد. همچنین فرض کنید که $\{v_i\}_{i=1}^{n}$ نشان‌دهنده برچسب‌های واقعی هر سلول (نوع سلولی صحیح) باشد. با توجه به دو تعریف ذکر شده، ARI طبق رابطه (۸) محاسبه شد [۱۷]:

$$ARI = \frac{\sum_{ls} \binom{n_{ls}}{2} - \left[\sum_{l} \binom{n_l}{2} \sum_{s} \binom{n_s}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_{l} \binom{n_l}{2} + \sum_{s} \binom{n_s}{2}\right] - \left[\sum_{l} \binom{n_l}{2} \sum_{s} \binom{n_s}{2}\right] / \binom{n}{2}} \quad (۸)$$

در این رابطه، l و s اندیس مورد اشاره به k خوشه بودند و $n_l = \sum_i^n I(u_i = l)$، $n_s = \sum_i^n I(v_i = s)$ و $n_{ls} = \sum_{i,j}^n I(u_i = l) I(v_i = s)$. در این روابط، $I(x = y)$ همان تابع همانی است که وقتی $x = y$ باشد مقدار آن ۱ می‌باشد و در غیر این صورت صفر است. به طور خلاصه اگر برچسب خوشه‌های تولید شده توسط یک روش خوشه‌بندی کاملاً متناظر با برچسب‌های واقعی باشد آنگاه مقدار ARI برابر ۱ می‌باشد و در غیر این صورت مقدار ARI به نسبت عدم تطابق‌های موجود کاهش می‌یابد.

NMI(*Normalized Mutual Index*): فرض کنید $p_l = \frac{n_l}{n}$، $q_s = \frac{n_s}{n}$ و $z_{ls} = \frac{n_{ls}}{n}$. حال آنتروپی مربوط به هر جواب خوشه‌بندی (مربوط به v و u) را می‌توان این‌گونه تعریف کرد: $h(u) = -\sum_l p_l * \log p_l$ و $h(v) = -\sum_s q_s * \log q_s$. همچنین میزان اطلاعات منحصربه‌فرد بین این دو جواب خوشه‌بندی به‌صورت $i(u,v) = \sum_{l,s} z_{ls} \log(z_{ls}/p_l/q_s)$ تعریف می‌شود. حال با توجه به این روابط، می‌توان معیار NMI را به صورت رابطه زیر تعریف نمود [۱۷]:

$$NMI = i(u,v) / \overline{h(u)\, h(v)} \quad (۹)$$

همانند ARI، اگر تطابق ۱۰۰ درصدی بین دو جواب خوشه‌بندی u و v وجود داشته باشد، مقدار NMI نیز ۱ می‌شود. به‌طور خلاصه هرچه مقادیر ARI و NMI مربوط به یک روش خوشه‌بندی اعمال شده بر روی یک مجموعه داده به ۱ نزدیک‌تر باشد نشان‌دهنده کیفیت بالای خوشه‌بندی می‌باشد.

خلوص (Purity): معیار خلوص برای خوشه‌هایی که دارای یک کلاس واحد هستند، اندازه‌گیری می‌شود. برای محاسبه آن، برای هر خوشه، تعداد نقاط داده از کلاس معمول در خوشه مورد نظر شمرده می‌شود، سپس تمام خوشه‌ها جمع شده و بر تعداد نقاط داده تقسیم می‌شود [۱۷]:

$$Purity(\Omega, C) = \frac{1}{N} \sum_{k,j} \max |w_k \cap c_j| \quad (۱۰)$$

که در آن $\Omega = \{w_1, w_2, \ldots, w_k\}$ و $C = \{c_1, c_2, \ldots, c_j\}$ مجموع دسته‌ها (جواب‌های درست خوشه‌بندی) می‌باشد.

نتایج مربوط به تخمین تعداد خوشه‌ها

با توجه به روش پیشنهادی، اولین مرحله در شناسایی جمعیت‌های سلولی تخمین تعداد خوشه‌ها یا همان جمعیت‌های سلولی می‌باشد که نتایج روش مبتنی بر گراف پیشنهادی در مقایسه با سایر روش‌ها در زمینه تخمین تعداد خوشه‌ها در جدول ۲ خلاصه شد. به‌طورکلی، روش‌های رایج تعیین تعداد خوشه‌ها به دو دسته روش‌های مستقیم و روش‌های مبتنی بر آزمون‌های آماری تقسیم می‌شوند. روش‌های مستقیم به دنبال بهینه‌سازی یک معیار به‌خصوص، مانند مجموع مربعات فواصل درون خوشه‌ای (Within-cluster Sum of Square) یا سیلوئت میانگین هستند؛ از جمله این روش‌ها می‌توان به روش Elbow و روش‌های مبتنی بر معیار سیلوئت اشاره کرد [۱۷].

ایده‌اصلی روش Elbow به‌دست‌آوردن تعداد خوشه‌ها به نحوی است که مجموع فواصل درون خوشه‌ای داده‌ها (یا مجموع مربعات فواصل درون خوشه‌ای) حداقل شود. این روش مجموع فواصل درون خوشه‌ای داده‌ها را به عنوان تابعی از تعداد خوشه‌ها در نظر می‌گیرد. به این ترتیب تعداد خوشه‌ها به نحوی انتخاب می‌شوند که افزودن یک خوشه دیگر، بهبودی در حداقل‌سازی WSS ایجاد نکند.

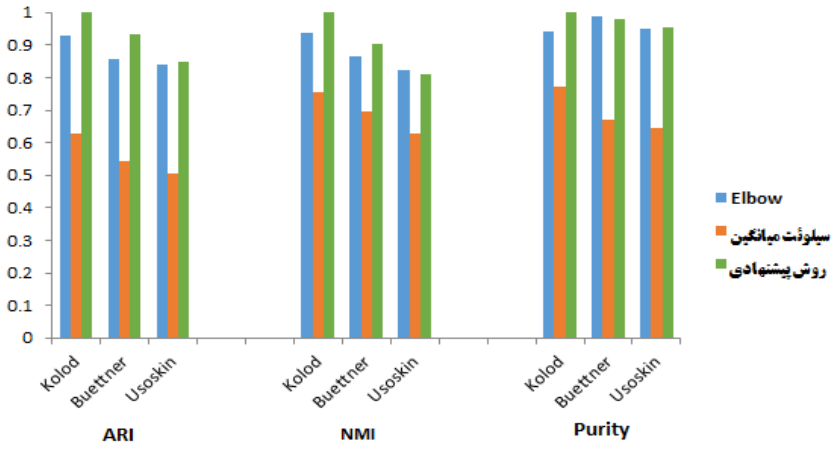
در روش سیلوئت میانگین، دو معیار فواصل درون خوشه‌ای و برون خوشه‌ای به‌طور هم‌زمان در نظر گرفته می‌شود. در واقع این معیار مشخص می‌کند که پراکندگی داده‌ها در خوشه‌ها به چه صورت است. هر چه مقدار سیلوئت بالاتر باشد، کیفیت خوشه‌بندی نیز بالاتر است. در روش سیلوئت میانگین، الگوریتم خوشه‌بندی به ازای مقادیر مختلف K اجرا شده و به ازای هر اجرا، معیار سیلوئت برای هر یک از اعضای خوشه‌ها محاسبه می‌شود. سپس از سیلوئت‌های به‌دست آمده معدل گرفته می‌شود. مقدار بهینه K مقداری است که به ازای آن، سیلوئت میانگین ماکزیمم شود.

نتایج مربوط به تخمین تعداد خوشه‌ها توسط روش‌های رایج Elbow و سیلوئت میانگین به‌همراه روش مبتنی بر گراف پیشنهادی در جدول ۲ نشان داده شد.

جدول ۲: نتایج تخمین تعداد خوشه‌ها (K)

		روش تخمین تعداد خوشه‌ها			
		تعداد واقعی خوشه‌ها	سیلوئت میانگین	Elbow	روش پیشنهادی مبتنی بر گراف
مجموعه داده	Kolod	۳	۲	۴	۳
	Buettner	۳	۲	۴	۴
	Usoskin	۴	۲	۴	۶

روش پیشنهادی جهت شناسایی جمعیت‌های سلولی بر روی سه مجموعه داده مذکور در جدول ۱ و به ازای تعداد خوشه‌های تخمین زده شده در جدول ۲ اجرا شد و نتایج به دست آمده برای معیارهای ARI، NMI و Purity در شکل ۴ خلاصه شد. همان‌طور که مشاهده شد، روش پیشنهادی با تخمین مناسب تعداد خوشه‌ها و استخراج ویژگی‌های مناسب در مرحله کاهش ابعاد، با دقت بالایی اقدام به خوشه‌بندی داده‌ها می‌کند.



شکل ۴: نتایج خوشه‌بندی روش پیشنهادی جهت شناسائی جمعیت‌های سلولی با استفاده از روش‌های Elbow، سیلوئت میانگین و روش مبتنی بر گراف پیشنهادی جهت تخمین تعداد خوشه‌ها

ارزیابی دقت خوشه‌بندی روش پیشنهادی

نتایج حاصل بر روی سه مجموعه داده مختلف نشان می‌دهد که روش پیشنهادی در مقایسه با روش‌های سنتی و روش‌های جدید ارائه شده برای داده‌های توالی‌یابی RNA سلول-منفرد، علاوه بر تخمین مناسب تعداد جمعیت‌های سلولی، دقت خوشه‌بندی را نیز افزایش می‌دهد.

برای این منظور، روش پیشنهادی را با سه روش خوشه‌بندی سنتی kmeans، GMM و روش خوشه‌بندی سلسله مراتبی HCLUST(Hierarchical Clustering) و همچنین سه روش خوشه‌بندی جدید pcaReduce، Sincera و SNN-Cliq که مختص داده‌های توالی‌یابی RNA سلول-منفرد طراحی شده‌اند، مقایسه شد. نتایج به دست آمده برای معیارهای ARI، NMI و Purity به صورت تفصیلی در جدول‌های ۳، ۴ و ۵ نشان داده شد. همان‌طور که مشاهده شد، دقت خوشه‌بندی روش پیشنهادی تنها در مجموعه داده Kolod با روش Sincera برابری می‌کند و در سایر مجموعه داده‌ها در مقایسه با روش‌های دیگر (حتی Sincera)، دقت به مراتب بالاتری را حاصل می‌کند.

جدول ۳: نتایج پیاده‌سازی روش پیشنهادی و مقایسه با روش‌های معروف (ARI%)

		روش						
		HCLUST	GMM	kmeans	SNN-Cliq	pcaReduce	Sincera	روش پیشنهادی
مجموعه داده	Kolod	۶۱/۸	۹۰/۶	۷۷/۳	۷۳,۵	۸۷/۶	۱۰۰	۱۰۰
	Buettner	۳۱/۸	۴۶/۳	۵۳/۳	۴۴/۸	۴۴/۱	۸۰/۳	**۹۳/۴۷**
	Usoskin	۲۲/۳	۵۵/۲	۴۲/۶	۷۲/۶	۶۰/۴	۷۵/۶	**۸۴/۶۹**

جدول ۴: نتایج پیاده‌سازی روش پیشنهادی و مقایسه با روش‌های معروف (NMI%)

		روش						
		HCLUST	GMM	Kmeans	SNN-Cliq	pcaReduce	Sincera	روش پیشنهادی
مجموعه داده	Kolod	**۵۷/۷۸**	**۹۳/۳۳**	**۸۰/۱۱**	**۷۸/۳۶**	**۹۰/۲۳**	۱۰۰	۱۰۰
	Buettner	۳۸/۹۳	۴۴,۱۱	۵۵,۴۵	۴۲/۵۵	۵۱/۶۳	۸۳,۷۱	**۹۰/۴۸**
	Usoskin	۱۸/۲۶	۶۰/۲۳	۵۰/۴۹	۷۳/۵۵	۶۱/۳۷	۷۷,۵۳	**۸۱/۰۹**

جدول ۵: نتایج پیاده‌سازی روش پیشنهادی و مقایسه با روش‌های معروف (Purity%)

		روش						
		HCLUST	GMM	Kmeans	SNN-Cliq	pcaReduce	Sincera	روش پیشنهادی
مجموعه داده	Kolod	۶۳/۴۲	۹۶/۱۸	۸۱/۲۳	۸۰,۳۲	۹۱,۵۱	۱۰۰	۱۰۰
	Buettner	۴۵/۲۵	۵۴/۷۳	۵۸/۱۴	۵۰/۷۹	۴۸/۹۳	۸۵/۲۲	**۹۷/۸**
	Usoskin	۲۹/۳۲	۶۶/۰۸	۵۰/۳۹	۸۱/۳۸	۶۶/۷۲	۸۲/۱۷	**۹۵/۳۴**

بحث و نتیجه‌گیری

در این پژوهش به مسئله خوشه‌بندی داده‌های توالی‌یابی RNA سلول–منفرد پرداخته شد که منجر به شناسایی جمعیت‌های سلولی شد. همان‌طور که اشاره شد تجزیه‌وتحلیل این نوع از داده‌ها با چالش‌هایی مثل وجود نویز، ابعاد بالا و عدم در اختیار داشتن اطلاعات قبلی دقیق مثل تعداد و نوع جمعیت‌های سلولی همراه است.

در سال‌های اخیر، روش‌های محاسباتی جدید جهت غلبه بر این چالش‌ها معرفی‌شده‌اند. بدون تردید، اطلاعات دقیق و فوق‌العاده ارزشمندی که فناوری سلول–منفرد فراهم می‌کند، به دلیل کسب داده‌های پیچیده، نیازمندی‌های مربوط به ذخیره داده‌های فراوان به علاوه موانع موجود برای پردازش و مدیریت این داده‌ها، هزینه قابل‌توجهی را در بر می‌گیرد.

انتظار می‌رود که بزرگی داده‌های سلول–منفرد در آینده نزدیک به میلیون‌ها سلول در پروژه‌های جدید به عنوان اطلس سلول انسان (Human Cell Atlas) افزایش یابد. داده‌ها در این مقیاس، مشکلاتی را در پردازش پایه‌ای همانند محاسبه ماتریس کواریانس و در ارزیابی آماری و استحکام نتایج، به وجود می‌آورند. در نتیجه، برای پردازش این داده‌ها، نیاز به پهنای باند بالا، الگوریتم‌های موازی، روش‌های محاسباتی و کامپیوترهایی با عملکرد بالا اجتناب‌ناپذیر است.

پیشرفت روزافزون علم یادگیری ماشین سبب شده که کاربرد آن در زمینه تحلیل داده‌های سلول–منفرد بیش از پیش مورد توجه قرار گیرد؛ بنابراین روش‌های تحلیلی مبتنی بر یادگیری ماشین نقش مهم و تعیین کننده‌ای در این زمینه بازی می‌کند.

در همین راستا، روش پیشنهادی با کمک یک روش یادگیری ماشین بدون نظارت و مبتنی بر تئوری گراف، برای غلبه بر این چالش‌ها ارائه شد. با توجه به نتایج به دست آمده، مزایای روش پیشنهادی را می‌توان از چند جنبه مورد بررسی قرار داد.

اولاً این روش بدون داشتن هیچ اطلاعات قبلی در مورد تعداد و نوع جمعیت‌های سلولی، اقدام به تخمین مناسب تعداد خوشه‌ها با استفاده از یک روش مبتنی بر گراف می‌کند. این موضوع می‌تواند در شناسایی جمعیت‌های سلولی نادر مثل بیماری سرطان به محققان کمک کند. اگرچه تعداد جمعیت‌ها

می‌تواند با استفاده از اطلاعات زیست‌شناسی پیشین، تعیین شود؛ اما تعداد ثابت برای داده‌های سلول–منفرد مناسب نیست؛ زیرا ممکن است زیرمجموعه‌های ناشناخته‌ای وجود داشته باشد که شناسایی آن‌ها بتواند در ایجاد دانش جدید در این زمینه مؤثر باشد. یکی از دلایل این است که برخی از بیماری‌ها مثل سرطان ممکن است ترکیب سلولی را تغییر دهند و جمعیت‌های سلولی جدید و ناشناخته را که باید مورد توجه قرار گیرد، معرفی می‌کنند.

ثانیاً در بخش کاهش ابعاد مسئله با استفاده از یک کرنل گاوسی مبتنی بر گراف پیشنهادی، یک روش کاهش ابعاد بدون نظارت غیرخطی ارائه شد که با استخراج مؤلفه‌ها و ویژگی‌های اصلی به‌عنوان نشانگرهای مهم سلولی، فرآیند خوشه‌بندی و شناسایی جمعیت‌های سلولی را بهبود بخشد. این موضوع نیز از آن جهت حائز اهمیت است که معمولاً کشف یک رابطه خطی در مجموعه داده‌های توالی‌یابی سلول–منفرد تقریباً غیرممکن است و در نتیجه استفاده از روش‌های خطی کاهش ابعاد کمک چندانی به افزایش کیفیت خوشه‌بندی نمی‌کند.

نتایج حاصل از خوشه‌بندی بر روی سه مجموعه داده سلول–منفرد معروف Kolodziejczyk و همکاران [۲۰]، Buettner و همکاران [۲۱] و Usoskin و همکاران [۲۲] نشان می‌دهد که روش پیشنهادی با دقت بالایی اقدام به شناسایی جمعیت‌های سلولی می‌نماید. برای ارزیابی نهایی، روش پیشنهادی با سه روش خوشه‌بندی سنتی kmeans [۱۷]، GMM [۱۷] و روش خوشه‌بندی سلسله مراتبی (HCLUST) [۱۷] و همچنین سه روش خوشه‌بندی جدید و مختص داده‌های سلول–منفرد به نام‌های pcaReduce [۱۱]، Sincera [۱۲] و SNN-Cliq [۱۳] مقایسه شد. نتایج مربوط به پارامتر ARI، به‌عنوان مهم‌ترین پارامتر خوشه‌بندی، برای روش پیشنهادی و بر روی سه مجموعه داده Kolod، Buettner و Usoskin به ترتیب برابر ۱۰۰، ۹۳/۴۷ و ۸۴/۶۹ به دست آمد که در مقایسه با سایر روش‌های مذکور، بالاترین دقت مربوط به روش Sincera با مقادیر ۱۰۰، ۸۰,۳ و ۷۵/۶ بود. نتایج حاصل از پارامتر NMI نیز برای روش پیشنهادی به ترتیب ۱۰۰، ۹۰/۴۸ و ۸۱/۰۹ بودند و در مقایسه با سایر روش‌ها نیز Sincera با نتایج ۱۰۰، ۸۳/۷۱ و ۷۷/۵۳ بالاترین دقت را به دست آورد. همچنین نتایج مربوط به پارامتر purity نیز مشابه دو پارامتر قبلی به دست آمد، به‌طوری که روش پیشنهادی دقتی معادل ۱۰۰، ۹۷/۸ و ۹۵/۳۴ در مقایسه با دقت ۱۰۰، ۸۵/۲۲ و ۸۲/۱۷ روش Sincera به دست آورد. همان‌طور که نتایج به دست آمده برای پارامترهای معروف خوشه‌بندی ARI، NMI و purity نشان داد، روش پیشنهادی نسبت به سایر روش‌ها بر روی دو مجموعه داده Buettner و Usoskin از دقت به مراتب بالاتری برخوردار است و تنها در مجموعه داده Kolod دقت برابری با روش Sincera به دست آمد.

نکته‌ای که در انتها باید به آن اشاره کرد این است که در روش پیشنهادی، در چندین مرحله اقدام به محاسبه ماتریس فاصله و در نتیجه ماتریس تشابه سلول–سلول شد. در روش پیشنهادی از ضریب همبستگی پیرسون برای محاسبه این ماتریس استفاده شد که معمولاً برای داده‌هایی با توزیع نرمال یا تعداد داده‌های زیاد استفاده می‌شود. درصورتی‌که با داده‌هایی مواجه شدیم که دارای این شرایط نباشند، شاید لازم باشد که از سایر معیارها نظیر ضریب همبستگی اسپیرمن استفاده کرد. به همین دلیل شاید بتوان با توسعه روش پیشنهادی و تلفیق مناسب این ضرایب همبستگی، روشی مستقل از داده ارائه داد که فارغ از توزیع آماری حاکم بر داده‌ها و تعداد آن‌ها اقدام به خوشه‌بندی دقیق سلول‌ها نماید.

تعارض منافع

این مقاله هیچ‌گونه تضاد منافعی ندارد.

References

1. Wang Z, Gerstein M, Snyder M. RNA-Seq: A Revolutionary Tool for Transcriptomics. Nat Rev Genet 2009;10(1):57-63. doi: 10.1038/nrg2484.

2. Nagalakshmi U, Waern K, Snyder M. RNA-Seq: A Method for Comprehensive Transcriptome Analysis. Curr Protoc Mol Biol 2010;Chapter 4:Unit 4.11.1-13. doi: 10.1002/0471142727.mb0411s89.

3. Eberwine J, Sul JY, Bartfai T, Kim J. The promise of single-cell sequencing. Nat Methods 2014;11(1):25-7. doi: 10.1038/nmeth.2769.

4. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq Whole-Transcriptome Analysis of a Single Cell. Nat Methods 2009;6(5):377-82. doi: 10.1038/nmeth.1315.

5. Poirion OB, Zhu X, Ching T, Garmire L. Single-Cell Transcriptomics Bioinformatics and Computational Challenges. Front Genet 2016;7:163. doi: 10.3389/fgene.2016.00163.

6. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression

analysis. Nat Methods 2014;11(7):740-2. doi:10.1038/nmeth.2967
7. Yanglan G, Ning L, Guobing Z, Yongchang X, Jihong G. Identification of cancer subtypes from single-cell RNA-seq data using a consensus clustering method. BMC Med Genomics 2018;11(Suppl 6):117. doi: 10.1186/s12920-018-0433-z.
8. Nguyen A, Khoo WH, Moran I, Croucher PI, Phan TG. Single Cell RNA Sequencing of Rare Immune Cell Populations. Front Immunol 2018;9:1553. doi: 10.3389/fimmu.2018.01553.
9. Xiaoqing Yu, Y. Ann Chen, Jose R. Conejo-Garcia, Christine H. Chung, Xuefeng Wang. Estimation of immune cell content in tumor using single-cell RNA-seq reference data. BMC Cancer 2019;19(1):715. doi: 10.1186/s12885-019-5927-3.
10. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science 2014;343(6172):776-9. doi: 10.1126/science.1247651.
11. Zurauskiene J, Yau C. pcaReduce: Hierarchical Clustering of Single Cell Transcriptional Profiles. BMC Bioinformatics 2016;17:140. doi: 10.1186/s12859-016-0984-y.
12. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. Sincera: A pipeline for single-cell RNA-seq profiling analysis. PLoS Comput Biol 2015;11(11):e1004575. doi: 10.1371/journal.pcbi.1004575.
13. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics 2015;31(12):1974-80. doi: 10.1093/bioinformatics/btv088.
14. Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, Clifford Stein. Introduction to Algorithms. 3rd ed. Cambridge: MIT Press; 2009.
15. Bie R, Mehmood R, Ruan S, Sun Y, Dawood H. Adaptive fuzzy clustering by fast search and find of density peaks. Personal and Ubiquitous Computing 2016;20(5):1-9. doi: 10.1007/s00779-016-0954-4
16. Pouyan MB, Kostka D. Random forest based similarity learning for single cell RNA sequencing data. Bioinformatics 2018;34(13):i79-i88. doi: 10.1093/bioinformatics/bty260.
17. Alpaydin E. Introduction to Machine Learning. 3rd ed. Cambridge: MIT Press; 2015.
18. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 2014;11(3):333-7. doi: 10.1038/nmeth.2810.
19. Arthur D, Vassilvitskii S. K-means++: The advantages of careful seeding. Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms; 2007 Jun; Philadelphia, PA, USA: Society for Industrial and Applied Mathematics; 2007. p. 1027-35. https://dl.acm.org/doi/10.5555/1283383.1283494
20. Kolodziejczyk AA, Kim JK, Tsang JC, Ilicic T, Henriksson J, Natarajan KN, et al. Single cell rna-sequencing of pluripotent states unlocks modular transcriptional variation. Cell Stem Cell 2015;17(4):471-85. doi: 10.1016/j.stem.2015.09.011.
21. Buettner F, Natarajan K, Casale F, Proserpio V, Scialdone A, Theis F, et al. Computational Analysis of Cell-To-Cell Heterogeneity in Single-Cell RNA-sequencing Data Reveals Hidden Subpopulations of Cells. Nat Biotechnol 2015;33(2):155-60. doi: 10.1038/nbt.3102.
22. Usoskin D, Furlan A, Islam S, Abdo H, Lönnerberg P, Lou D, et al. Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. Nat Neurosci 2015;18(1):145-53. doi: 10.1038/nn.3881.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2020; 7(1): 60-72

A Graph-Based Clustering Approach to Identify Cell Populations in Single-Cell RNA Sequencing Data

Einipour Amin[1], Mosleh Mohammad[2]*, Ansari-Asl Karim[3]

• **Received:** 16 Jul, 2019 • **Accepted:** 14 Dec, 2019

Introduction: The emergence of single-cell RNA-sequencing (scRNA-seq) technology has provided new information about the structure of cells, and provided data with very high resolution of the expression of different genes for each cell at a single time. One of the main uses of scRNA-seq is data clustering based on expressed genes, which sometimes leads to the detection of rare cell populations. However, the results of the proposed methods mainly depend on the shape of the cell populations and the dimensions of the data. Therefore, it is very important to develop a method that can identify cell populations regardless of these obstacles.

Method: In the proposed method, which was a library method, at first, the number of clusters (cell populations) was estimated. Estimating the number of clusters is important because in the real world, basic information such as the number and type of cell populations is not available. Thereafter, using a graph-based Gaussian kernel, while reducing the dimensions of the problem, the cell populations were identified by means of the kmeans++ clustering.

Results: The results of the implementation showed that the proposed method can achieve an acceptable improvement compared to other machine learning methods presented in this regard. For example, for the ARI criterion, values of 100, 93.47 and 84.69 were obtained for Kolod, Buettner, and Usoskin single-cell data sets, respectively.

Conclusion: The proposed method can cluster and thus identify cell populations with high accuracy and quality without having any basic information about the number and type of cell populations, regardless of the high dimensions of the problem.

Keywords: Single-cell RNA-sequencing, Clustering, Identification of Cell Populations, Graph-based Gaussian Kernel

• **Citation:** Einipour A, Mosleh M, Ansari-Asl K. A Graph-Based Clustering Approach to Identify Cell Populations in Single-Cell RNA Sequencing Data. Journal of Health and Biomedical Informatics 2020; 7(1): 60-72. [In Persian]

1. Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran
2. Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran
3. Department of Electrical Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran

***Corresponding Author:** Mohammad Mosleh
Address: Computer Engineering Dept., Faculty of Engineering, Dezful Branch, Islamic Azad University, Daneshgah Blvd, Dezful, Iran. P.O. Box: 313
• **Tel:** 061-42420601 • **Email:** mosleh@iaud.ac.ir