دانشگاه یزد

دانشکده‌ی ریاضی

گروه علوم کامپیوتر

پایان‌نامه

جهت دریافت درجه‌ی کارشناسی ارشد

علوم کامپیوتر

تحلیل خوشه‌بندی برای داده بیان ژن با استفاده از تجزیه‌ی ماتریس نامنفی

استاد راهنما:

دکتر سید ابوالفضل شاهزاده‌فاضلی

استاد مشاور:

دکتر مهدیه هاشمی‌نژاد

پژوهش ونگارش:

فاطمه السادات فاطمیون

مهر ۱۳۹۰


بسم الله الرحمن الرحيم

سپاس‌گزاری

سپاس خدا را که مرا به راه دانش رهنمون شد و هموست که فرصت شاگردی اساتید گرامی را بر من ارزانی داشت تا چراغی بر تاریکی جهلم باشند.

از استاد بزرگوارم، جناب آقای دکتر سید ابوالفضل شاهزاده‌فاضلی، استاد راهنمای این پژوهش، که افق‌های جدیدی بر پنجره‌ی ذهنم گشودند، سپاس‌گزارم و رهنمودهای ایشان را ارج می‌نهم.

از خانم دکتر مهدیه هاشمی‌نژاد که زحمت مطالعه و مشاوره این پایان‌نامه را تقبل فرمودند تقدیر و تشکر می‌کنم.

از مادرم که پشتیبان همیشگی و پدرم که بزرگ‌ترین مشوق من در راه فراگیری دانش بوده‌اند کمال سپاس و امتنان را دارم و توفیق جبران هر قطره از دریای بی‌کران مهر این عزیزان را غنیمت می‌شمارم.

چکیده

امروزه حجم عظیمی از مطالعات پزشکی در جهت شناسایی و درمان بیماری‌هایی است که از طریق ژن منتقل می‌شود. برای بررسی و نگهداری اطلاعات ژنتیکی، فناوری‌های مفیدی به وجود آمده است که یکی از آن‌ها، فناوری ریزآرایه می‌باشد. تجزیه و تحلیل اطلاعات به دست آمده از ریزآرایه‌ها به کمک روش‌های داده‌کاوی انجام می‌شود. یکی از این روش‌ها خوشه‌بندی است که می‌تواند در یافتن گروه‌های واقعی و نهفته در داده‌ها مؤثر باشد. همچنین با استفاده از روش‌های کاهش بعد می‌توان مجموعه داده‌هایی با حجم کوچک‌تر از مجموعه داده‌های اصلی تولید کرد و آن را به‌عنوان ورودی روش خوشه‌بندی به‌کار برد.

در این رساله از تجزیه‌ی ماتریس نامنفی (NMF) برای کاهش بعد داده‌های ریزآرایه استفاده می‌شود. همچنین برای مقداردهی اولیه این تجزیه روش‌های تصادفی، تحلیل مولفه اصلی (PCA) و تجزیه‌ی مقدار تکین مضاعف نامنفی (NNDSVD) به‌کار می‌رود. پس از آن با به‌کارگیری روش k-متوسط داده‌های کاهش‌یافته خوشه‌بندی می‌گردد. تحلیل‌های انجام شده در این تحقیق نشان می‌دهد که خوشه‌بندی داده‌های حاصل از NMF+PCA نتایج بهتری را ارائه می‌دهد.

فهرست مطالب

لیست تصاویر

فصل ۱

مقدمه

مقدمه

علم زیست شناسی از قدیمی‌ترین علومی است که بشر به آن توجه داشته است؛ اما از حدود یک قرن پیش، این علم وارد مرحله جدیدی به نام ژنتیک شد. با پیشرفت علم ژنتیک، ژن به‌عنوان عامل کنترل‌کننده ویژگی و عملکرد سلول‌ها شناخته شد. در علم ژنتیک، هنگامی که یک ژن در سلول مورد استفاده قرار می‌گیرد، می‌گویند آن ژن بیان شده یا روشن است. تغییر بیان ژن[۱] سلول در اثر جهش‌های ژنتیکی و شرایط خاص موجب فعالیت غیر عادی سلول و بروز بیماری می‌شود. امروزه حجم عظیمی از مطالعات پزشکی درجهت شناسایی و درمان بیماری‌هایی است که از طریق ژن منتقل می‌شود. همچنین برای بررسی و نگهداری اطلاعات ژنتیکی، فناوری‌های مفیدی به وجود آمده‌است که یکی از این تکنولوژی‌ها، ریزآرایه[۲] می‌باشد.

در یک آزمایش ریزآرایه، توالی ژن‌ها به وسیله رنگ‌های فلورسنت متفاوت علامت‌گذاری شده و روی یک آرایه شیشه‌ای یا پلاستیکی هیبرید می‌شوند. پس از آن، آرایه مزبور توسط اشعه لیزر پویش شده و تصاویری شامل هزاران لکه رنگی ایجاد می‌شود که انجام محاسبات عددی روی آن‌ها اندازه‌ی بیان ژن را نتیجه می‌دهد. سپس داده‌های حاصل از آزمایش در ماتریس بیان ژن ذخیره می‌شود. سطرهای این ماتریس را ژن‌ها و ستون‌های آن را افراد نمونه تشکیل می‌دهد. ماتریس مزبور در کشف و شناسایی پدیده‌های پزشکی ناشناخته، رفع اثرات ناهنجار ژنتیکی و پیشگیری و درمان بیماری‌ها مورد استفاده قرار می‌گیرد.

فناوری ریزآرایه امکان بررسی بیان هزاران ژن را به‌صورت همزمان فراهم می‌سازد و حجم عظیمی از داده‌های بیان ژنی را تولید می‌کند که می‌توان با استفاده از روش‌های داده‌کاوی[۳] دانش مورد نیاز جهت شناسایی و درمان بیماری‌ها را از آن استخراج نمود. در داده‌های ریزآرایه، تعداد ژن‌ها بسیار زیاد و تعداد نمونه‌ها کم می‌باشد. در تجزیه و تحلیل داده‌ها، وجود همه‌ی ژن‌ها ضروری نیست، یعنی ممکن است بسیاری از آن‌ها نیاز نباشند و تعدادی دیگر حاوی اطلاعات تقریباً یکسانی باشند و باعث افزونگی[۴] شوند، پس عمل ساده‌سازی و انتخاب

---

[۱] Gene expression

[۲] Microarray

[۳] Data mining

[۴] Redundancy

ژن‌های مناسب امری ضروری است که این عمل به کمک روش‌های کاهش بعد[۵] انجام می‌شود.

داده‌کاوی یکی از مراحل کشف دانش[۶] است. درکشف دانش ابتدا داده‌های مناسب با توجه به نوع کاربرد جمع آوری و سپس پیش‌پردازش می‌شود بدان معنا که مقادیر گم‌شده[۷] و پرت[۸] آن‌ها جایگزین شده، تبدیل و نرمال‌سازی روی آن‌ها انجام می‌شود. در جایی‌که حجم مجموعه داده بزرگ باشد عمل کاهش بعد داده‌ها نیز انجام می‌شود. هدف در این عمل، تولید مجموعه‌ای کوچک‌تر از داده‌های اولیه است به‌گونه‌ای که نتایج حاصل از این مجموعه با نتایج حاصل از داده اولیه تقریباً یکسان باشد. پس از اجرای مراحل پیش‌پردازش و کاهش بعد داده، فرآیند داده‌کاوی انجام می‌شود. با استفاده از روش‌های داده‌کاوی می‌توان دانش مفید را از داده‌ها استخراج نمود. از روش‌های کاهش بعد می‌توان برای شناسایی ابعاد و ویژگی‌های مطلوب برای نمایش داده‌ها در فرآیند داده‌کاوی استفاده نمود.

تاکنون روش‌های زیادی از جمله روش‌های تحلیل مؤلفه اصلی[۹]، تحلیل مؤلفه مستقل[۱۰]، تجزیه‌ی نیمه‌گسسته[۱۱]، تجزیه‌ی مقادیر تکین[۱۲] و تجزیه‌ی ماتریس نامنفی[۱۳] برای کاهش بعد داده‌ها معرفی شده است. در برخی از کاربردها به‌عنوان مثال تحلیل و داده‌کاوی داده‌های ریزآرایه، لازم است به منظور کاهش بعد از روش‌هایی استفاده نمود که شرایط نامنفی بودن ماتریس‌های کاهش یافته را حفظ کند که از بین تمام روش‌های مزبور تنها روش تجزیه‌ی ماتریس نامنفی این شرایط راحفظ می‌کند.

تجزیه‌ی ماتریس نامنفی برای اولین بار در سال ۱۹۷۰ تحت عنوان «تجزیه‌ی منحنی‌های خود سازمانده»[۱۴] ارائه شد. این تجزیه با نام تجزیه‌ی ماتریس مثبت به صورت گسترده توسط محققان اسپانیایی مطالعه گردید.

---

[۵] Dimension reduction

[۶] Knowledge discovery

[۷] Missing values

[۸] Outlier

[۹] Principal component analysis(PCA)

[۱۰] Independent component analysis(ICA)

[۱۱] Semi-discrete decomposition(SSD)

[۱۲] Singular value decomposition(SVD)

[۱۳] Nonnegative matrix factorization(NMF)

[۱۴] Self modeling curve resolution

در سال ۱۹۹۹تحقیقات «لی» و «سونگ» [۱۵]موجب توجه محققان به این تجزیه و کاربرد آن در زمینه داده‌کاوی و یادگیری ماشین شد. پس از انتشار مقاله آن‌ها[۴۲] ، تجزیه‌ی مزبور با نام تجزیه‌ی ماتریس نامنفی معرفی گردید. این تجزیه بسیاری از ساختارهای داده‌ی اصلی را حفظ و نامنفی بودن هر دو ماتریس پایه و ضرایب را ضمانت می‌کند. در تحلیل داده‌های بیان ژن ضروری است که پایه‌ها نامنفی نگه داشته شود، بنابراین برای کاهش بعد این داده‌ها از روش تجزیه‌ی ماتریس نامنفی استفاده شده است.

پس از کاهش بعد داده‌ها از روش‌های داده‌کاوی جهت استخراج الگوهای مفید از داده‌ها استفاده می‌شود. خوشه‌بندی[۱۶] یکی از روش‌های داده‌کاوی است که به مفهوم تقسیم داده‌ها به گروه‌هایی از داده‌های مشابه می‌باشد. قابل ذکر است که اعتبار نتایج این روش به انتخاب معیار شباهت بین داده‌ها وابسته است. انتخاب معیار شباهت مناسب، در داده‌هایی که دارای تعدد ویژگی یا بعد زیاد می‌باشد، کاری مشکل است و ممکن است به نتایج نامطلوب منجر شود. همچنین پیچیدگی محاسباتی روش‌های خوشه‌بندی زیاد بوده و در برخورد با مجموعه داده‌های بزرگ ناپایدار است. جهت رفع این مشکل‌ها می‌توان از روش‌های کاهش بعد در جهت تولید مجموعه داده‌هایی با حجم کوچک‌تر و ویژگی‌های مطلوب از مجموعه داده‌های اصلی استفاده نمود و آن را به‌عنوان ورودی روش خوشه‌بندی به‌کار برد. روش‌های خوشه‌بندی به دو دسته کلی خوشه‌بندی سلسله مراتبی[۱۷] و خوشه‌بندی افرازی[۱۸] تقسیم می‌شوند. روش خوشه‌بندی k-متوسط[۱۹] یکی از روش‌های خوشه‌بندی افرازی است. الگوریتم این روش بسیار ساده و قابل فهم است و عملکرد آن شبیه عملکرد روش تجزیه‌ی ماتریس نامنفی است و همچنین دارای نتایج مطلوب در زمینه‌ی خوشه‌بندی داده‌ها می‌باشد [۳۲]، [۵۱].

در این تحقیق روش تجزیه‌ی ماتریس نامنفی به‌عنوان یک روش تکراری در کاهش بعد داده‌های بیان ژن استفاده خواهد شد. ایده اصلی در یک روش تکراری این است که ابتدا مسئله به یک شکل معادل نوشته شود و سپس با شروع از یک جواب اولیه، دنباله‌ای همگرا به جواب اصلی تولید گردد. در روش تجزیه‌ی ماتریس نامنفی می‌توان از روش‌های مختلفی برای مقداردهی اولیه استفاده نمود. الگوریتمی که توسط «لی» و«سونگ» ارائه شد از روش تصادفی جهت مقداردهی اولیه عوامل تجزیه استفاده می‌نمود. تاکنون مطالعات زیادی جهت

---

[۱۵] Lee and Seung

[۱۶] Clustering

[۱۷] Hierarchical clustering

[۱۸] Partition clustering

[۱۹] k-means

استفاده از یک روش مقداردهی اولیه مناسب برای تجزیه‌یNMF انجام گرفته است. نتایج این مطالعات نشان داده است که روش تجزیه‌ی مقدار تکین مضاعف نامنفی[۲۰] نسبت به روش‌های دیگر دارای نتایج مطلوب‌تری است. در این رساله، تجزیه‌ی ماتریس نامنفی با مقداردهی اولیه‌ی تصادفی و تجزیه‌ی مقدار تکین مضاعف نامنفی مطالعه و روش تحلیل مؤلفه اصلی به منظور مقداردهی اولیه تجزیه‌ی NMFپیشنهاد و بررسی خواهد شد. پس از کاهش بعد داده‌های ریزآرایه، از روش k-متوسط برای خوشه‌بندی داده‌ها استفاده می‌شود.

اهمیت و ضرورت تحقیق

امروزه بیماری‌هایی که از طریق ژن منتقل می‌شوند و از جمله آن‌ها انواع سرطان‌ها و بیماری‌های قلبی، یکی از عوامل مهم مرگ و میر است. به‌همین دلیل حجم عظیمی از مطالعات پزشکی جهت شناسایی، درمان و پیشگیری از این بیماری‌ها می‌باشد و برای این منظور ضروری است که عوامل بیماری شناسایی و طبقه‌بندی شود تا مراحل درمانی مناسب با نوع خاص بیماری انجام شود. فناوری ریزآرایه و روش‌های داده‌کاوی کمک می‌کند تا با تحلیل تغییرات بیان هزاران ژن، به صورت همزمان بتوان فرایند ایجاد بیماری‌هایی مثل سرطان را شناسایی کرد و در درمان این بیماری‌ها گام‌های مهمی برداشت.

یکی از عوامل بسیار مهم سرطان، جهش‌های ژنتیکی است. این جهش‌ها بسیار شبیه به هم هستند و نیاز به آزمایش‌های دقیق جهت شناسایی آن‌ها از یکدیگر وجود می‌باشد. با شناسایی کامل ژنوم انسانی و ژن‌هایی که در این جهش‌ها موثر هستند، انتظار می‌رود تشخیص بین این جهش‌ها با استفاده از روش‌هایی بر مبنای بیان ژن، امکان‌پذیر باشد. محققان تلاش می‌کنند تا با خوشه‌بندی بیماری‌های ژنتیکی بر اساس رفتارها و عوامل ژنتیکی در زیر گروه‌های همگن، فرایند تشخیص و درمان بیماری‌ها را سرعت بخشند.

اهداف اصلی تحقیق

اهداف اصلی این پژوهش عبارتند از:

۱. مطالعه‌ی تجزیه‌ی ماتریس نامنفی

۲. مطالعه و بررسی روش تحلیل مؤلفه اصلی جهت مقداردهی اولیه‌ی تجزیه‌ی ماتریس نامنفی

[۲۰] Nonnegative double singular value decomposition(NNDSVD)

۳. مطالعه و بررسی روش تجزیه‌ی مقدار تکین مضاعف نامنفی جهت مقداردهی اولیه‌ی تجزیه‌ی ماتریس نامنفی

۴. تحلیل نتایج خوشه‌بندی داده‌های حاصل از روش تجزیه‌ی ماتریس نامنفی

ساختار تحقیق

در این تحقیق، فصل اول مقدمات و کلیات موضوع را ارائه داده است. فصل دوم شامل تعاریف و پیش زمینه‌های لازم از جمله بیوانفورماتیک، پیش زمینه‌های زیستی، مفاهیم داده‌کاوی و تعاریف ریاضی می‌باشد. فصل سوم روش‌های تحلیل مؤلفه اصلی و تجزیه‌ی ماتریس نامنفی را جهت کاهش بعد داده‌ها و روش تجزیه‌ی مقدار تکین مضاعف نامنفی را جهت مقداردهی اولیه‌ی روش تجزیه‌ی ماتریس نامنفی شرح می‌دهد. در فصل چهارم خوشه‌بندی و روش‌های آن توضیح و روش خوشه‌بندی k-متوسط از بین این روش‌های شرح داده می‌شود. فصل پنجم ابتدا به آماده‌سازی و پیش‌پردازش داده‌ها پرداخته است و سپس نتایج کاهش بعد داده‌ها را با استفاده از هر یک از الگوریتم‌های معرفی شده در پایان‌نامه مورد بررسی و مقایسه قرار می‌دهد و نتایج خوشه‌بندی داده‌های بیان ژن یکی از پایگاه‌داده‌ها را مورد تحلیل قرار می‌دهد.

فصل ۲

تعاریف و پیش‌نیازها