دانشگاه صنعتی اصفهان

دانشکده مهندسی برق و کامپیوتر

شناسایی ژن‌های موثر در بروز بیماری با استفاده از داده‌کاوی ریزآرایه و آنتولوژی ژن

پایان‌نامه کارشناسی ارشد هوش مصنوعی و رباتیک

آزاده محمدی

استاد راهنما

دکتر محمد حسین سرایی

۱۳۸۸

دانشگاه صنعتی اصفهان
دانشکده مهندسی برق و کامپیوتر

پایان نامه کارشناسی ارشد رشته هوش مصنوعی و رباتیک خانم آزاده محمدی

تحت عنوان

شناسایی ژن‌های موثر در بروز بیماری با استفاده از داده‌کاوی ریزآرایه و آنتولوژی ژن

در تاریخ ۱۳۸۸/۲/۵ توسط کمیته تخصصی زیر مورد بررسی و تصویب نهایی قرار گرفت.

۱- استاد راهنمای پایان نامه — دکتر محمد حسین سرایی

۲- استاد مشاور پایان نامه — دکتر منصور صالحی

سرپرست تحصیلات تکمیلی دانشکده — دکتر علی محمد دوست حسینی

سپاس پروردگار، که هرچه دارم همه از لطف اوست

بر خود لازم می دانم از استاد ارجمند جناب آقای دکتر سرایی که در تمامی مراحل پایان نامه راهنما و مشوق من بودند، تشکر و قدردانی نمایم، همچنین از جناب آقای دکتر صالحی به خاطر رهنمودهای ارزنده و حمایت های دلسوزانه شان بسیار سپاسگزارم.

از جناب آقای دکتر میرلوحی و سرکار خانم مهندس آقایی که زحمت داوری این پایان نامه را تقبل فرمودند و نیز جناب آقای دکتر دوست حسینی رئیس محترم تحصیلات تکمیلی سپاسگزاری می نمایم.

همچنین در این مجال از تمامی اساتید بزرگوار دانشکده برق و کامپیوتر دانشگاه صنعتی اصفهان که افتخار شاگردی آنها را داشته ام، تشکر و قدردانی می نمایم.

در نهایت از خانواده عزیزم تشکر می کنم که در همه دوران زندگی پشتیبان و همراه من بوده اند و همواره مدیون زحمات و لطف بی دریغ آنها هستم.

آزاده محمدی

اردیبهشت ۱۳۸۸

تقدیم به برترین هدیه‌های الهی

پدر عزیز و مادر مهربانم

فهرست مطالب

عنوان **صفحه**

چکیده

امروزه استفاده از اطلاعات ژنتیکی افراد در تشخیص و رده‌بندی انواع بیماری‌ها از جمله سرطان‌ها، مورد توجه قرار گرفته است. یکی از بهترین و دقیق‌ترین روش‌ها در این زمینه، بررسی مقادیر بیان ژنی در افراد مختلف توسط فناوری ریزآرایه می‌باشد. یکی از مشکلات داده‌های ریزآرایه کم بودن تعداد نمونه‌ها در مقایسه با تعداد ژن‌ها است. این مسئله سبب کاهش دقت رده‌بندی و افزایش هزینه‌های محاسباتی و آزمایشگاهی می‌شود، در عین حال بسیاری از این ژن‌ها در ایجاد بیماری مورد بررسی نقشی ندارند، در نتیجه تشخیص و انتخاب ژن‌های موثر در بروز بیماری علاوه بر آنکه سبب افزایش دقت رده‌بندی وکاهش هزینه‌ها می‌شود، از نظر زیستی نیز از اهمیت ویژه‌ای برخوردار است و می‌تواند اطلاعات مفیدی درباره علل و نحوه درمان بیماری‌ها در اختیار محققین قرار دهد. تشخیص و انتخاب ژن‌های موثر در بروز بیماری، از میان هزاران ژن مورد بررسی در آزمایش ریزآرایه، انتخاب ژن نام دارد.

در این پایان‌نامه با بررسی روش‌های مختلف انتخاب ژن، تلاش شده است با بهره‌گیری از مزایای روش‌های موجود، چارچوب جدیدی برای انتخاب ژن‌های موثر در بروز بیماری ارائه شود، به‌نحوی که نقاط ضعف روش‌های متداول پوشش داده شوند. در روش پیشنهادی، علاوه بر داده‌های بیان ژنی از یکی دیگر از منابع معتبر موجود درباره ژن‌ها یعنی آنتولوژی ژن نیز کمک گرفته شده است. استفاده از آنتولوژی ژن در کنار مجموعه داده‌های بیان ژنی تا حدی می‌تواند محدودیت‌های ریزآرایه یعنی کم بودن تعداد نمونه‌ها و خطای احتمالی در مقادیر اندازه‌گیری شده را جبران نماید. در چارچوب ارائه شده ابتدا بخش عمده‌ای از ژن‌های غیرمرتبط با کمک روش فیلتری (فیشر) حذف می‌شوند، اما روش‌های فیلتری همبستگی موجود بین ژن‌ها را مدنظر قرار نمی‌دهند در نتیجه ژن‌های باقیمانده دارای حجم بالایی از افزونگی می‌باشند. به‌منظور کاهش افزونگی در ژن‌های باقیمانده، یک رویکرد حریصانه برای حذف ژن‌های مشابه پیشنهاد شده است. در این رویکرد میزان مشابهت ژن‌ها با در نظر گرفتن اطلاعات آنتولوژی ژن و داده‌های بیان ژنی و بر اساس یک معیار تلفیقی محاسبه می‌شود و سپس بر اساس این معیار، ژن‌های افزونه از مجموعه ژن‌ها حذف می‌شوند. در نهایت ژن‌های باقیمانده از این مرحله، به‌عنوان ژن‌های کاندید به‌طور دقیق‌تر توسط روش SVMRFE مورد بررسی قرار می‌گیرند تا مجموعه ژن‌های نشانگر بیماری بدست آید. روش پیشنهادی بر روی دو مجموعه داده سرطان DLBCL و سرطان کلون اعمال شده است. نتایج بدست آمده نمایانگر تاثیر مثبت روش پیشنهادی بر کارایی رده‌بندی است، به‌علاوه مقایسه این روش با روش‌های انتخاب ژن متداول، نشان می‌دهد که روش ارائه شده به ازای تعداد ژن‌های مساوی، از کارایی بهتری برخوردار است.

همچنین از آنجایی که بسیاری از مجموعه داده‌های ریزآرایه به دلایل مختلف از جمله وجود خراش یا گرد و غبار بر روی اسلاید، بروز خطا در حین آزمایش، اختلال در تصویر ریزآرایه و پایین بودن قدرت تفکیکی تصاویر، شامل مقادیر گمشده می‌باشند در این پایان‌نامه با استفاده از تلفیق روش خوشه‌بندی CST و آنتولوژی ژن روش نوینی برای تخمین مقادیر گمشده در مرحله پیش‌پردازش ارائه گردیده است. عملکرد روش پیشنهادی بر روی مجموعه داده سرطان DLBCL و به ازای درصدهای مختلفی از مقادیر گمشده مورد بررسی قرار گرفته است. مقایسه نتایج حاصل از روش پیشنهادی با نتایج سایر روش‌های تخمین مقادیر گمشده، نشان می‌دهد که روش پیشنهادی می‌تواند مقادیر گمشده را با دقت بالاتری تخمین بزند.

واژه‌های کلیدی: ۱- انتخاب ژن ۲- آنتولوژی ژن ۳- بیان ژن ٤- ریزآرایه ۵- مقدار گمشده

فصل اول

مقدمه

۱ ـ ۱ داده‌کاوی زیستی[1]

در سال‌های اخیر، با ظهور فناوری‌های زیستی پیشرفته، حجم داده‌های زیستی به‌طور قابل‌ملاحظه‌ای افزایش یافته است. ذخیره‌سازی و تحلیل این حجم از داده‌ها بدون استفاده از کامپیوتر، پیچیده یا تقریباً غیرممکن است. این مسئله منجر به ایجاد یک مبحث بین رشته‌ای به نام بیوانفورماتیک[2] شده است. بیوانفورماتیک علم نوینی است که با تلفیق علوم زیست‌شناسی، کامپیوتر و ریاضیات (به‌ویژه آمار)، تلاش می‌کند به مسائل زیستی در زمینه‌های سلولی و مولکولی پاسخ دهد [۱]. به‌طور کلی می‌توان مباحث بیوانفورماتیک را به سه دسته عمده ژنومیک[3]، پروتئومیک[4] و ترانسکریپتومیک[5] تقسیم نمود.

ژنومیک شامل تجزیه و تحلیل داده‌های ژنتیکی است. مجموعه اطلاعات ژنتیکی هر موجود زنده ژنوم[6] نامیده می‌شود. اطلاعات ژنتیکی کد شده در ژنوم، خصوصیات موجودات زنده را تعیین می‌کند و از یک نسل به نسل بعدی

---

[1] Biological data mining
[2] Bioinformatics
[3] Genomics
[4] Proteomics
[5] Transcriptomics
[6] Genome

منتقل می‌شود. عناصر وراثتی ژن[1] نامیده می‌شوند. به‌طور خلاصه می‌توان گفت که ژنومیک شامل توالی‌یابی[2] و تحلیل عملکرد ژن‌ها و مقایسه ساختمان ژنی در موجودات مختلف است. تحلیل توالی DNA[3] می‌تواند برای مقایسه اعضای مختلف از یک گونه جانداران یا به‌منظور مقایسه گونه‌های مختلف جانداران مورد استفاده قرار گیرد، همچنین تحلیل توالی‌ها می‌تواند تصویر واضح‌تری از نحوه تکامل موجودات زنده و چگونگی ارتباط آنها با یکدیگر فراهم آورد. پروژه ژنوم انسان[4] که از سال ۱۹۹۶ تا سال ۲۰۰۳ به طول انجامید، نمونه‌ای از تحلیل توالی است، در این پروژه، کل ژنوم انسان تعیین توالی گردید و درون یک پایگاه داده قرار گرفت [۲].

رفتار سلولی و تمام فعالیت‌هایی که در سلول انجام می‌شود بر عهده پروتئین‌ها است، بنابراین برای توجیه رفتار سلولی، باید ساختار و عملکرد پروتئین‌ها شناسایی شود. پروتئین‌های هر موجود بر اساس اطلاعات ژنتیکی کد شده در ژنوم او، ساخته می‌شوند. به مجموعه پروتئین‌های یک موجود، پروتئوم[5] گفته می‌شود. پروتئومیک شاخه‌ای از علم بیوانفورماتیک است که ساختار و عملکرد پروتئین‌ها را بررسی می‌کند و داده‌های پروتئینی را مورد تجزیه و تحلیل قرار می‌دهد. در این شاخه از علم بیوانفورماتیک، پیش‌بینی ساختار پروتئین‌ها، تعیین عملکرد پروتئین‌های ناشناخته و نحوه تعامل پروتئین‌های مختلف با یکدیگر، مورد بررسی قرار می‌گیرد [۳].

ترانسکریپتوم[6] مجموعه کامل دستورات لازم برای ساخت پروتئین‌های مختلف در یک سلول یا موجود است. همان‌طور که بیان شد ژنوم مجموعه کامل ژن‌های یک موجود و پروتئوم مجموعه پروتئین‌های قابل‌تولید از ژنوم است. ترانسکریپتوم همانند یک پل ارتباطی بین ژنوم و پروتئوم است و بررسی آن مشخص می‌نماید که در هر مرحله از رشد یا تحت شرایط مختلف، کدام ژن‌ها و چگونه برای ساخت پروتئین‌های لازم، فعال شده‌اند. تحلیل این دسته از داده‌ها، ترانسکریپتومیک نامیده می‌شود. در این شاخه از علم بیوانفورماتیک تحلیل الگوهای بیان ژن[7]، چگونگی نمو و تمایز سلول‌ها و سازگاری با محیط و شرایط متغیر مورد بررسی قرار می‌گیرد. با ظهور فناوری ریزآرایه[8]، این شاخه از علم بیوانفورماتیک اهمیت زیادی پیدا کرده است. برخلاف ژنوم که تنها دربردارنده اطلاعات ایستا درباره توالی ژن‌ها است، آزمایش‌های ریزآرایه با اندازه‌گیری مقادیر بیان ژنی، اطلاعات پویایی درباره عملکرد سلولی ارائه می‌کنند [۳].

---

[1] Gene
[2] Sequence alignment
[3] Deoxyribo Nucleic Acid
[4] Human genome project
[5] Proteome
[6] Transcriptome
[7] Gene expression
[8] Microarray

به طور کلی بیوانفورماتیک سه هدف عمده را دنبال می‌کند. هدف اول عبارت است از سازمان‌دهی داده‌ها به‌نحوی که محققان بتوانند به اطلاعات موجود دسترسی یافته و داده‌های جدید را به این اطلاعات اضافه نمایند. هدف دوم، توسعه ابزار و منابعی است که به تحلیل داده‌ها کمک نمایند. توسعه این‌گونه ابزارها مستلزم مهارت در تئوری محاسبات و زیست‌شناسی است. هدف سوم، استفاده از این ابزارها برای تحلیل داده‌ها و تفسیر نتایج به‌صورتی است که از نظر زیستی بامعنی باشد [۱]. منظور از داده‌ها در بحث بیوانفورماتیک داده‌های زیستی مانند توالی ژنوم، داده‌های بیان ژن و توالی اسیدهای آمینه در پروتئین‌ها، می‌باشد.

به‌طور کلی بیوانفورماتیک می‌تواند به‌صورت کاربرد فناوری کامپیوتر در مدیریت داده‌های زیستی، تعریف شود [۳]. بیوانفورماتیک به محققین اجازه می‌دهد از پیشرفت‌های علوم کامپیوتر و آمار در تحلیل داده‌های زیستی استفاده کنند؛ با این‌حال هر چه حجم داده‌ها بیشتر می‌شود، استفاده از روش‌های پیچیده‌تری برای مدیریت داده‌ها، ضرورت پیدا می‌کند، بنابراین در حال حاضر هدف اصلی بیوانفورماتیک استخراج اطلاعات نهفته در این حجم انبوه از داده‌ها است. کشف این اطلاعات، می‌تواند درک روشن‌تری از پدیده‌های زیستی برای محققین فراهم آورده و منجر به ایجاد فرضیات جدیدی درباره نحوه عملکرد ژن‌ها و پروتئین‌ها، دلایل بروز بیماری‌ها و نحوه درمان آنها شود. استخراج این اطلاعات توسط روش‌های هوشمند کامپیوتری از جمله داده‌کاوی[1] میسر می‌شود.

داده‌کاوی عبارت است از استخراج اطلاعات ناشناخته و مخفی در داده‌ها که به‌طور بالقوه مفید می‌باشند. بنابراین هدف اصلی داده‌کاوی کشف اطلاعات نهفته در داده‌ها است. پژوهش جدی روی موضوع داده‌کاوی از اوایل دهه ۹۰ شروع شد [۲]. چهار مرحله اساسی در فرایند داده‌کاوی وجود دارد که عبارتند از: جمع‌آوری داده‌ها، پیش‌پردازش[2] داده‌ها، اعمال روش‌های داده‌کاوی و تفسیر اطلاعات. در مرحله اول، داده‌ها از منابع مختلف جمع‌آوری می‌شوند. داده‌های جمع‌آوری شده معمولاً دارای داده‌هایی با مقادیر نادرست یا نامعلوم (مقادیر گمشده[3]) می‌باشند، در نتیجه پیش از اعمال روش‌های داده‌کاوی بر روی داده‌ها، داده‌های خام باید پیش‌پردازش شوند تا مشکلات گفته شده مرتفع شود. پس از پیش‌پردازش داده‌ها، با اعمال روش‌های داده‌کاوی، الگوهای مخفی در داده‌ها شناسایی می‌شوند. در این مرحله الگوریتم‌های مختلفی می‌توانند مورد استفاده قرار گیرند. انتخاب روش مناسب، با توجه به نوع مسئله صورت می‌گیرد. در نهایت اطلاعات و الگوهای استخراج شده بر اساس معیارهای مختلف ارزیابی می‌شوند، همچنین این اطلاعات می‌تواند برای تحلیل و تفسیر، در اختیار افراد خبره در زمینه مرتبط، قرار گیرد. بررسی اطلاعات بدست‌آمده از داده‌کاوی توسط افراد خبره، از اهمیت ویژه‌ای برخوردار است. این کار

---

[1] Data mining
[2] Preprocessing
[3] Missing value

اگرچه زمان‌بر است اما اطمینان می‌دهد اطلاعات استخراج شده دقیق و سودمند هستند، به‌علاوه تحلیل و تفسیر صحیح نتایج را امکان‌پذیر می‌سازد.

امروزه استفاده از روش‌های داده‌کاوی در علم بیوانفورماتیک به سرعت در حال افزایش است. حجم داده‌های زیستی بدست آمده از آزمایش‌ها و نیز وجود بانک‌های اطلاعاتی عظیم از این نوع داده‌ها، منجر به ظهور زمینه تحقیقاتی جدیدی به نام داده‌کاوی زیستی شده است که هدف آن پردازش و تحلیل داده‌های زیستی و استخراج دانش نهفته در این داده‌ها می‌باشد [۳].

۱ ۴ اهمیت و تاریخچه تحلیل ریزآرایه

یکی از مهم‌ترین زمینه‌های تحقیقاتی پزشکی، شناسایی عوامل موثر در بروز بیماری است. تعیین این عوامل، می‌تواند روند تشخیص و درمان بیماری‌ها را بهبود بخشد. بسیاری از بیماری‌ها منشا ژنتیکی دارند، به‌عنوان مثال سلول‌های نرمال ممکن است در اثر جهش‌های ژنتیکی، به سلولهای سرطانی بدخیم تبدیل شوند. این تغییرات، مقادیر بیان ژن‌ها را تحت تاثیر قرار می‌دهند. بیان ژن، فرایندی است که طی آن واحدهای وراثتی یعنی ژن‌ها به واحدهای عملیاتی یعنی پروتئین‌ها تبدیل می‌شوند [۴].

یکی از فناوری‌های جدید برای اندازه‌گیری سطح بیان ژنی، ریزآرایه می‌باشد. این فناوری امکان ارزیابی تعداد بسیار زیادی از ژن‌ها را به‌طور هم‌زمان فراهم می‌کند [۵]. داده‌های حاصل از آزمایش‌های ریزآرایه می‌توانند اطلاعات ارزشمندی درباره نحوه عملکرد سلولی، در اختیار محققین قرار دهند. تحلیل داده‌های ریزآرایه یکی از مسائل مطرح در زمینه بیوانفورماتیک است که در زیرشاخه ترانسکریپتومیک قرار می‌گیرد. هدف از تحلیل داده‌های ریزآرایه بررسی تغییرات بیان ژنی در شرایط مختلف است. استخراج اطلاعات از این حجم انبوه داده‌ها با استفاده از روش‌های داده‌کاوی امکان‌پذیر می‌شود، از جمله استفاده از روش‌های خوشه‌بندی[1] و رده‌بندی[2] در تحلیل داده‌های ریزآرایه بسیار مورد توجه می‌باشد.

هدف از خوشه‌بندی داده‌های بیان ژن، تعیین ژن‌ها با الگوهای بیانی مشابه و تشخیص الگوهای مخفی درون داده‌ها است. خوشه‌بندی برای تعیین ژن‌های مشابه، پیش‌بینی عملکرد ژن‌های ناشناخته و کاهش پیچیدگی مورد استفاده قرار می‌گیرد [۶]. با توجه به اهمیت شناسایی الگوهای بیان ژنی، تا کنون روش‌های متفاوتی برای خوشه‌بندی داده‌های بیان ژنی مورد استفاده قرار گرفته‌اند که از آن جمله می‌توان به خوشه‌بندی سلسله مراتبی[3] [۷, ۸] خوشه‌بندی

---

[1] Clustering
[2] Classification
[3] Hierarchical

K میانگین[1] [۶] و نقشه خودسازمان‌ده[2] [9]اشاره نمود. علاوه بر روش‌های مرسوم خوشه‌بندی، الگوریتم‌هایی برای خوشه‌بندی داده‌های بیان ژنی به‌طور خاص ارائه شده‌اند که از آن جمله می‌توان به الگوریتم CAST[3] [10] و CST[4] [۱۱] اشاره نمود. این روش‌ها در برابر نویز[5] بهتر عمل کرده و قادر به شناسایی خودکار تعداد خوشه‌ها می‌باشند. به‌طور کلی بررسی‌ها نشان داده‌اند که الگوریتم‌های خوشه‌بندی که به‌طور خاص برای داده‌های بیان ژنی طراحی شده‌اند، در اکثر موارد نسبت به الگوریتم‌های متداول خوشه‌بندی مانند خوشه‌بندی K میانگین، SOM و غیره موفق‌تر عمل می‌نمایند [۶].

یکی دیگر از کاربردهای داده‌کاوی در ریزآرایه، رده‌بندی نمونه‌ها بر اساس مقادیر بیان ژنی است. با داده‌کاوی ریزآرایه، می‌توان الگوهایی جهت تشخیص بیماری‌ها استخراج نمود. تشخیص زودهنگام بیماری باعث افزایش احتمال بهبود و کاهش هزینه‌های درمانی می‌گردد، به‌علاوه تشخیص بیماری با استفاده از الگوهای استخراج شده از داده‌های ریزآرایه دارای دقت بالایی است. بسیاری از بیماری‌ها علائم ظاهری و بالینی مشابهی را بروز می‌دهند، در حالی‌که ممکن است در سطح سلولی متفاوت بوده و نیاز به رویکردهای درمانی مختلفی داشته باشند. با بررسی بیماری در سطح سلولی می‌توان گونه‌های مختلف یک بیماری را تفکیک و رده‌بندی نمود. تا کنون روش‌های مختلفی برای رده‌بندی داده‌های بیان ژنی مورد استفاده قرار گرفته است که از آن جمله می‌توان به روش K نزدیک‌ترین همسایه[6] [۱۲]، شبکه‌های عصبی مصنوعی[7] [۱۳]، درخت‌های تصمیم[8] [۳] و ماشین‌های بردار پشتیبان[9] [۱۴, ۱۵] اشاره کرد.

همان‌گونه که بیان شد داده‌های بیان ژنی می‌توانند در تشخیص بیماری‌ها مورد استفاده قرار گیرند، اما مشکل اساسی در مورد این داده‌ها، محدود بودن تعداد نمونه‌های مورد آزمایش در مقایسه با تعداد ژن‌ها است. بسیاری از این ژن‌ها در ایجاد بیماری مورد بررسی نقشی ندارند، از این‌رو تشخیص ژن‌های موثر در بروز بیماری می‌تواند علت و نحوه ایجاد بیماری را مشخص نموده و پزشکان و داروسازان را در اتخاذ روش‌های درمانی مناسب و طراحی داروهای مفید، یاری کند. با تعیین ژن‌های موثر در بروز بیماری، آزمایش‌های تشخیصی و رده‌بندی بیماران، می‌تواند

---

[1] K-means
[2] Self Organizing Map
[3] Cluster Affinity Search Technique
[4] Correlation Search Technique
[5] Noise
[6] K Nearest Neighbor
[7] Artificial Neural Network
[8] Decision tree
[9] Support Vector Machines

تنها با استفاده از این ژن‌ها صورت گیرد؛ این امر سبب کاهش هزینه‌های آزمایشگاهی و افزایش دقت می‌شود. تشخیص و انتخاب زیرمجموعه‌ای از ژن‌ها به‌عنوان عوامل موثر در بروز بیماری، انتخاب ژن نامیده می‌شود.

تا کنون روش‌های مختلفی برای تعیین ژن‌های موثر در بروز بیماری، ارائه شده‌اند. روش‌های انتخاب ژن به‌طور کلی در سه دسته فیلتری[۱]، پوششی[۲] و ادغامی[۳] قرار می‌گیرند [۱۶]. در روش‌های فیلتری، انتخاب ژن مستقل از رده‌بندی انجام می‌شود. این روش‌ها معمولاً بر پایه روش‌های آماری استوار هستند، بدین ترتیب که قدرت هر ژن در تفکیک نمونه‌ها بر اساس یک معیار آماری محاسبه می‌شود و سپس ژن‌هایی که بر اساس معیار محاسبه شده قدرت تفکیک بهتری دارند، به‌عنوان مجموعه‌ژن‌های موثر در بروز بیماری انتخاب می‌شوند. از جمله این روش‌ها می‌توان به انتخاب ژن با استفاده از معیار SNR[۴] [۱۷]، انتخاب ژن با استفاده از معیار فیشر[۵] [۱۸] و انتخاب ژن با استفاده از معیار مجموع رتبه ویلکوکسون[۶] [۱۹] اشاره کرد. این روش‌ها ساده وسریع هستند اما همبستگی بین ژن‌ها را درنظر نمی‌گیرند و مجموعه ژن‌های انتخابی دارای حجم بالایی از افزونگی[۷] می‌باشند.

در روش‌های انتخاب ژن پوششی، فضای ویژگی[۸] (ژن‌ها) برای یافتن مجموعه ژن‌های موثر در بروز بیماری جستجو می‌شود. در این روش‌ها با استفاده از یک مکانیزم جستجو، در هر مرحله یک زیرمجموعه ژن انتخاب شده و کیفیت آن بر اساس کارایی رده‌بند[۹]، مورد ارزیابی قرار می‌گیرد. زیرمجموعه‌ای از ژن‌ها که بالاترین کارایی را در رده‌بند ایجاد نماید به‌عنوان مجموعه ژن‌های متمایزکننده انتخاب می‌شود. به‌عنوان مثال در [۲۰] از روش انتخاب رو به جلو[۱۰] برای انتخاب ژن‌های موثر در بیماری استفاده شده است. در [۲۱, ۲۲] نیز، از الگوریتم ژنتیک[۱۱] برای انتخاب ژن استفاده شده است. این روش‌ها تعامل ژن‌ها را درنظر می‌گیرند اما پیچیدگی محاسباتی بالایی دارند.

دسته دیگر از روش‌های انتخاب ژن روش‌های ادغامی هستند که در آنها مرحله انتخاب ژن با مرحله رده‌بندی ادغام می‌شود. روش SVMRFE[۱۲] یکی از موفق‌ترین روش‌های انتخاب ژن است که توسط گیون[۱۳] و همکاران ارائه شده است [۲۳]. در این روش در هر مرحله کم‌اهمیت‌ترین ژن‌ها در تفکیک نمونه‌ها، به‌صورت بازگشتی حذف

---

[1] Filter
[2] Wrapper
[3] Embedded
[4] Signal to Noise Ratio
[5] Fisher
[6] Wilcoxon rank sum test
[7] Redundancy
[8] Feature sapce
[9] Classifier
[10] Forward selection
[11] Genetic algorithm
[12] Support Vector Machine Recursive Feature Elimination
[13] Guyon

می‌شوند. تعیین اهمیت ژن‌ها در این روش بر اساس ماشین بردار پشتیبان صورت می‌گیرد. این روش تعامل بین ژن‌ها را درنظر می‌گیرد و در عین حال پیچیدگی محاسباتی آن نسبت به روش‌های پوششی کمتر است.

هدف از انتخاب ژن، حذف ویژگی‌های غیرمرتبط[۱] و افزونه[۲] می‌باشد. ویژگی‌های غیرمرتبط ویژگی‌هایی هستند که در تفکیک نمونه‌ها نقش چندانی ندارند و بدون تاثیرگذاشتن بر کارایی یادگیری می‌توان آنها را از مجموعه ژن‌ها حذف کرد. ویژگی افزونه خصیصه‌ای است که به‌تنهایی مرتبط با تفکیک رده‌ها است اما وجود یک ویژگی دیگر در مجموعه ژن‌ها سبب می‌شود که حذف این ویژگی تغییری در کارایی یادگیری ایجاد نکند. اکثر روش‌های مطرح شده در انتخاب ژن به‌دنبال حذف ویژگی‌های غیرمرتبط هستند اما اکثر روش‌ها بخصوص روش‌های فیلتری افزونگی ژن‌ها را درنظر نمی‌گیرند. وجود افزونگی در مجموعه ژن‌های انتخابی علاوه بر آن‌که سبب افزایش هزینه محاسباتی می‌شود، کارایی رده‌بندی را نیز کاهش می‌دهد، همچنین افزونگی ژن‌ها سبب انتخاب ژن‌هایی با عملکرد مشابه می‌شود. این ژن‌ها حاوی اطلاعات مشابهی هستند، در نتیجه انتخاب ژن‌های افزونه، اطلاعات بیشتری را فراهم نمی‌کند [۲۴].

تا کنون روش‌های مختلفی برای کاهش افزونگی در مجموعه ژن‌های انتخابی پیشنهاد شده است. از جمله در [۲۵, ۲۶] استفاده از روش‌های خوشه‌بندی برای تعیین ژن‌های مشابه و کاهش افزونگی پیشنهاد شده است. در [۲۵] پیش از اعمال روش‌های رتبه‌بندی، ژن‌های مشابه با استفاده از خوشه‌بندی فازی[۳] تعیین و حذف می‌شوند و رتبه‌بندی ژن‌ها تنها بر روی ژن‌های باقیمانده انجام می‌شود، اما در [۲۶]، خوشه‌بندی سلسله‌مراتبی پس از انتخاب ژن توسط روش‌های فیلتری رتبه‌بندی اعمال شده است تا افزونگی موجود در ژن‌های انتخابی را کاهش دهد. در [۲۷] از روش مارکوف بلانکت[۴] برای تشخیص و کاهش افزونگی ژن‌ها استفاده شده است. دینگ[۵] و پنگ[۶] در [۲۸] روشی به نام MRMR[۷] برای انتخاب ژن ارائه داده‌اند که هدف آن انتخاب مجموعه ژنی است که به‌طور هم‌زمان کمترین میزان افزونگی و بیشترین میزان ارتباط با رده[۸] مورد بررسی را داشته باشد.

خطاهای آزمایشگاهی در روند انجام آزمایش ریزآرایه سبب ایجاد مقادیر نادرست یا نامعلوم (مقادیر گمشده) در داده‌های بیان ژنی می‌شوند، لذا پیش از اعمال روش‌های داده‌کاوی، لازم است داده‌ها با استفاده از روش‌های پیش‌پردازش اصلاح شوند. مراحل پیش‌پردازش وابسته به خصوصیات مجموعه داده است. در بسیاری از مجموعه

---

[1] Irrelevant
[2] Redundant
[3] Fuzzy clustering
[4] Markov Blanket
[5] Ding
[6] Peng
[7] Minimum Redundancy Maximum Relevance
[8] Class

داده‌ها، مقادیر بیان ژنی در برخی از نمونه‌ها نامعلوم (گمشده) است. اکثر روش‌های داده‌کاوی برای عملکرد صحیح نیاز به مجموعه کاملی از داده‌ها دارند، از این‌رو مقادیر گمشده باید به شکل صحیح مدیریت شوند، همچنین به‌منظور حداقل کردن تاثیرات ناشی از خطای آزمایش، بر مقادیر بیان ژنی باید از روش‌های نرمال‌سازی استفاده نمود [۵].

تا کنون برای مقابله با مقادیر گمشده روش‌های مختلفی ارائه شده‌اند. از جمله در [۷] جایگزینی مقادیر گمشده با یک مقدار ثابت مانند صفر یا میانگین مقادیر ژن در نمونه‌ها پیشنهاد شده است. این روش‌ها ساده و سریع هستند اما به علت نادیده گرفتن همبستگی بین داده‌ها نتایج دقیقی تولید نمی‌کنند. ترویانسکایا[1] و همکاران در [۲۹] روش نسبت‌دهی K نزدیک‌ترین همسایه (KNNImpute[2]) را برای تخمین مقادیر گمشده پیشنهاد داده‌اند. در این روش مقادیر گمشده هر ژن با توجه به مقادیر K ژنی که بیشترین میزان مشابهت را با آن دارند، تخمین زده می‌شود. در [۳۰] روشی به نام نسبت‌دهی K نزدیک‌ترین همسایه متوالی (SKNNImpute[3])برای تخمین مقادیر گمشده ارائه شده است. این روش مشابه با روش K نزدیک‌ترین همسایه می‌باشد با این تفاوت که انتخاب نزدیک‌ترین همسایه‌ها برای هر ژن حاوی مقدار گمشده، از بین ژن‌های فاقد گمشدگی صورت می‌پذیرد و پس از تخمین مقادیر گمشده یک ژن، آن ژن نیز می‌تواند در مراحل بعدی در تخمین مقادیر گمشده سایر ژن‌ها مورد استفاده قرار گیرد. کیم[4] و همکاران در [۳۱] روشی به نام نسبت‌دهی حداقل مربعات محلی (LLSImpute[5]) ارائه داده‌اند که در آن مقادیر گمشده بر اساس ترکیب خطی K نزدیک‌ترین همسایه تخمین زده می‌شوند. در [۳۲, ۳۳] از روش‌های مبتنی بر خوشه‌بندی برای تخمین مقادیر گمشده استفاده شده است. در این روش‌ها، با استفاده از خوشه‌بندی، ژن‌های مشابه تعیین شده و سپس مقادیر گمشده هر ژن با استفاده از مقادیر متناظر در ژن‌های هم‌گروه آن تخمین زده می‌شوند.

مشکل عمده روش‌های فوق وابسته بودن آنها به پارامترهای ورودی مانند تعداد همسایه‌ها یا تعداد خوشه‌ها است. همچنین تخمین مقادیر گمشده در تمامی روش‌های مذکور تنها بر اساس اطلاعات بیان ژن صورت می‌گیرد، در حالی‌که به علت کمی نمونه‌ها و وجود خطا در داده‌های ریزآرایه تخمین مقادیر گمشده بر پایه این داده‌ها نمی‌تواند چندان دقیق باشد. در این حالت بهره‌گیری از سایر منابع اطلاعاتی موجود، می‌تواند سبب بهبود تخمین شود.

---

[1] Troyanaskaya
[2] K Nearest Neighbor Imputation
[3] Sequential KNNImpute
[4] Kim
[5] Local Least Squares Imputation

۱ ۳ روش پیشنهادی و اهداف پایان‌نامه

همان‌گونه که بیان شد، از داده‌های بیان ژنی می‌توان در تشخیص و رده‌بندی بسیاری از بیماری‌ها مانند سرطان استفاده کرد، اما این داده‌ها دو مشکل عمده دارند: یکی آن‌که داده‌های حاصل از آزمایش ریزآرایه در اثر خطاهای آزمایشگاهی ممکن است مقادیر غیرصحیح یا نامعلوم (مقادیر گمشده) داشته باشند. از این رو پیش‌پردازش داده‌های بیان ژنی به منظور رفع این مشکلات پیش از انجام رده‌بندی ضروری است. مشکل دیگر داده‌های بیان ژنی، بیشتر بودن تعداد ژن‌ها در مقایسه با تعداد نمونه‌ها است که سبب کاهش دقت رده‌بندی و افزایش هزینه‌های محاسباتی و آزمایشگاهی می‌شود. در این پایان‌نامه سعی می‌شود با توجه به نقاط ضعف و قوت روش‌های موجود، روش‌های جدیدی برای برخورد با این مسائل ارائه شود.

هر چند فناوری ریزآرایه حاوی اطلاعات بسیار ارزشمندی می‌باشد، اما به دلیل امکان بروز خطا در مراحل مختلف آزمایش و آماده‌سازی داده‌ها و نیز به علت کم بودن تعداد نمونه‌های آزمایشی، تصویر کاملی از نحوه عملکرد ژن‌ها ارائه نمی‌کند؛ از این رو بهره‌گیری از سایر منابع اطلاعاتی موجود درباره ژن‌ها، می‌تواند تا حدی محدودیت‌های ریزآرایه را جبران نماید. یکی از مهم‌ترین منابع اطلاعاتی موجود درباره ژن‌ها، آنتولوژی ژن[1] است که به شکل ساخت‌یافته عملکرد ژن‌ها و چگونگی ارتباط آنها را توصیف می‌کند. تلفیق این اطلاعات با داده‌های حاصل از آزمایش ریزآرایه می‌تواند نتایج قابل‌اطمینان‌تری تولید نماید، زیرا در این حالت تحلیل عملکرد ژن‌ها بر اساس منابع اطلاعاتی گسترده‌تری صورت می‌گیرد.

در این پایان‌نامه سعی می‌شود با تلفیق اطلاعات آنتولوژی ژن و داده‌های بیان ژن، مقادیر گمشده در داده‌های ریزآرایه تخمین زده شوند. در روش ارائه شده، با استفاده از خوشه‌بندی CST و نیز بهره‌گیری از اطلاعات آنتولوژی ژن، ژن‌های مشابه در قالب خوشه‌ها گروه‌بندی می‌شوند، سپس مقادیر گمشده هر ژن با توجه به مقادیر ژن‌های هم‌دسته آن تخمین زده می‌شوند. با این روش علاوه بر داده‌های حاصل از آزمایش ریزآرایه می‌توان از دانش موجود درباره ژن‌ها، که به صورت آنتولوژی ژن وجود دارد نیز بهره گرفت، به‌علاوه این روش نیازی به تعیین پارامتر توسط کاربر ندارد و تعداد ژن‌های مورد نیاز برای تخمین مقادیر هر ژن به‌طور خودکار تعیین می‌شود.

همچنین در این پایان‌نامه یک چارچوب جدید جهت شناسایی و انتخاب ژن‌های موثر در بروز بیماری ارائه می‌شود. در این روش ابتدا با استفاده از معیار فیشر بخشی از ژن‌ها حذف می‌شوند، سپس با بهره‌گیری از اطلاعات آنتولوژی ژن و نیز داده‌های بیان ژنی، همبستگی بین ژن‌ها بررسی می‌شود و با استفاده از رویکردی حریصانه[2]، ژن‌های افزونه از مجموعه ژن‌های باقیمانده حذف می‌شوند. در نهایت با استفاده از روش SVMRFE بهترین

---

[1] Gene Ontology
[2] Greedy

زیرمجموعه ژنی انتخاب می‌شود. این ژن‌ها به‌عنوان مجموعه ژن‌های موثر در بروز بیماری می‌توانند در رده‌بندی نمونه‌ها مورد استفاده قرار گیرند. کارایی روش پیشنهادی بر روی مجموعه داده‌های سرطان DLBCL[1] و کلون[2] مورد بررسی قرار گرفته است.

۱ ۴ روند ارائه مطالب

در فصل دوم، مقدمه مختصری در مورد علم ژنتیک و مفاهیم پایه زیست‌شناسی مولکولی بیان شده و سپس مفهوم بیان ژنی مطرح می‌گردد، در ادامه فناوری ریزآرایه، نحوه استخراج داده‌های بیان ژنی و کاربردهای آن شرح داده می‌شود. همچنین در این فصل آنتولوژی ژن به‌عنوان یک منبع اطلاعاتی مفید در تحلیل ژن‌ها، معرفی می‌شود.

در فصل سوم، مراحل تحلیل داده‌های ریزآرایه مورد بررسی قرار می‌گیرد. در این فصل ابتدا مروری بر روش‌های پیش‌پردازش داده‌های ریزآرایه مانند روش‌های تخمین مقادیر گمشده و نرمال‌سازی داده‌ها می‌شود، سپس انواع روش‌های مطرح در خوشه‌بندی و رده‌بندی داده‌های بیان ژنی بررسی می‌شوند.

در فصل چهارم، اهداف و دلایل انتخاب ژن مورد بررسی قرار گرفته و انواع روش‌های انتخاب ژن و مزایا و معایب هریک بررسی می‌شود، همچنین مشکلات ناشی از وجود افزونگی در مجموعه ژن‌های انتخابی، مطرح شده و تعدادی از مهم‌ترین روش‌های مقابله با افزونگی معرفی می‌شوند. در پایان این فصل چگونگی ارزیابی روش‌های انتخاب ژن مورد بررسی قرار می‌گیرد.

در فصل پنجم یک روش جدید برای تخمین مقادیر گمشده ارائه می‌شود. روش ارائه شده یک روش نسبت‌دهی مبتنی بر خوشه‌بندی است که در آن برای خوشه‌بندی ژن‌ها علاوه بر داده‌های بیان ژن از اطلاعات آنتولوژی ژن نیز کمک گرفته می‌شود. همچنین یک چارچوب جدید برای شناسایی و انتخاب ژن‌های موثر در بروز بیماری ارائه می‌شود. این چارچوب با ترکیب روش‌های فیلتری و ادغامی از مزایای هر دو روش بهره می‌گیرد، همچنین با تلفیق اطلاعات آنتولوژی ژن و داده‌های بیان ژنی، افزونگی در مجموعه ژن‌های انتخابی را کاهش می‌دهد.

نتایج پیاده‌سازی روش‌ها بر روی دو مجموعه داده سرطان DLBCL و سرطان کلون در فصل ششم ارائه می‌شود. در نهایت در فصل هفتم نتایج بدست‌آمده جمع‌بندی شده و پیشنهادهایی جهت ادامه کار ارائه می‌شود. لیست مقالات برگرفته شده از این پایان‌نامه در انتهای فصل هفتم آورده شده است.

---

[1] Diffuse Large B-Cell Lymphoma
[2] Colon