مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره پنجم شماره اول. ص ۳۴-۲۵، ۱۳۹۷

تشخیص نوع لوسمی به کمک یادگیری ماشین: کاهش ابعاد و متوازن‌سازی

زینب قرائتی[۱]، محمدرضا پژوهان[۲]*

• **دریافت مقاله:** ۹۶/۸/۲۵ • **پذیرش مقاله:** ۹۷/۲/۱۷

مقدمه: ترکیب تکنیک‌های محاسباتی هوش مصنوعی و داده‌کاوی در پزشکی به پیشرفت‌های قابل توجهی در پیش‌گیری و تشخیص بیماری‌ها منجر شده است. در تشخیص لوسمی حاد از اطلاعات ژنتیکی، مدل‌های پیچیده‌ای تاکنون ارائه شده؛ اما نتایج قابل توجهی را ارائه نکرده است. این مطالعه به تشخیص نوع سرطان خون با بررسی محدوده گسترده‌ای از توابع پارامتری و غیرپارامتری و به منظور افزایش قابلت تعمیم آن‌ها در یادگیری با استخراج ویژگی‌های ذاتی کم‌تر از نمونه‌ها می‌پردازد.

روش: این مطالعه توصیفی- تحلیلی، بر روی داده‌های Leukemia1 از دانشگاه واندربیلت آمریکا انجام شد. این داده‌ها مجموعه‌ای از نمونه‌های مغز استخوان و خون بیماران لوسمی است که برای طبقه‌بندی بر اساس سه زیر گروه سرطان خون ALL B-cell، ALL T-cell و AML استفاده می‌شود. دسته‌بندی پارامتری با الگوریتم‌های خطی، بیز ساده، فاصله اقلیدسی، نزدیک‌ترین میانگین، تطبیق قالب و دسته‌بندی غیرپارامتری با الگوریتم‌های تخمین‌گرهای پایه، هسته، k –همسایه نزدیک‌تر و k –همسایه نزدیک‌تر مبتنی برهسته انجام گردید.

نتایج: با در نظر گرفتن تمامی ویژگی‌ها بهترین الگوریتم نزدیک‌ترین میانگین بود که به دقت پیش‌بینی ۹۲/۸۶٪ رسید. با اعمال روش کاهش ویژگی PCA، باز هم بهترین نتیجه مربوط به الگوریتم نزدیک‌ترین میانگین بود و با متوسط تعداد ویژگی ۶/۸ به دقت ۹۶٪ دست یافت. در نهایت با متوازن‌سازی داده‌های Leukemia1، متوسط تعداد ویژگی و دقت توسط الگوریتم درجه ۲ به ترتیب ۵/۴۱ و ۹۸/۵۹ حاصل گردید.

نتیجه‌گیری: نتایج به دست آمده بیانگر اثربخشی استخراج ویژگی‌های ذاتی و متوازن‌سازی در بهبود دقت مدل مبتنی بر قاعده بیز و برتری آن نسبت به مدل‌های پیچیده‌تر کنونی می‌باشد.

کلید واژه‌ها: داده‌های ژنتیکی، تشخیص نوع سرطان خون، داده‌کاوی، متوازن‌سازی داده‌ها، کاهش ابعاد

• **ارجاع:** قرائتی زینب، پژوهان محمدرضا. تشخیص نوع لوسمی به کمک یادگیری ماشین: کاهش ابعاد و متوازن‌سازی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۷؛ ۵(۱): ۳۴-۲۵.

۱. دانشجوی کارشناسی ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران
۲. دکتری مهندسی کامپیوتر، استادیار، گروه مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران
* **نویسنده مسئول:** یزد، دانشگاه یزد، پردیس فنی و مهندسی، گروه مهندسی کامپیوتر
• **شماره تماس:** ۰۳۵-۳۱۲۳۲۳۵۸
• **Email:** Pajoohan@yazd.ac.ir

مقدمه

در حال حاضر حجم عظیمی از داده‌های پزشکی در اَشکال مختلف در دسترس است. این داده‌ها با روش‌های مختلف مورد تجزیه و تحلیل قرار می‌گیرند تا نتایجی را تولید کنند که در تصمیم‌گیری کارآمد و بهبود کیفیت خدمات به پزشک کمک کند. ترکیب تکنیک‌های محاسباتی هوش مصنوعی در پزشکی به پیشرفت‌های قابل توجهی در پیش‌گیری و تشخیص بیماری منجر شده است. سرطان، دومین علت اصلی مرگ‌ومیر در سراسر جهان بعد از بیماری‌های قلبی- عروقی است [۱]. یکی از اساسی‌ترین مشکلات در درمان بیماری سرطان، عدم وجود روشی مناسب در تشخیص زودهنگام آن می‌باشد. این در حالی است که اگر این بیماری به سرعت تشخیص داده شود، درمان بسیار موفق‌تری را به همراه خواهد داشت. از طرف دیگر به دلیل هزینه زیاد درمان سرطان در مراحل پیشرفته بیماری، سرمایه‌گذاری در امر تحقیقات سرطان از نظر اقتصادی نیز مقرون به صرفه است. این مطالعه با رویکرد یادگیری ماشین به تشخیص نوع سرطان خون که جهت درمان را مشخص می کند، پرداخته است.

چهارنوع اصلی سرطان خون وجود دارد که عبارت‌اند از [۲].

۱- لوسمی حاد لنفوئیدی (Acute Lymphoblastic Leukemia): این نوع لوسمی سلول‌های لنفاوی یا لنفوسیت‌ها را تحت تأثیر قرار می‌دهد و روندی حاد دارد.

۲- لوسمی حاد میلوئیدی (Acute Myeloid Leukemia): این نوع لوسمی سلول‌های مغز استخوان یا میلوسیت‌ها را تحت تأثیر قرار می‌دهد و روندی حاد دارد. در این بیماری، مغز استخوان میلوبلاست (نوعی گلبول سفید)، گلبول قرمز یا پلاکت‌های غیرطبیعی می‌سازد.

۳- لوسمی مزمن لنفوئیدی (Chronic Lymphocytic Leukemia): این نوع لوسمی سلول‌های لنفاوی یا لنفوسیت‌ها را تحت تأثیر قرار می‌دهد که بافت‌های لنفاوی را می‌سازند و روندی مزمن دارد.

٤- لوسمی مزمن میلوئیدی (Chronic Myeloid Leukemia): این نوع لوسمی سلول‌های مغز استخوان یا میلوسیت‌ها را تحت تأثیر قرار می‌دهد که بافت‌های مغز استخوان را می‌سازند و روندی مزمن دارد.

در سال‌های گذشته، پژوهش‌های زیادی در زمینه تشخیص نوع سرطان خون انجام شده است. به عنوان نمونه روش ABC (Artificial Bee Colony) برای کاهش ویژگی با طبقه‌بند kNN (k-Nearest Neighbor) [۳]، همچنین روش انتخاب ویژگی MFA score(Marginal Fisher Analysis score and redundancy excluding) با روش بازشناسی الگو SVM (Support Vector Machine) پیاده‌سازی شده است [۴]. روش RS (Relative Simplicity) استفاده شده که با طبقه‌بند DC (Direct Classifier) همراه بوده است [۵]. یا به عنوان مثالی دیگر، پس از انتخاب ویژگی توسط روش VIMs (Variable Importance Measures)، نمونه‌های موجود با ویژگی‌های حاصل، توسط جنگل تصادفی طبقه‌بندی شدند [۶].

در این مقاله، با توجه به اثر بخشی الگوریتم‌های یادگیری ماشین در بسیاری از پژوهش‌ها [۱۱-۷]، به بررسی عملکرد الگوریتم‌های پارامتریک و غیرپارامتریک درفرآیند تشخیص نوع سرطان خون پرداخته شد. همچنین اثر بخشی کاهش نویز در داده‌ها را با حذف ویژگی‌های نامناسب مورد ارزیابی قرار داده شد. از آنجا که میزان پیچیدگی داده‌ها مشخص نیست، هدف این پژوهش مشخص کردن مناسب‌ترین الگوریتم با توجه به پیچیدگی داده‌ها می‌باشد، به گونه‌ای که پدیده بیش‌برازش و کم برازش رخ ندهد.

روال کار به این صورت است که ابتدا تمام ویژگی‌ها توسط متخصص سرطان خون، برای تشخیص بهتر نوع سرطان مشخص می‌شود. معمولاً هرچه تعداد ویژگی‌ها بیشتر باشد فرضیه‌های یادگیری که می‌توانند باتوجه به نمونه‌های آموزشی مدل یادگیری را ایجاد کنند، به صورت نمایی افزایش می‌یابد؛ لذا ویژگی‌های کم‌تر سبب می‌شود تا استخراج فرضیه نهایی، که عملاً مدل یادگیری را شکل می‌دهد، شباهت بیشتری به مدل ذاتی داده‌ها داشته باشد؛ بنابراین با رویکرد کاهش ابعاد می‌توان انتظار داشت تا دقت تعمیم در یادگیری افزایش یابد. همچنین حذف ویژگی‌های نامناسب می‌تواند سبب کاهش نویز در داده‌ها گردد. در این پژوهش از (Principal Component Analysis) PCA که یک روش مؤثر در استخراج ویژگی است به منظور کاهش ابعاد استفاده شد [۱۲].

به طور خلاصه، نوآوری اصلی این پژوهش پیاده‌سازی مجموعه‌ای از روش‌های پارامتری و غیرپارامتری در تشخیص نوع سرطان خون، بررسی اثربخشی آن‌ها و بهبود نتایج توسط کاهش ابعاد و متوازن‌سازی داده‌ها می‌باشد. نتایج تجربی، بیانگر بهینگی روش دسته‌بندی نزدیک‌ترین میانگین و روش درجه دو از مجموعه روش‌های پارامتری در مجموعه داده‌های "Leukemia1"می‌باشد. با اعمال روش کاهش ویژگی

PCA و سپس اعمال الگوریتم‌های نزدیک‌ترین میانگین و درجه دو، دقت تشخیص نسبت به پژوهش‌های قبلی [۶-۳] بهبود یافت. در نهایت داده‌های نامتوازن انتخاب شده با روش Synthetic Minority Over-sampling Technique)SMOTE + TL(TomekLink) متوازن گردید که موجب افزایش قابل توجه دقت پیش‌بینی نوع سرطان خون شد.

با عنایت به اهمیت ویژه بیماری سرطان خون و تشخیص زودهنگام آن و با توجه به مزیت مضاعف روش‌های یادگیری ماشین، در این پژوهش برای مجموعه داده‌های "Leukemia1" یک مدل پیش‌بینی نوع سرطان خون، به وسیله روش‌های یادگیری ماشین، ارائه گردید. پیش‌بینی نوع سرطان خون از اهمیت بالایی برخوردار است چرا که عدم تشخیص زودهنگام و تشخیص اشتباه نوع آن موجب مرگ سریع بیمار می‌گردد. همچنین با تشخیص زودهنگام نوع سرطان خون بیمار اقدامات درمانی لازم را سریع‌تر آغاز می‌کند.

روش

این مطالعه از نوع توصیفی-تحلیلی می‌باشد. در این بخش از کار که به عنوان روش کار و پیاده‌سازی شناخته می‌شود سعی شده با ارائه یک الگوریتم مناسب و اجرای آن بر روی داده‌ها، نوع سرطان خون افراد مبتلا را با دقت بیشتری تشخیص داد. در ادامه به شرح دادگان مورد استفاده و مراحل الگوریتم پیشنهادی پرداخته می‌شود.

داده‌های "Leukemia1" یکی از شناخته‌شده‌ترین داده‌ها در تشخیص نوع سرطان خون است که توسط گروه انفورماتیک پزشکی دانشگاه واندربیلت ایالات متحده آمریکا تدوین شده است. در این داده‌ها، مجموعه‌ای از نمونه‌های مغز استخوان و نمونه خون بیماران لوسمی برای تشخیص تفاوت بین لوسمی حاد لنفوئیدی و لوسمی حاد میلوئیدی استفاده می‌شود. این مجموعه شامل ۷۲ نمونه با ۵۳۲۷ ژن است که ۳۸ نمونه مربوط به لوسمی میلوئیدی حاد، ۹ نمونه مربوط به لوسمی لنفوئیدی حاد نوع B و ۲۵ نمونه مربوط به لوسمی لنفوئیدی نوع T می‌باشد. طبقه‌بندی بر اساس ۳ زیر گروه سرطان خون ALL B-cell، ALL T-cell و AML می‌باشد [۱۲].

با توجه به اثربخشی الگوریتم‌های یادگیری ماشین در بسیاری از پژوهش‌ها [۱۱-۸]، در ابتدا به بررسی الگوریتم‌های پارامتریک و غیرپارامتریک در فرآیند تشخیص نوع سرطان خون پرداخته شد. روش‌های پارامتری، یادگیری را با استفاده از استخراج پارامترهای آماری از داده‌ها انجام می‌دهند. در مقابل، روش‌های غیرپارامتری سعی می‌کنند با توجه به ارتباط بین نمونه‌های مجموعه داده‌ها، عمل یادگیری را انجام دهند [۱۲].

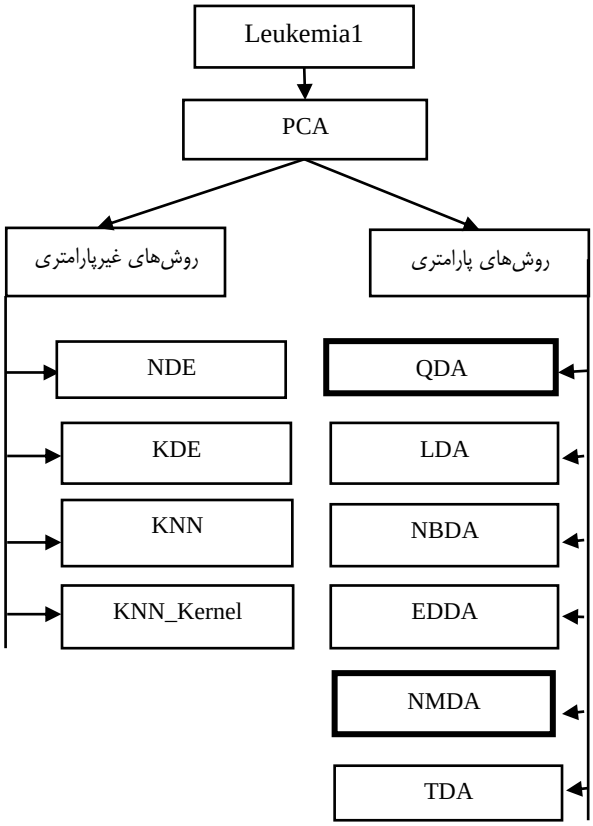
در روش‌های پارامتری تمام نمونه‌های آموزش، تأثیرگذار هستند؛ در حالی‌که در روش‌های غیرپارامتری، تنها نمونه‌هایی از داده‌های آموزش در یادگیری تأثیرگذارند که به نمونه جدید نزدیک می‌باشند. از جمله عیب‌های روش‌های پارامتری می‌توان به کند بودن و مصرف حافظه زیاد اشاره کرد.

در میان روش‌های پارامتری از روش‌های تحلیل تفکیک‌کننده خطی LDA(Linear Discriminant Analysis)، درجه دو QDA(Quadratic Discriminant Analysis) ، بیز ساده (Naïve Bayes Discriminant Analysis) NBDA، فاصله اقلیدسی (Euclidean Distance) EDDA(Discriminant Analysis)، نزدیک‌ترین میانگین (Nearest Mean Discriminant Analysis) NMDA و تطبیق قالب (Template Discriminant Analysis)TDA استفاده شد. همچنین در روش‌های غیرپارامتریک از روش‌های تخمین‌گر پایه (Naïve Density Estimation)NDE، مبتنی بر هسته (Kernel Density Estimation)KDE، k-همسایه نزدیک‌تر (K-Nearest Neighbor)KNN و k-همسایه نزدیک‌تر مبتنی بر هسته (KNN-Kernel) استفاده شد [۱۲].

در روش‌های غیرپارامتری برای محاسبه مقدار مناسب برای k (پارامتر تعداد نزدیک‌ترین همسایه‌ها در روش‌های KNN و KNN-Kernel) و h (پارامتر پهنای بازه در روش‌های NDE و KDE) از اعتبارسنجی متقابل با ۵-زیرمجموعه (5-fold cross validation) استفاده شد. مقدار h برای هر روش، از میان بازه بیشترین فاصله و کم‌ترین فاصله بین نمونه ها انتخاب شد. همچنین برای k، از میان مقادیر ۱ تا ۱۰، بهترین جواب که معیار دقت را بیشینه می‌کند، برای هر روش انتخاب می‌گردد. در روش KNN برای محاسبه فاصله نمونه ورودی با نمونه‌های آموزشی، از فاصله اقلیدسی استفاده شد. در روش KDE، برای ایجاد یک تخمین مناسب‌تر از نمونه‌ها، هسته گوسی مورد استفاده قرار گرفت.

داده‌های مورد استفاده دارای ۵۳۲۷ ویژگی هستند. هرچه تعداد ویژگی‌ها بیشتر باشد فرضیه‌های یادگیری که می‌توانند باتوجه به نمونه‌های آموزشی، مدل یادگیری را ایجاد کنند، به صورت نمایی افزایش می‌یابد. از آنجایی که این افزایش نمایی، به

تعداد ویژگی‌ها وابسته است؛ لذا ویژگی‌های کم‌تر سبب می شود تا استخراج فرضیه نهایی، که عملاً مدل یادگیری را شکل می‌دهد، شباهت بیشتری به مدل ذاتی داده‌ها داشته باشد؛ بنابراین با رویکرد کاهش ابعاد می‌توان انتظار داشت تا دقت تعمیم در یادگیری افزایش یابد. همچنین حذف ویژگی‌های نامناسب می‌تواند سبب کاهش نویز در داده‌ها گردد. در این مطالعه از روش PCA به منظور کاهش ویژگی استفاده شد. در این روش از فضای ویژگی‌های کنونی نگاشتی به یک فضا با ابعاد کمتر ایجاد می‌شود. در این روش ابتدا یک بردار ویژه (Eigen vector) به همراه مقادیر ویژه متعلق به آن از کواریانس داده‌های آموزشی محاسبه می‌شود. سپس از میان مقادیر ویژه حاصل، مقادیر ویژه‌ای که از یک حد آستانه (که بیانگر n درصد از ویژگی‌های مهم با بالاترین میزان واریانس می‌باشد) بیشتر باشد انتخاب شد [۱۲]. شکل ۱ نمودار گردش کار الگوریتم‌های اعمال‌شده بر روی داده Leukemia1 را نشان می‌دهد.



شکل ۱. نمودار گردش کار الگوریتم‌های اعمال شده

در بین نمونه‌های Leukemia1، فقط تعداد اندکی از آن‌ها لنفوئیدی حاد نوع B هستند. در عوض، تعداد بسیاری از آن‌ها لوسمی میلوئیدی حاد و لنفوئیدی حاد نوع T می‌باشند؛ بنابراین لنفوئیدی حاد نوع B که به نسبت کل نمونه‌ها تعداد کم‌تری دارند، به عنوان نمونه‌های کلاس اقلیت معرفی می‌شوند. در مقابل، لوسمی میلوئیدی حاد و لنفوئیدی حاد نوع T که بیشترین نمونه‌های سرطانی را تشکیل می‌دهند، به عنوان نمونه‌های کلاس اکثریت شناخته می‌شوند. به این مسئله در علم یادگیری ماشین، مسئله عدم توازن گفته می‌شود.

وقتی تعداد زیادی از نمونه‌های آموزشی مربوط به کلاس اکثریت باشد، نمونه‌های کلاس اقلیت کم‌تر مورد توجه قرار می‌گیرد. در نتیجه، مدل یادگیری در تخمین درست نمونه‌های اقلیت دقت کم‌تری دارد؛ لذا با متوازن‌سازی داده‌ها سعی می‌شود تا دقت پیش‌بینی نمونه‌های لوسمی افزایش یابد.

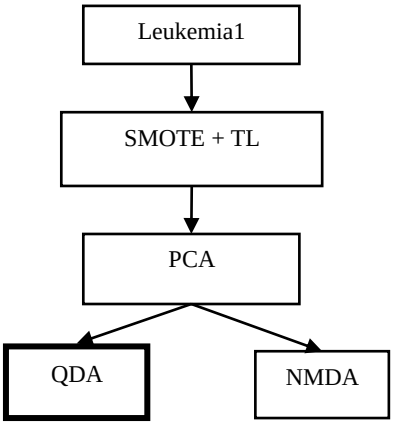
برای حل مسئله عدم توازن و مشکلاتی که از آن نشأت می‌گیرد، روش‌هایی پیشنهاد شده است [۱۴]. از متداول‌ترین این روش‌ها، SMOTE و روش‌های مبتنی بر آن است. این روش با استفاده از همسایه‌های هر نمونه از کلاس اقلیت، نمونه‌های مصنوعی جدیدی می‌سازد. به این صورت که در مرحله اول، به ازای هر نمونه i از کلاس اقلیت، k همسایه نزدیکترش در همان کلاس را پیدا می‌کند. در مرحله دوم به ازای تمام خصوصیات هر همسایه j در k همسایه نزدیک‌تر، فاصله نمونه i تا j را محاسبه می‌کند. سپس در مرحله سوم، مقداری بین ۰ تا ۱ را به عنوان شکاف (Gap) در نظر گرفته، آن را در فاصله i تا j ضرب کرده و با مقادیر خصوصیات i جمع می‌کند. در آخر مقادیر جدید به دست آمده را به عنوان مقادیر خصوصیات نمونه مصنوعی جدید در نظر می‌گیرد.

در پژوهش‌های انجام شده سعی شده است برای بهبود روش SMOTE، با افزودن ماژول‌های دیگری کارایی روش را افزایش دهند [۱۵]. در این پژوهش، برای حذف نویز از روش TL در کنار SMOTE استفاده شد. روش TL، از نزدیک ترین همسایه عضو کلاس دیگر استفاده می‌کند. در این روش دو نمونه را TL گویند اگر: ۱- از دو کلاس متفاوت باشند و ۲- هیچ نمونه‌ای دیگر از کلاس متفاوت، نسبت به آن‌ها نزدیک‌تر نباشد. به بیانی دیگر، نزدیک‌ترین دو نمونه از دو کلاس متفاوت، یک TL هستند. اگر دو نمونه، TL باشند، یا یکی از آن‌ها نمونه نویز است، یا هر دو نمونه، روی خط مرز هستند.

از این روش هم می‌توان به عنوان روشی برای کاهش نمونه و هم به عنوان روشی برای پاک‌سازی داده‌ها استفاده کرد. اگر در

TL، فقط نمونه متعلق به کلاس اکثریت حذف شود، رویکرد کاهش نمونه انجام شده. چنانچه هر دو نمونه حذف شوند، رویکرد پاک‌سازی داده‌ها انجام شده است.

این بار با استفاده از روش "SMOTE + TL" داده‌ها را متوازن کرده، سپس کاهش ویژگی PCA را اعمال نموده و دو الگوریتم QDA و NMDA بر روی داده‌های متوازن و دادگان با ویژگی کاهش‌یافته اجرا گردید (شکل ۲). نتایج نشان داد که بهترین نتیجه مربوط به روش درجه ۲ می‌باشد.



شکل ۲: نمودار گردش کار بهترین الگوریتم‌های اعمال‌شده پس از متوازن‌سازی و کاهش ویژگی

نتایج

در این تحقیق به منظور مقایسه‌ای عادلانه برای تشخیص نوع سرطان خون، از مجموعه داده‌های یکسان (مجموعه داده‌گان Leukemia1) استفاده شد. متغیر هدف در این مطالعه، تشخیص نوع سرطان خون است که در مورد هر کدام از افراد مورد بررسی یکی از سه حالت ALL B-cell، ALL T-cell و AML می‌باشد.

معیار مورد استفاده در سایر پژوهش‌ها برای سنجش و ارزیابی، دقت (Accuracy) می‌باشد [۶- ۳]. در این پژوهش نیز از همین معیار استفاده شد که در آن مثبت حقیقی (TP=True Positive)، موارد دارای سرطان است که به درستی پیش‌بینی شده‌اند و منفی حقیقی (TN=True Negative)، موارد منفی (بدون سرطان) است که به درستی پیش‌بینی شده‌اند و ALL تعداد کل موارد پیش‌بینی است. این میعار در رابطه ۱ نشان داده شد.

$$Accuracy = \frac{TP + TN}{ALL} \qquad (۱)$$

با توجه به عدم توازن داده‌ها، در این پژوهش از معیارهای صحت (Precision)، حساسیت (Recall) و سنجه F (F-measure) نیز استفاده می‌شود. صحت بیانگر تعداد نمونه‌هایی است که به عنوان مبتلا پیش‌بینی شده‌اند و در واقعیت نیز این چنین هستند و می‌تواند توانایی روش در کاهش نرخ مثبت کاذب را نشان دهد. صحت از طریق رابطه ۲ محاسبه می‌گردد. در این رابطه FP (False Positive) موارد مثبت (دارای سرطان) که به اشتباه سالم پیش‌بینی شده‌اند را بیان می‌کند.

$$Precision = \frac{TP}{TP + FP} \qquad (۲)$$

فراخوانی مجدد، درصدی از موارد که مبتلا تشخیص داده شده اند را محاسبه می‌کند که می‌تواند توانایی روش در کاهش نرخ منفی کاذب را بیان کند. فراخوانی مجدد از طریق رابطه ۳ محاسبه می‌گردد. در این رابطه FN (False Negative) موارد منفی (بدون سرطان) است که به اشتباه مبتلا پیش‌بینی شده‌اند.

$$Recall = \frac{TP}{TP + FN} \qquad (۳)$$

سنجه F، از ترکیب دقت و حساسیت حاصل می‌گردد. معیار F1 از طریق رابطه ۴ محاسبه می‌گردد:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \qquad (۴)$$

جهت پیاده‌سازی الگوریتم پیشنهادی و محاسبه معیارهای مورد نظر از نرم‌افزار متلب و سیستمی با پردازنده Core i7 با ۴ گیگا بایت رم و تحت ویندوز ۷ استفاده شد.

نتایج حاصل از روش‌های پارامتری، مطابق جدول ۱ می‌باشد. همان‌طور که مشاهده می‌شود روش طبقه‌بند NMDA، بهترین پاسخ را در روش‌های پارامتری به خود اختصاص می

دهد. روش QDA، به‌دلیل عدم معکوس‌پذیری در ابعاد بالا، قادر به محاسبه ماتریس کواریانس نمی‌باشند و برای داده‌های با ابعاد بالا کارایی ندارد. نتایج آزمایش‌ها نشان می‌دهد که روش‌هایی با پیچیدگی بالا مانند LDA، بر روی این داده‌ها موجب پدیده بیش برازش شده و در فرآیند پیش‌بینی نوع سرطان خون پاسخگو نمی‌باشند. همچنین روش‌های ساده مانند TDA به‌دلیل وقوع کم‌برازش به دقت مناسبی دست نیافته است.

جدول ۱: نتایج روش‌های پارامتری

	LDA	NBDA	EDDA	NMDA	TDA
بدون نرمال‌سازی	۳۱/۴۳	۵۱/۴۳	۵۱/۴۳	۹۳/۸۶	۵۴/۲۶
نرمال‌سازی zscore	۲۵/۷۱	۵۱/۴۳	۵۱/۴۳	۸۱/۴۳	۷۸/۵۷
نرمال‌سازی min-max	۲۷/۱۴	۵۱/۴۳	۵۱/۴۳	۳۴/۲۹	۲۵/۷۱

نتایج حاصل از پیاده‌سازی توابع غیرپارامتری بر روی داده‌های نرمال سازی شده و بدون نرمال‌سازی مطابق جدول ۲ می‌باشد. همان‌طور که مشاهده می‌شود روش‌های غیرپارامتری برای این داده‌ها عملکرد مطلوبی نداشته است.

جدول ۲: نتایج روش غیرپارامتری

الگوریتم	مقدار بهینه h	دقت		روش‌های نرمال‌سازی
NDE	۱۸۰۴۳۰	۴۰/۶۰	اعتبار سنجی	بدون نرمال‌سازی
		۴۰/۲۸	آزمون	
	۷۷	۵۳/۴۶	اعتبارسنجی	z-score
		۵۲/۷۸	آزمون	
KDE	۱۷۵۴۳۰	۵۲/۷۵	اعتبار سنجی	بدون نرمال‌سازی
		۵۲/۷۸	آزمون	
	۱۷۷	۵۲/۷۵	اعتبارسنجی	z-score
		۳۴/۷۲	آزمون	
الگوریتم	**مقدار بهینه K**	**دقت**		**روش‌های نرمال‌سازی**
KNN	۶	۵۸/۳۰	اعتبار سنجی	بدون نرمال‌سازی
		۵۲/۷۸	آزمون	
	۶	۴۴/۷۴	اعتبارسنجی	z-score
		۴۴/۴۴	آزمون	
KNN-kernel	۵	۳۴/۷۵	اعتبار سنجی	بدون نرمال‌سازی
		۳۴/۷۲	آزمون	
	۵	۳۴/۷۵	اعتبارسنجی	z-score
		۳۴/۷۲	آزمون	

در ادامه الگوریتم استخراج ویژگی با روش PCA پیاده‌سازی شد. روش استخراج ویژگی PCA بر روی الگوریتم‌های پارامتری، که در آن نتایج مطلوب بودند، اعمال شد. همچنین برنامه را ۱۰ بار تکرار کرده و از نتایج میانگین گرفته شد. بهترین نتایج توسط الگوریتم‌های NMDA و روش QDA حاصل شد (جدول ۳).

همان‌طور که در روش NMDA مشخص است، بهترین مقدار n توسط اعتبارسنجی ۵ زیرمجموعه، ۰/۵۶ به دست آمد. با مشخص شدن n، با تعداد ویژگی ۶/۸ می‌توان با دقت ۹۶٪ نوع سرطان خون را تشخیص داد. همچنین در روش QDA مشاهده می‌شود که با نگهداری ۴۲٪ از اطلاعات (n=۰/۴۵) به طور متوسط ۴/۲ ویژگی به دست آمد و به دقت ۹۵/۳۳٪ رسید.

جدول ۳: میانگین نتایج بهترین الگوریتم‌های پارامتری با روش PCA

الگوریتم		میانگین
NMDA	دقت اعتبارسنجی	۹۲/۹۸
	دقت آزمون	۹۶
	پارامتر PCA آزمون	۵۶
	ابعاد آزمون	۶/۸
	ابعاد اعتبارسنجی	۷/۴
QDA	دقت اعتبارسنجی	۸۹/۱۲
	دقت آزمون	۹۵/۳۳
	پارامترPCA آزمون	۴۲
	ابعاد آزمون	۴/۲
	ابعاد اعتبارسنجی	۴/۹۴

دو الگوریتم QDA و NMDA، که در آن‌ها نتایج بهتری به دست آمده است ۱۰۰ بار دیگر اجرا شد. سپس از نتـایج اجـرای مجدد آن‌ها میانگین گرفته شد (جداول ۴ و ۵).

جدول ٤: میانگین معیارهای ارزیابی روش NMDA در ۱۰۰ اجرا

	دقت	باخوانی مجدد	صحت	معیار_اف
کلاس۱	۹۳/۴	۹۶/۱۰	۹۲/۹۴	۹۴/۵۰
کلاس ۲	۹۳/۴	۸۸/۴۶	۸۷/۵	۸۷/۹۸
کلاس۳	۹۳/۴	۹۱/۰۱	۹۶/۳۶	۹۳/۶۱

جدول ٥: میانگین معیارهای ارزیابی روش QDA در ۱۰۰ اجرا

	دقت	باخوانی مجدد	صحت	معیار_اف
کلاس۱	۹۰/۸۸	۹۴/۷۲	۹۳/۱۲	۸۷/۸۸
کلاس ۲	۹۰/۸۸	۸۲/۱۱	۷۲/۸۹	۹۵/۱۹
کلاس۳	۹۰/۸۸	۸۷/۸۸	۷۷/۲۳	۹۱/۳۹

در مرحله بعد با استفاده از روش "SMOTE + TL" داده‌ها را متوازن کرده، سپس کاهش ویژگی PCA را اعمال نمـوده و دو الگوریتم QDA و NMDA بـر روی داده‌هـای متـوازن و کاهش ویژگی یافته اجرا گردید. بهترین نتیجه مربوط بـه روش درجه ۲ می‌باشد که به دقت قابل ملاحظه ۹۸/۵۹٪ در ۱۰۰ مرحله تکرار رسیده است (میـانگین ۱۰۰ اجـرا). در اینجـا چون معیار ارزیابی سایر پژوهش‌ها دقت بود برای مقایسه تنهـا مقدار این معیار ارزیابی آورده شد (جدول ۶).

جدول ٦: میانگین دقت در ۱۰۰ اجرا با داده‌های متوازن شده

الگوریتم	دقت
روش NMDA	۹۴/۷۶
روش QDA	۹۸/۵۹

بحث و نتیجه‌گیری

در این پژوهش یک سیستم پیش‌بینی برای تشخیص نوع سرطان خون ارائه شد. برای ارزیابی از داده‌های Leukemia1 استفاده گردید. این سیستم برای دسته‌بندی افراد به عنوان بیمار سرطان خون نوع ALL B-cell، ALL T-cell و AML از روش پارامتری و غیر پارامتری با کاهش ویژگی PCA و متوازن‌سازی "SMOTE + TL" استفاده می‌نماید.

در پژوهش‌های گذشته، Prasartvit و همکاران از روش ABCبرای کاهش ویژگی به همراه الگوریتم kNN با انتخاب ۱۰ ویژگی به دقت ۹۷ رسیدند [۳].

Li و همکاران با روش انتخاب ویژگی MFA score و روش بازشناسی الگو SVM با انتخاب ۷ ویژگی به دقت ۹۲٪ رسیدند [۴].

Wang و همکاران با روش RS و طبقه‌بند DC با انتخاب ۲۰/۵۶ ویژگی به دقت ۹۴/۱۲٪ دست یافتند [۵].

همچنین Chen و همکاران پس از انتخاب ۱۰ویژگی توسط روش VIMs، نمونه‌های موجود با ویژگی‌های حاصل، توسط جنگل تصادفی طبقه‌بندی شدند و به دقت ۹۴٪ رسیدند [۶].

بررسی‌هایی بر روی روش‌های پارامتری مبتنی بر قاعده بیز و روش‌های غیرپارامتری مبتنی بر همسایه نمونه‌ها صورت پذیرفت. این روش‌ها بر روی داده‌های Leukemia1 بررسی گردید. با روش NMDA (از روش‌های پارامتری) بهترین دقت حاصل شد. دقت به دست آمده، ۹۲/۸۶٪، با تمامی ویژگی‌های موجود بود. با اعمال روش کاهش ویژگی PCA، با متوسط تعداد ویژگی ۶/۸، دقت روش NMDA، ۹۶٪ و با متوسط ویژگی ۴/۲ دقت الگوریتم QDA، ۹۵/۳۳ به دست آمد. بر روی داده‌های مربوطه روش "SMOTE + TL"، برای متوازن ساختن داده‌ها و سپس PCA اعمال گردید. بر روی داده‌های حاصل روش NMDA و QDA اجرا گردید. الگوریتم QDA با تعداد ژن ۵/۴۱ به دقت قابل ملاحظه ۹۸/۵۹٪ دست یافت.

جدول ۷ نتایج مربوط به مقالات اخیر و بهبودهای انجام شده در این پژوهش را نشان می‌دهد. همان‌طور که مشخص است، در روش‌های PCA + QDA، PCA + NMDA و QDA + PCA + Smote-TL در مقایسه با دیگر پژوهش‌ها [۶-۳]، حتی با استفاده از تعداد ژن‌های کم‌تر، دقت افزایش یافته است. به عنوان محدودیت روش پیشنهادی، به مشخص نکردن دقیق ژن‌های مؤثر در طبقه‌بندی نوع سرطان خون می‌توان اشاره کرد.

با توجه به این که توابع توزیع غیر پارامتری، کلاس هر نمونه را با توجه به نمونه‌های همسایه پیش‌بینی می‌کند، این که داده‌های ما با روش غیرپارامتری دقت پایینی دارد این موضوع را نشان می‌دهد که کلاس هر نمونه از داده‌های ما تنها وابسته به داده‌های همسایه نیست. مدل توزیع احتمال پارامتری، که با استفاده از تخمین تعداد کمی پارامتر مثل میانگین، واریانس و کواریانس ایجاد می‌شود، مؤثر بودن روش پارامتری که در کل مجموعه داده‌ها برقرار می‌باشد را نشان می‌دهد.

از طرفی ویژگی‌های کم‌تر سبب می‌شود تا استخراج فرضیه نهایی، که عملاً مدل یادگیری را شکل می‌دهد، شباهت بیشتری به مدل ذاتی داده‌ها داشته باشد؛ بنابراین با رویکرد کاهش ویژگی دقت تعمیم در یادگیری افزایش یافت.

از طرف دیگر، وقتی تعداد زیادی از نمونه‌های آموزشی مربوط به کلاس اکثریت باشد، نمونه‌های کلاس اقلیت کمتر مورد توجه قرار می‌گیرد. در نتیجه دقت مدل یادگیری در تخمین درست نمونه‌های اقلیت کم‌تر می‌شود؛ لذا متوازن‌سازی داده‌ها، موجب افزایش دقت پیش‌بینی نمونه‌های لوسمی گردید.

بنابراین در این مطالعه با استفاده از الگوریتم‌های یادگیری ماشین پارامتری و همچنین استفاده از رویکرد کاهش ابعاد و متوازن‌سازی داده‌ها به نتایج مطلوبی جهت پیش‌بینی نوع سرطان خون یافت شد. با استفاده از این مدل ارائه شده، بدون آزمایش‌های طاقت فرسای نمونه‌گیری از مغز استخوان بیمار و همچنین با هزینه کم می‌توان نوع سرطان خون بیمار را برای انتخاب روش معالجه بهتر و سریع‌تر پیش‌بینی کرد.

جدول ۷: دقت الگوریتم‌های مختلف در طبقه‌بندی سرطان خون

الگوریتم	دقت	تعداد ژن
PCA + QDA الگوریتم پیشنهادی (میانگین ۱۰ تکرار)	۹۵/۳۳	۴/۲
PCA + NMDA الگوریتم پیشنهادی (میانگین ۱۰ تکرار)	۹۶	۶/۸
PCA + Smote-TL + QDA الگوریتم پیشنهادی (میانگین ۱۰۰ تکرار)	۹۸/۵۹	۵/۴۱
VIMs based on random forest [۵]	۹۴	۱۰
RS-based DC [۶]	۹۴/۱۲	۲۰/۵۶
MFA score+ based on SVM [۴]	۹۲	۷
MFA score+ based on SVM [۴]	۹۳	۱۵
ABC+ based on kNN [۳]	۹۷	۱۰

در پژوهش‌های آینده سعی خواهد شد، از روش‌های انتخاب ویژگی، که نوع دیگری از روش‌های کاهش ویژگی است، به جای استخراج ویژگی استفاده نموده و سپس اثر الگوریتم‌های بیز را بر روی ویژگی‌های انتخاب شده بررسی کند. انتظار می‌رود این رویکرد بتواند با حفظ ویژگی‌های اصلی، ژن‌های مؤثر در ابتلا به هر نوع سرطان را نشان دهد.

References

1. Zhou Q, Zhou H, Li T. Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features. Knowledge-Based Systems 2016;95:1-11.

2. Minnie D, Srinivasan S. Preprocessing and generation of association rules for prediction of acute myeloid leukemia from bone marrow data. Journal of Theoretical and Applied Information Technology 2014; 70(2): 415-24.

3. Prasartvit T, Banharnsakun A, Kaewkamnerdpong B, Achalakul T. Reducing bioinformatics data dimension with ABC-kNN. Neurocomputing 2013;116:367-81.

4. Li J, Su L, Pang Z. A filter feature selection method based on MFA score and redundancy excluding and it's application to tumor gene expression data analysis. Interdisciplinary Sciences: Computational Life Sciences 2015; 7(4): 391-6.

5. Chen Y, Wang L, Li L, Zhang H, Yuan Z. Informative gene selection and the direct classification of tumors based on relative simplicity. BMC Bioinformatics 2016; 17: 44.

6. Wang H, Yang F, Luo Z. An experimental study of the intrinsic stability of random forest variable importance measures. BMC Bioinformatics 2016;17:60.

7. Voyant C, Notton G, Kalogirou S, Nivet, ML, Paoli C, Motte F, et al. Machine learning methods for solar radiation forecasting: A review. Renewable Energy 2017; 105: 569-82.

8. Heidari M, Seifossadat G, Razaz M. an intelligent-based islanding detection method using K-Nearest neighbors and discrete wavelet transform. Journal of Electrical Engineering 2013; 43(1): 6-15. Persian

9. Remagnino P, Mayo S, Wilkin P, Cope J, Kirkup D. Machine Learning for Plant Leaf Analysis. Computational Botany: Methods for Automated Species Identification. Berlin, Heidelberg: Springer Berlin Heidelberg; 2017. p. 57-79.

10. Mishra S, Sharma L, Majhi B, Sa PK. Microscopic Image Classification Using DCT for the Detection of Acute Lymphoblastic Leukemia (ALL). Proceedings of International Conference on Computer Vision and Image Processing: CVIP; 2016. p. 171-80.

11. Mousavizadegan M, Mohabatkar H. An evaluation on different machine learning algorithms for classification and prediction of antifungal peptides. Med Chem 2016;12(8):795-800.

12. Alpaydin E, Introduction to Machine Learning. . 3th ed. London, England: MIT press; 2014.

13. Leukemia1. [cited: 2016 Apr 1]. Available http://www.gems-system.org.

14. Barandela R, Sanchez JS, Garcia V, Rangel E. Strategies for learning in class imbalance problems. Pattern Recognition 2003; 36: 849-51.

15. Batista GE, Prati RC, Monard MC. A study of the behavior of several methods for balancing machine learning training data. ACM Sigkdd Explorations Newsletter 2004; 6(1): 20-9.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2018; 5(1): 25-34

Diagnosis of Leukemia Type by Machine Learning: Dimension Reduction and Balancing

Gharaati Zeinab[1], Pajoohan MohammadReza[2]*

• **Received:** 16 Nov, 2017 • **Accepted:** 7 May, 2018

Introduction: Combination of artificial intelligence and data mining has been resulted to considerable progress in the prevention and diagnosis of diseases. Complex models have been proposed for the diagnosis of acute leukemia from genetic information, but significant results have not been achieved. This study aimed to predict the type of blood cancer by examining a wide range of parametric and non-parametric methods and to increase the generalization of learning by extracting fewer essential features.

Methods: This descriptive and analytical study used Leukemia1 dataset from the Vanderbilt University of USA. This dataset contains a set of bone marrow and blood samples of patients having leukemia used for classification based on three subgroups of leukemia, namely ALL B-cell, ALL T-cell and AML. Parametric classification including linear algorithms, Naïve Bayes, Euclidean distance, nearest average, template matching as well as non-parametric classification using basic estimator algorithms, kernel, k-nearest neighbors and k-nearest neighbors based on the kernel has been used.

Results: Considering all features, the best method was nearest mean prediction method achieving the accuracy of 92.86%. By applying the PCA feature reduction method, too, the best result was related to the nearest mean algorithm and by average number of features of 6.8, the accuracy became 96%. Finally, using data-balancing methods and quadratic algorithm resulted in the average number of features and the accuracy of 5.41 and 98.59% respectively.

Conclusion: The results show the effectiveness of essential features extraction in improving the accuracy of Bayes-based models and its preference over the existing complex models.

Keywords: Genetics data, Diagnosis of type of blood cancer, Data mining, Data balancing, Dimension reduction.

• Citation: Gharaati Z, Pajoohan MR. Diagnosis of Leukemia Type by Machine Learning: Dimension Reduction and Balancing. Journal of Health and Biomedical Informatics 2018; 5(1): 25-34.

1. M.Sc. Student in Computer Engineering, Computer Engineering Dept., Yazd University, Yazd, Iran.
2. Ph.D in computer Engineering, Assistant Professor of Computer Engineering, Department of Computer Engineering Dept., Yazd University, Yazd, Iran.

***Correspondence**: Computer Engineering Dept., Yazd University, Yazd, Iran
• **Tel:** 035-31232358 • **Email:** pajoohan@yazd.ac.ir