نشریه مهندسی مکانیک ایران، دوره ۲۶، شماره ۲، شماره پیاپی۷۵، تابستان ۱۴۰۳، صفحه ۶۶-۴۶
انجمن مهندسان مکانیک ایران
مقاله علمی پژوهشی

DOI: https://doi.org/10.30506/ijmep.2023.2003519.1940

طبقه بندی پنج نوع داده سرطان بر اساس روش های شبکه عصبی و تحلیل و بررسی بیان ژن بر اساس روش همجوشی انتخاب ویژگی

فرنوش ترکی[۱]
دانشجوی دکتری

عاطفه خادم[۲]
کارشناسی ارشد

عبدالحسین جلالی آقچای[۳]
دانشیار

در داده های میکروآرایه با حجم بالا، تعداد کم نمونه ها و تغییرپذیری ذاتی در فرآیندهای بیولوژیکی باعث ایجاد مشکل افزایش هزینه محاسباتی و پیچیدگی طبقه بندی ها می شود. همچنین تفسیر ژن های عامل بیماری پیچیده است، چرا که از نظر بیولولوژیکی، تنها مجموعه کوچکی از ژن ها می توانند بیماری را با دقت بیشتری توصیف نمایند. اولین قدم در آنالیز داده های میکروآرایه، کاهش قابل توجه تعداد ژن ها یا به عبارتی انتخاب ژن های متمایز کننده در فرآیند طبقه بندی است. این مرحله انتخاب ژن نامیده می شود. در این مقاله از طبقه بندی بیان ژن پنج داده، سرطان روده، سرطان پستان، لوسمی، تومورهای پروستات و لنفوم های سلول های بزرگ پخش شده استفاده شده و هر یک از آن ها به تفکیک در چرخه انتخاب ویژگی و نیز دسته بندی با تعداد ویژگی های متغیر وارد می شوند.

واژه‌های راهنما: یادگیری ماشین، شبکه عصبی، تحلیل ژن های سرطان، طبقه بندی داده ها، بایومکانیک

۱- مقدمه

وظیفه اصلی طبقه‌بندی داده‌ها گرفتن یک بردار ورودی و اختصاص آن به یک کلاس جداگانه است. اساسی‌ترین نوع طبقه‌بندی، تمایز بین دو کلاس است که به آن طبقه‌بندی باینری می‌گویند و طبقه‌بندی پیچیده‌تر مربوط به بیش از دو طبقه است که به طبقه‌بندی چند طبقه‌ای معروف است. امروزه تحقیقات نشان می‌دهد که طبقه‌بندی میکروآرایه‌ها در تشخیص سرطان موثر است [۱].

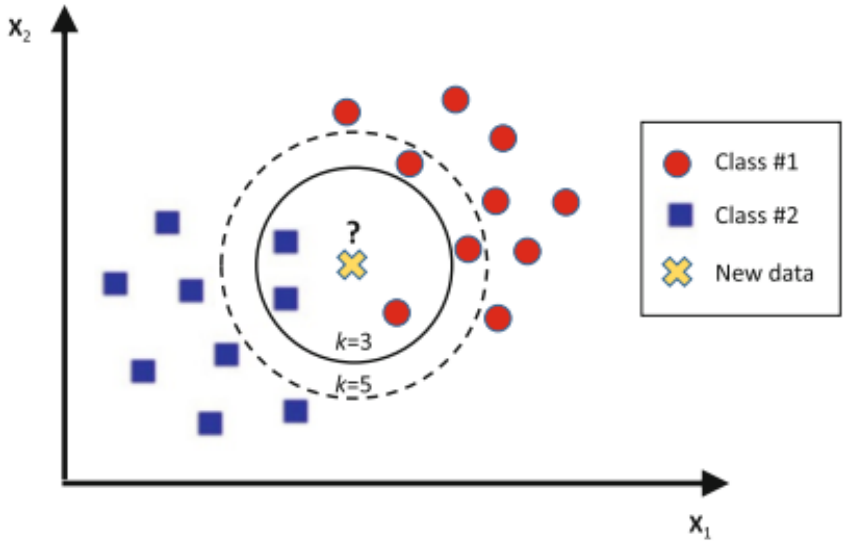
[۱] دانشجوی دکتری، دانشکده مهندسی مکانیک، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران، farnoosh.turki@email.kntu.ac.ir
[۲] کارشناسی ارشد، دانشکده مهندسی مکانیک، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران، atefeh.khadem@gmail.com
[۳] نویسنده مسئول، دانشیار، دانشکده مهندسی مکانیک، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران، jalali@kntu.ac.ir

تاریخ دریافت: ۱۴۰۲/۰۳/۰۸، تاریخ بازنگری: ۱۴۰۲/۰۶/۰۷، تاریخ پذیرش: ۱۴۰۲/۰۷/۱۵

داده‌های میکروآرایه دو ویژگی منحصر به‌فرد دارند که آن‌ها را از سایر روش‌های طبقه‌بندی خودکار متمایز می‌کند: ابعاد بالا و تعداد کم نمونه. این دلایل باعث ایجاد مشکلاتی در بسیاری از روش‌های یادگیری ماشین می‌شود. برازش بیش از حد داده‌ها یا ابعاد بالا در طبقه‌بندی رخ می‌دهد. بنابراین برای رفع مشکلات ذکر شده، بسیاری از روش‌هایی که در مرور تاریخچه تحقیق به آن‌ها اشاره خواهد شد، بر اساس مدل‌های ساده خطی یا غیرخطی تعریف شدند.

۱-۱- K نزدیکترین همسایگی[1]

این الگوریتم نوعی از الگوریتم یادگیری ماشینی نظارت شده است که در مسائل طبقه‌بندی و رگرسیون پیش‌بینی استفاده می‌شود و اغلب در مسائل طبقه‌بندی پیش‌بینی در صنعت استفاده می‌شود. K-nn یک الگوریتم یادگیری ساده است؛ زیرا مرحله یادگیری خاصی ندارد و در طول طبقه‌بندی از تمام داده‌ها برای یادگیری استفاده می‌کند. هم‌چنین، K-nn یک الگوریتم یادگیری بدون پارامتر است زیرا هیچ فرضی در مورد داده‌های اصلی ایجاد نمی‌کند. برای پیاده‌سازی هر الگوریتم، یک مجموعه داده مورد نیاز است. بنابراین، در مرحله اول K-nn، داده‌های آموزشی باید همراه با داده‌های آزمون بارگذاری شوند. سپس مقدار K باید به‌عنوان نزدیک‌ترین نقاط داده انتخاب شود. K می‌تواند هر عدد صحیحی باشد. با کمک هر یک از روش‌های اقلیدسی، فاصله همینگ[2] یا فاصله منهتن[3]، فاصله بین داده‌های آزمون و هر خط از داده‌های آموزشی محاسبه می‌شود. رایج‌ترین روش برای محاسبه فاصله، روش اقلیدسی است. مرحله بعدی مرتب کردن آن‌ها به‌ترتیب صعودی بر اساس مقدار فاصله است. اکنون بر اساس رایج‌ترین کلاس این خطوط، الگوریتم یک کلاس را به نقطه تست اختصاص می‌دهد. شکل (۱)، نمونه‌ای از الگوریتم K را نشان می‌دهد که در آن مقادیر K، ۳ و ۵ هستند. مهم‌ترین نکته در این الگوریتم بدست آوردن مقدار بهینه برای K است [۲].



شکل ۱- مثالی برای الگوریتم K-nn برای مقادیر K، ۳ و ۵ [۲]

[1] k-Nearest Neighbor (K-nn)
[2] Hamming
[3] Manhattan

۱-۲- شبکه ماشین بردار پشتیبان[1]

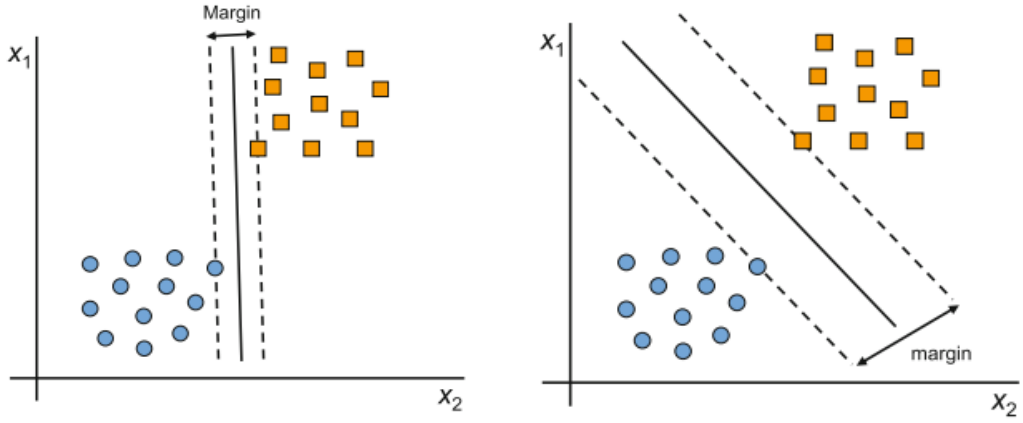
این روش یک الگوریتم یادگیری ماشینی نظارت شده است که نمونه‌های داده‌ای را که به‌صورت نقاط در فضا با استفاده از یک خط یا ابر صفحه نمایش داده شده‌اند، جدا می‌کند. این تفکیک به این صورت است که نقاط داده‌ای که در یک سمت خط قرار دارند مشابه یکدیگر هستند و در یک گروه قرار می‌گیرند. نمونه داده‌های جدید پس از اضافه شدن به‌همان فضا در یکی از دسته‌بندی‌های موجود قرار می‌گیرند.
در شکل (۲)، بخش (الف) نسبت به بخش (ب) حاشیه بزرگتری دارد. از نقطه نظر علمی، حداکثر حاشیه هایپرپلن با مشکل بیش از حد برازش کنار می‌آید و ظرفیت تعمیم خوبی دارد [۳].

۱-۳- درخت تصمیم[2]

یکی از پرکاربردترین الگوریتم‌های داده کاوی، الگوریتم درخت تصمیم است. در داده کاوی، درخت تصمیم یک مدل پیش‌بینی است به‌طوری که می‌توان از آن برای هر دو مدل رگرسیون و کلاس استفاده کرد. هنگامی که درخت برای وظایف طبقه‌بندی استفاده می‌شود، به‌عنوان درخت طبقه‌بندی شناخته می‌شود و زمانی‌که برای فعالیت‌های رگرسیون استفاده می‌شود، درخت تصمیم‌گیری رگرسیون نامیده می‌شود [۴،۵]. هنگامی‌که یک طبقه‌بندی ساخته می‌شود، اندازه‌گیری دقت آن از اهمیت بالایی برخوردار است. بهتر است از داده‌های آزمایشی برای اندازه‌گیری دقت یک طبقه‌بندی کننده استفاده کنید؛ یعنی پس از ساخت مدل بر روی داده‌های آموزشی، دقت مدل در تعیین برچسب کلاس نمونه‌ها بر روی داده‌های آزمون آزمایش شود.

۱-۴- تئوری بیز[3]

هنگامی که با مشکل طبقه‌بندی مواجه می‌شوید، یکی از ساده‌ترین روش‌های طبقه‌بندی می‌تواند استفاده از الگوریتم ساده بیز باشد.



شکل ۲- مثالی برای (**الف**) SVM با حاشیه بزرگ (**ب**) با حاشیه کوچک [۳]

[1] Support Vector Machines (SVM)
[2] Decision trees and random forests
[3] Bayes theorem

در بیشتر موارد، زمانی که تعداد متغیرها کم است اما مشاهدات زیاد هستند، الگوریتم ساده بیز برای تشخیص دسته‌ها مناسب است. البته واضح است که اساس الگوریتم طبقه‌بندی بیز، قضیه بیز است. این الگوریتم از نظریه بیز استفاده می‌کند که رابطه آن مطابق رابطه (۱) می‌باشد.

$$p(c/x)=\frac{p(x/c)p(c)}{p(x)} \quad (۱)$$

در رابطه (۱)، $p(c/x)$ احتمال پسین، $p(x/c)$ تابع درست‌نمایی، $p(c)$ احتمال کلاس پیشین و $p(x)$ احتمال پیش‌بینی است [۶].

۱-۵- مروری بر مطالعه‌های انجام شده

کاهش ابعاد، تبدیل داده‌های با ابعاد بالا به نمایش معنادار کاهش یافته است. تکنیک‌های کاهش ابعاد را می‌توان به دو دسته تقسیم کرد: استخراج ویژگی و انتخاب ویژگی.

جدول ۱- خلاصه مطالعه‌های انجام شده برای طبقه‌بندی میکروآرایه‌ها

	اعتبارسنجی	پیش‌پردازش	وظیفه	شماره منبع
SVM	Test set, HO, CV, LOOCV, Bootstrap	فضای عدم تشابه، درخت سه اتصاله ایزومپ، FE، FS و الگوریتم کلونی زنبورعسل[1]، FS بر اساس CSO[2]، بهینه‌سازی چند منظوره تکاملی، روش فازی[3]، روش انتخاب ژن بر اساس الگوریتم کلونی مورچه[4]، رگرسین PCA-Log	دوتایی، چندتایی، یک کلاسه	[۱۵-۸]
KNN	HO, CV, LOOCV	FPRS،، KFRS ،FS تئوری مارکو[5]	دوتایی، چندتایی	[۲۱-۱۶]
پیش‌بینی تصادفی	Test set, CV	FS،FS و CSO	دوتایی، چندتایی	[۲۴-۲۲]
درخت تصمیم	HO, CV	FS, PCA, MGSACO	دوتایی، چندتایی	[۲۷-۲۵]
یادگیری ماشین ویژه	Test set, Ho, CV	FS, PCA-Log	دوتایی، چندتایی	[۳۰-۲۸]
تئوری بیز	Test set, Ho, CV	FE, FS, FPRS, KFRS, PCA-Log	دوتایی، چندتایی	[۳۲-۳۱]
لاسو[6]	HO, Bootstrap	ندارد	دوتایی	[۳۴-۳۳]

[1] Artificial Bee Colony (ABC)
[2] Cat swarm optimization
[3] Fuzzy preference-based rough set
[4] Microarray Gene Selection based on Ant Colony (MGSACO)
[5] Markov blanket
[6] Lasso

استخراج ویژگی(FE) فرآیندی است که مجموعه‌ای از ویژگی‌های جدید را از طریق نقشه‌برداری عملکردی، از ویژگی‌های اصلی استخراج می‌کند.
در چند دهه گذشته روش کاهش ابعادی داده‌ها با تمرکز بر استخراج ویژگی می‌باشد [۷]. روش‌های جدید انتخاب ویژگی به‌طور مداوم در حال توسعه هستند؛ بنابراین مجموعه وسیعی در دسترس محققان است. به‌طور معمول، اکثر روش‌های انتخاب ژن در رویکرد فیلترینگ قرار می‌گیرند.
در حال حاضر روش‌های طبقه‌بندی متعددی وجود دارد که در زمینه میکروآرایه‌ها به‌کار گرفته شده است، در این قسمت سعی شد که این منابع از منظر تکنیک‌های پیش‌پردازش و اعتبارسنجی مورد استفاده، در جدول (۱) دسته‌بندی شود. جدول (۱)، متداول‌ترین روش‌های یادگیری ماشین را نشان می‌دهد که با بررسی‌های انجام شده روش SVM معمول‌ترین روش بکارگرفته شده می‌باشد.

۲- رویه کار پژوهشی

پنج مجموعه داده محک برای انجام آزمایش‌ها شامل استفاده از لنفوم‌های سلول‌های بزرگ پخش شده[1] [۳۵] سرطان لوسمی [۳۶] و سرطان پروستات [۳۷]، سرطان روده [۳۸] و سرطان پستان [۳۹] است. در مجموعه داده‌ها، DLBCL و لنفوم فولیکولار (FL) دو نوع بدخیمی هستند که باید طبقه‌بندی شوند. مجموعه داده DLBCL حاوی ۷۰۷۰ ژن از ۷۷ نمونه است که در آن ۵۸ نمونه توسط DLBCL انتخاب شده و سایرین دارای FL هستند. مدل‌های طبقه بندی با استفاده از بیان ژن ساخته می‌شوند تا بین این دو لنفوم تمایز ایجاد کنند. مجموعه داده‌های لوسمی حاوی نمونه‌های لوسمی حاد لنفوبلاست (ALL) و لوسمی حاد میلوئید (AML) از مغز استخوان و محیط کشت خون است. این مجموعه داده شامل ۷۲ نمونه (۴۷ ALL و ۲۵ AML) است که بیان ژن در آنها بر روی ۷۱۲۹ نمونه ژن اندازه‌گیری می‌شوند. در مجموعه داده سوم، ۱۰۲ نمونه (۵۰ نمونه طبیعی و ۵۲ نمونه بافت پروستات) وجود دارد که هر کدام از آنها شامل ۱۲۵۳۳ ژن است. تمام داده‌های ژن در مقطع عرضی با روش نرمالیزه سازی کوانتایل هنجارسازی می‌شوند. همچنین داده سرطان روده شامل ۲۰۰۰ ژن توصیف‌کننده از ۶۲ بافت (۶۲ نمونه) است که ۲۲ نمونه (معادل ۳۵/۵ ٪ کل داده) سالم و ۴۰ نمونه (معادل ۶۴/۵ ٪ کل داده) سرطانی هستند. در نهایت داده‌های سرطان پستان شامل ۱۳۲ نمونه بافت و ۱۹۲۶ ژن بیان کننده است و ۱۱ نمونه سالم (معادل ۸/۳ ٪ کل داده) و ۱۲۲ داده از بافت سرطانی (معادل ۹۲ ٪ کل داده) می‌باشد. توضیح کوتاهی از داده‌های نمونه به همراه شاخصه‌هایی چون تعداد ویژگی‌ها، تعداد نمونه‌ها و تعداد کلاس‌های مربوطه در جدول (۲)، به نمایش در آمده است.
در مرحله طبقه‌بندی از ۴ دسته‌بند خانواده شبکه عصبی شامل شبکه عصبی دینامیکی، شبکه ماشین بردار پشتیبان، تئوری بیز، K نزدیکترین همسایگی و در نهایت درخت تصمیم استفاده شده است. انتخاب ویژگی نیز مبتنی بر رای گیری اکثریت به روش‌های ریلایف[2]، پی سی سی، روش F-Score و Term Variance نتایج حاکی از آن است در مجموع بکارگیری شیوه همجوشی میان ویژگی‌ها تا حد چشمگیری بر دقت اثر مطلوب گذاشته است. در ادامه در جداول و شکل‌ها عملکرد برای هر داده به تفکیک نمایش داده شده است.

---

[1] Diffuse large B-cell lymphomas (DLBCL)
[2] Relief

جدول ۲- داده‌های نمونه به همراه شاخصه‌هایی چون تعداد ویژگی‌ها، تعداد نمونه‌ها و تعداد کلاس‌های مربوطه

داده‌های از دست رفته	شماره‌کلاس	شماره ویژگی‌ها	نام نمونه‌ها	نام نمونه ژن‌ها	شماره دسته داده‌ها
ندارد	۲	۷۰۷۰	۷۷	لنفوم سلول‌های پخش‌شده	۱
ندارد	۲	۱۲۵۳۳	۱۰۲	سرطان پروستات	۲
ندارد	۲	۷۱۲۹	۷۲	سرطان لوسمی	۳
ندارد	۲	۲۰۰۰	۶۲	سرطان روده	۴
ندارد	۲	۱۹۲۶	۱۳۲	سرطان سینه	۵

۲-۱- روش‌های انتخاب ویژگی مورد استفاده

الف) روش ریلایف

روش ریلایف از یک راه حل آماری برای انتخاب ویژگی استفاده می‌کند؛ همچنین یک روش مبتنی بر وزن است که از الگوریتم‌های مبتنی بر نمونه الهام گرفته است. روش کار به این صورت است که از میان مجموعه نمونه‌های آموزشی، یک زیر مجموعه نمونه انتخاب می‌کنیم. تعداد نمونه‌ها در این زیرمجموعه را باید مشخص کنیم و آن را به عنوان ورودی به الگوریتم ارائه دهیم. الگوریتم به‌صورت تصادفی یک نمونه از این زیرمجموعه را انتخاب می‌کند، سپس برای هر یک از ویژگی‌های این نمونه، نزدیکترین برخورد و نزدیکترین شکست را بر اساس معیار اقلیدسی پیدا می‌کند.

الگوریتم ریلایف از طریق فرآیند نمونه‌گیری و مقایسه تکراری عمل می‌کند. برای هر نمونه در مجموعه داده، نزدیک‌ترین همسایگان را از کلاس‌های یکسان و متفاوت بر اساس برخی پارامترهای فاصله شناسایی می‌کند. تفاوت بین مقادیر ویژگی نمونه و همسایگان آن به وزن اختصاص داده شده به هر ویژگی کمک می‌کند. ویژگی‌هایی که به طور مداوم برای نمونه‌هایی با یک کلاس در مقایسه با ویژگی‌هایی با کلاس‌های مختلف تفاوت‌های بزرگ‌تری دارند، مرتبط‌تر در نظر گرفته می‌شوند و احتمال بیشتری برای انتخاب دارند.

ایده اصلی در این الگوریتم این است که هر چه اختلاف بین اندازه یک ویژگی در نمونه انتخاب شده و نزدیک‌ترین برخورد کمتر باشد، این ویژگی بهتر است و بعلاوه یک ویژگی خوب آن است که اختلاف بین اندازه آن ویژگی و نزدیک‌ترین شکست آن بیشتر باشد. دلیل کار هم خیلی ساده است، ویژگی‌هایی که به خوبی دو کلاس (یا یک کلاس از سایر کلاس‌ها) را از هم تمیز می‌دهند، برای نمونه‌های متعلق به دو کلاس متفاوت، مقادیری نزدیک به هم نمی‌دهند و فاصله معناداری بین مقادیری که به نمونه‌های یک کلاس می‌دهند و مقادیری که به سایر کلاس(ها) می‌دهند، وجود دارد.

ب) روش ضریب همبستگی پیرسون

در مباحث آماری، ضریب همبستگی پیرسون (PCC) میزان همبستگی خطی بین دو متغیر تصادفی را اندازه-گیری می‌کند. مقدار این ضریب بین ۱- تا ۱ متغیر است، در آن ۱ به معنای همبستگی کامل مثبت است؛ ۰ به‌معنی عدم همبستگی و ۱- به‌معنی همبستگی منفی کامل است. این ضریب که در آمار بسیار مورد استفاده قرار می‌گیرد، توسط کارل پیرسون بر اساس ایده اصلی فرانسیس گالتون گردآوری شده است و برای انتخاب

ویژگی و یافتن همبستگی میان ویژگی‌ها استفاده می‌شود. این شیوه مبتنی بر همبستگی است و یکی از روش‌های رایج انتخاب ویژگی می‌باشد، در این روش، اهمیت یک ویژگی بر حسب قدرت همبستگی بین یک ویژگی و متغیر کلاس اندازه‌گیری می‌شود. زائد بودن یک ویژگی نیز بر اساس قدرت همبستگی بین ویژگی و ویژگی‌های دیگر تعریف می‌شود.

ج) روش F-Score

در تجزیه و تحلیل آماری و طبقه‌بندی باینری، روش F-score با اندازه‌گیری صحت یک آزمون انجام می‌شود. برای محاسبه نمره هم دقت p و هم فراخوان r را در نظر می‌گیرد: p عدد نتایج مثبت صحیح است که بر اساس تعداد کل نتایج مثبت برگردان تقسیم می‌شود و r تعداد نتایج صحیح مثبت تقسیم شده توسط تعداد نمونه‌های مربوطه است (کلیه نمونه‌هایی که می‌بایست مثبت باشند). با این خاصیت می‌توان ارتباط بین ویژگی‌های مناسب را یافت.

د) روش Term Variance

در تئوری و آمار احتمال واریانس امید ریاضی، مربع انحراف یک متغیر تصادفی نسبت به میانگین آن است. در محاسبات مرسوم، این تکنیک خاصیتی را از ویژگی‌ها اندازه‌گیری می‌کند که بر اساس آن، مشخص می‌شود چگونه مجموعه‌ای از اعداد (تصادفی) از مقدار متوسط آن‌ها جدا می‌شوند. واریانس در آمار نقش اساسی دارد، تا جایی‌که برخی از ایده‌هایی که از آن استفاده می‌کنند شامل آمار توصیفی، استنباط آماری، آزمایش فرضیه، خوب بودن نمونه‌گیری و در نهایت انتخاب ویژگی می‌باشد. واریانس ابزاری مهم در علوم یادگیری ماشینی است که هدف اصلی آن تجزیه و تحلیل آماری داده‌ها است. مادامی که بتوان میزان واریانس میان ویژگی‌ها را برآورد کرد، خروجی‌های مناسبی را می‌توان جهت حذف ویژگی‌های با پراکندگی کم بافت استخراج کرد و این منجر به ایجاد مجموعه‌ای غیر تکراری خواهد شد.

۲-۲- طبقه‌بندی

بهترین مدل دینامیکی که بتواند متناسب با نوع داده‌های جدید کارایی داشته باشد، مدل شبکه‌های عصبی است. دلیل این انتخاب آن است که می‌تواند با تنظیم وزن‌ها به مدل‌های کارآمدی رسید. نکته قابل توجه آن است که نمی‌توان پارامترهای تنظیمی دقیقی را در این مدل مفروض و اثرگذار دانست؛ جز تعداد نورون‌های هر لایه و نیز تعداد لایه‌های متناسب با مسئله. به‌همین منظور ما از فرآیند تکرار به‌عنوان یکی از روش‌های گیر انداختن بهترین مدل طبقه‌بند مبتنی بر شبکه‌های عصبی به قسمی استفاده می‌کنیم که بتواند بالاترین سطح دقت را فراهم آورد. تکرار هدفمند است و بر پایه یافتن مدل تصادفی نیست؛ به‌عبارت بهتر آن‌که در ابتدا با تعداد لایه‌های کم و نیز تعداد نورون‌های کم در هر لایه برازش انجام می‌شود و سپس بر تعداد نورون‌ها در هر لایه و بعد تعداد لایه‌ها افزوده می‌شود.

راهکار در عین سادگی بسیار کارآمد است و تکرار با اتکا بر دستیابی به بالاترین سطح دقت صورت می‌پذیرد. مدل پیشنهادی موارد زیر را حل نموده است:

الف) حل مشکل بیش برازش: مدل تنها با استفاده از الگوهای آموزشی شکل بگیرد، اما اگر داده‌ای دیگری که حتی مقدار کمی از مجموعه آموزشی فاصله دارد و به مدل اعمال شود، مدل قادر نیست به درستی پاسخی برای داده‌های جدید بیابد و آن‌ها را با اشتباه زیادی طبقه‌بندی می‌کند.

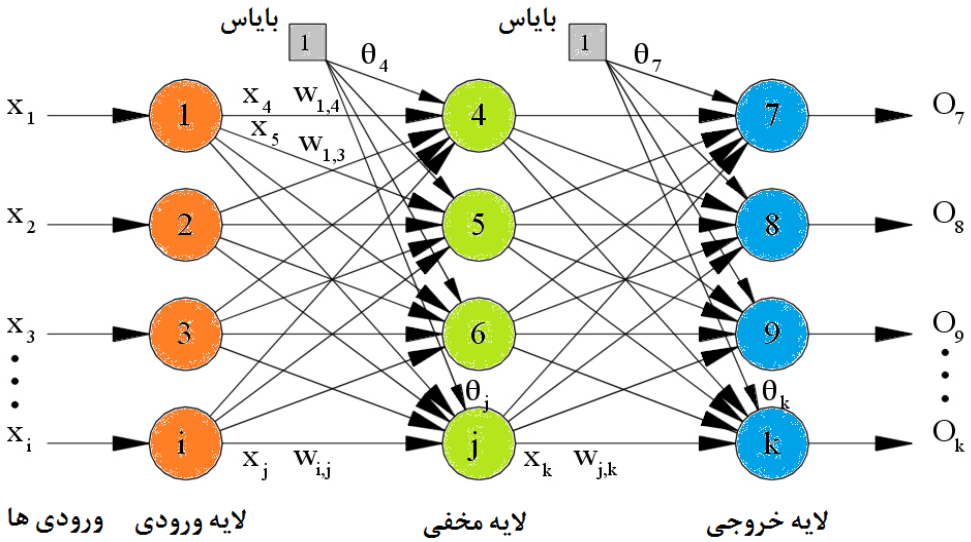
ب) حل مشکل زیر برازش: مدل تنها با استفاده از الگوهای آموزشی شکل بگیرد؛ اما حتی با استفاده از همان داده‌های آموزشی نیز خطا داشته باشد. در واقع الگوریتم، یک مدل کلی از مجموعه آموزشی به دست می‌آورد ولی نمی‌تواند در مورد داده‌هایی که با آنها الگوی تصمیم‌گیری را ساخته، تصمیم‌گیری کند. به‌طور معمول ساختار شبکه‌های عصبی از زیربرازش می‌گریزند.

ج) امکان ذخیره بهترین مدل: به سبب آنکه پس از یافتن بهترین مدل نیازمند به آزمایش هستیم، مدل را ذخیره می‌کنیم تا بتواند با داده‌های جدید تصمیم‌گیری کند.

د) تقسیم‌بندی داده‌ها به سه بخش داده‌های آموزشی، اعتبار و آزمایش: مدل با استفاده از داده‌های آموزشی شکل می‌گیرد و داده‌های اعتبار در بخش اعتبارسنجی متقاطع K-fold استفاده می‌شوند و در نهایت مدل انتخاب شده توسط داده‌های آزمایشی، امتحان می‌شود.

۲-۳- اعتبارسنجی

یکی از بهترین مدل‌های اعتبارسنجی، مدل K-fold یا تقسیم داده‌ها به K بخش مجزاست. در روش اعتبارسنجی K-fold مجموعه داده‌ها را به K بخش مجزا تقسیم می‌کنیم. فرآیند مدل‌سازی را برای K مرتبه تکرار می‌کنیم و در هر مرتبه K-۱ بخش از داده‌ها برای فرآیند آموزش استفاده می‌شود و یک بخش از داده‌ها که در فرآیند آموزش، شرکت داده نشده، برای فرآیند آزمایش و اعتبارسنجی مدل پیش‌بینی کننده، مورد استفاده قرار می‌گیرد. در خاتمه از خطای پیش‌بینی محاسبه شده در هریک از K مرحله متوسط‌گیری می‌شود. مزیت استفاده از زیرمجموعه‌سازی تصادفی داده‌ها در این روش سبب می‌شود تأثیر نحوه توزیع داده‌ها برای فرآیند مدل‌سازی حذف شود. هم در مرحله انتخاب ویژگی و هم طبقه‌بندی از این شیوه استفاده شده و در هر دو مرحله داده‌های آموزشی و آزمایشی و نیز اعتبار استفاده شدند.



شکل ۲- مدل دینامیکی که ساختار نورون و لایه‌ها بر اساس نوع مسئله تغییر می‌کند و انتخاب آنها تصادفی است.

۲-۴- پارامترهای خروجی مدل

معیار پایه‌ای که برای ارزیابی مدل استفاده می‌شود اغلب دقت است که تعداد پیش‌بینی‌های صحیح را در همه پیش بینی‌ها توصیف می‌کند برای داده‌های سرطان مطابق رابطه (۲) محاسبه می‌شود.

$$Acc = \frac{Nc}{N} \quad (۲)$$

که در این رابطه Nc و N به‌ترتیب تعداد پیش‌بینی‌های صحیح به تعداد کل داده‌ها می‌باشد.
صحت[1] معیاری است که نشان می‌دهد چند پیش‌بینی مثبت انجام شده درست هستند (مثبت‌های واقعی). مطابق رابطه (۳) بدست می‌آید.

$$P = \frac{Nc}{NPC} \quad (۳)$$

در این رابطه کمیت NPC، مقادیر پیش‌بینی شده موقعیت‌های ابتلا به سرطان توسط مدل است.
حساسیت[2] معیاری است که نشان می‌دهد طبقه‌بندی‌کننده چه تعداد از موارد مثبت را بر روی همه موارد مثبت در داده‌ها به درستی پیش‌بینی کرده است. رابطه کلی برای بدست آوردن این پارامتر مطابق رابطه (۴) می‌باشد.

$$R = \frac{Nc}{N_{cancer}} \quad (۴)$$

در این رابطه کمیت مخرج تعداد مبتلایان به سرطان در داده‌ها را نشان می‌دهد.
F1-Score معیاری است که صحت و حساسیت را با هم ترکیب می‌کند. به‌طور کلی به‌عنوان میانگین هارمونیک این دو توصیف می‌شود. میانگین هارمونیک روش دیگری برای محاسبه میانگین مقادیر است که اغلب برای نسبت‌ها (مانند صحت و حساسیت) مناسب‌تر از میانگین حسابی سنتی توصیف می‌شود. فرمول مورد استفاده برای F1 در رابطه (۵) نشان داده شده است.

$$F1 = 2 * \frac{P * R}{P + R} \quad (۵)$$

هدف آن است که یک معیار واحد ارائه شود که دو نسبت (صحت و حساسیت) را به‌روشی متعادل وزن کند و برای افزایش ارزش امتیاز F1، هر دو باید مقدار بیشتری داشته باشند.

---

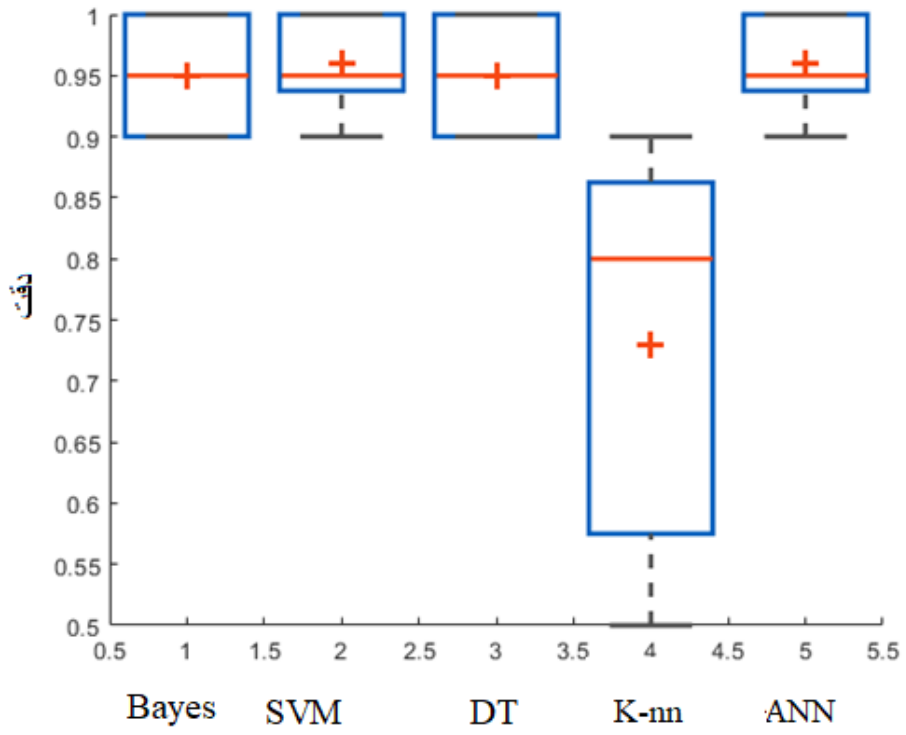
[1] Precision
[2] Recall

۳- بررسی نتایج

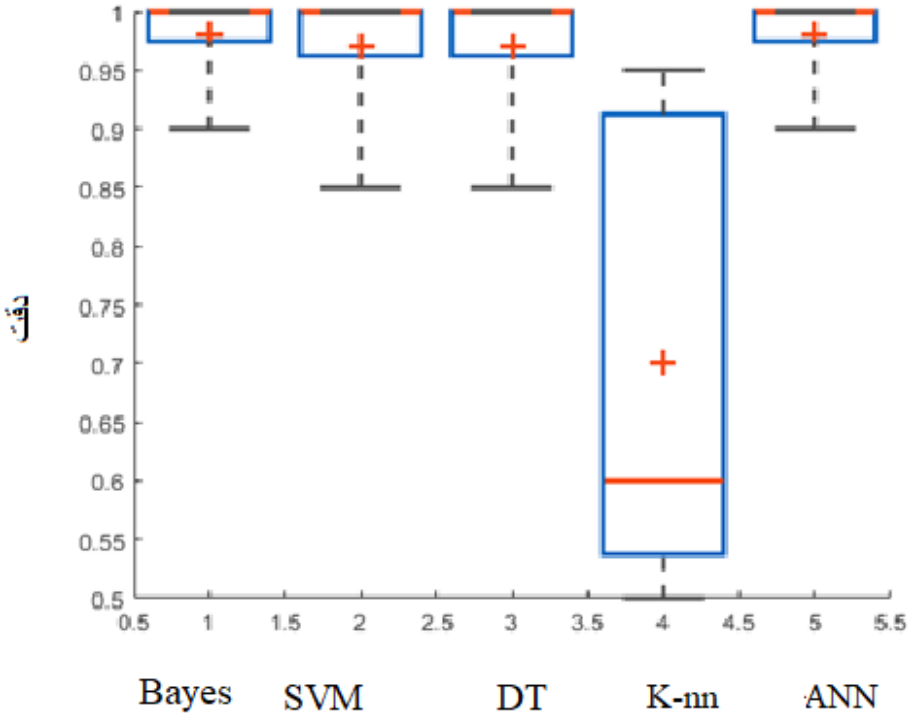
در این بخش خروجی‌های بدست آمده بررسی شده‌اند. در ابتدا نتایج اعتبارسنجی انجام شده بر اساس روش k-fold گردآوری شده‌اند.

جدول ۳- تنظیمات با کی فولد برابر با ۵ در طبقه‌بندی و شبکه‌های عصبی چهارگانه

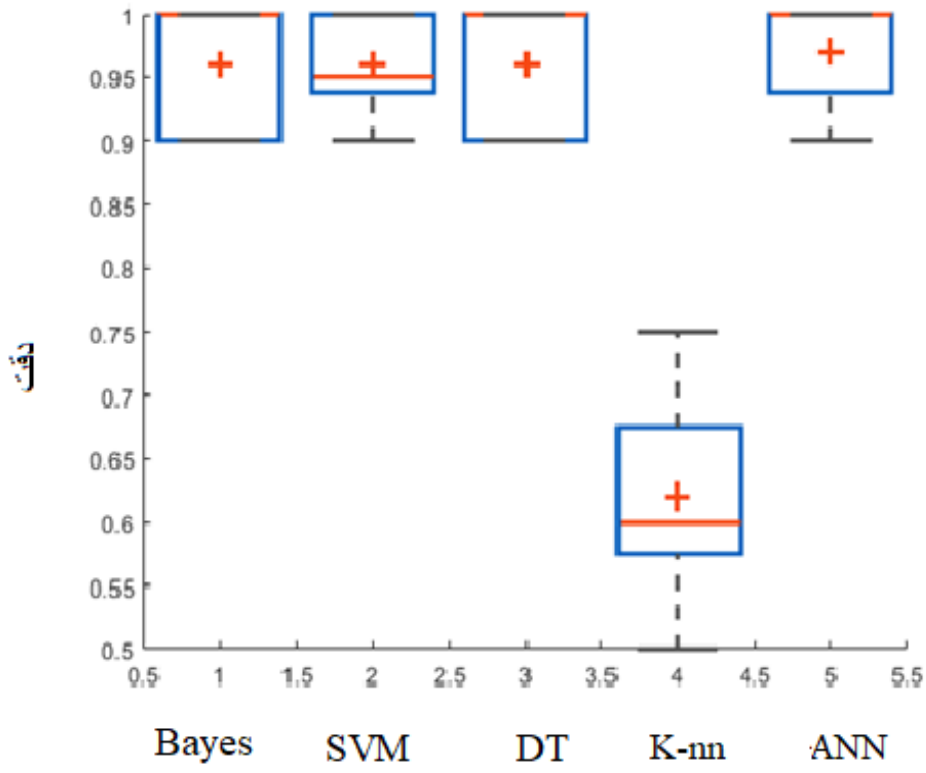
	تعداد ویژگی‌ها	۲	۳	۴	۵	۷	۱۲	۱۴	۲۱	۲۵
	موفقیت در مقایسه با سایر طبقه‌بندها	خیر	بلی	بلی	بلی	خیر	خیر	بلی	بلی	بلی
CV=5	دقت طبقه‌بندی در ۱۰ بار تکرار شبکه‌ها	٪ ۹۳	٪ ۹۶	٪ ۹۸	٪ ۹۷/۳	٪ ۹۶	٪ ۹۶	٪ ۹۷/۶	٪ ۹۵	٪ ۹۶
	دقت طبقه‌بندی در ۲۰ بار تکرار شبکه‌ها	٪ ۹۵	٪ ۹۶	٪ ۹۷	٪ ۹۶	٪ ۹۸	٪ ۹۸	٪ ۹۹	٪ ۹۴	٪ ۹۵
CV=10	دقت طبقه‌بندی در ۱۰ بار تکرار شبکه‌ها	٪ ۹۴	٪ ۹۵	٪ ۹۶	٪ ۹۸	٪ ۹۷	٪ ۹۵	٪ ۹۶	٪ ۹۴	٪ ۹۵
	دقت طبقه‌بندی در ۲۰ بار تکرار شبکه‌ها	٪ ۹۴	٪ ۹۶	٪ ۹۷	٪ ۹۶	٪ ۹۸	٪ ۹۸	٪ ۹۸	٪ ۹۷	٪ ۹۵
	ویژگی‌های انتخاب شده	-۳۲۰۰ ۴۸۲۳	-۳۲۰۰ -۴۴۴۷ ۴۸۲۳	-۴۸۲۳ -۷۴۵۱ -۷۲۲۹ -۸۵۴۵	-۷۴۵۱ -۴۸۲۳ -۸۷۶۵ -۳۲۰۰ ۷۵۱۵	-۳۶۹۹ -۳۰۷۷ -۶۱۴۴ -۸۲۲۰ -۱۰۳۲۴ -۸۵۳۶ ۷۰۶۱	-۸۲۲۰ -۴۸۲۳ -۲۷۵ -۳۰۵۹ -۳۰۷۷ -۳۶۹۹ -۴۴۴۷ -۷۰۶۱ -۷۴۵۱ -۴۷۰۰ -۵۴۶۱ ۶۱۴۴	-۴۸۲۳ -۷۴۵۱ -۸۷۶۵ -۸۴۶۸ -۷۵۱۵ -۲۷۱۸ -۶۸۲۱ -۷۵۳۱ -۱۰۱۳۰ -۵۸۱۵ -۱۲۰ -۷۶۵۲ ۹۹۴۹	-۸۲۲۰ -۴۸۲۳ -۷۳۷۲ -۷۶۵۲ -۳۹۸ -۹۳۸ -۱۵۱۴ -۳۶۷۳ -۳۷۷۴ -۴۰۱۱ -۴۴۴۷ -۴۷۰۰ -۴۹۸۶ -۵۲۳۷ -۵۶۴۸ -۷۵۳۱ -۲۷۱۴ -۳۲۰۰ -۷۳۴۶ ۷۴۵۱	-۸۲۲۰ -۴۸۲۳ -۷۳۷۲ -۷۶۵۲ -۳۹۸ -۹۳۸ -۱۵۱۴ -۳۶۷۳ -۳۷۷۴ -۴۰۱۱ -۴۴۴۷ -۴۷۰۰ -۴۹۸۶ -۵۲۳۷ -۵۶۴۸ -۷۵۳۱ -۲۷۱۴ -۳۲۰۰ -۷۳۴۶ ۷۴۵۱



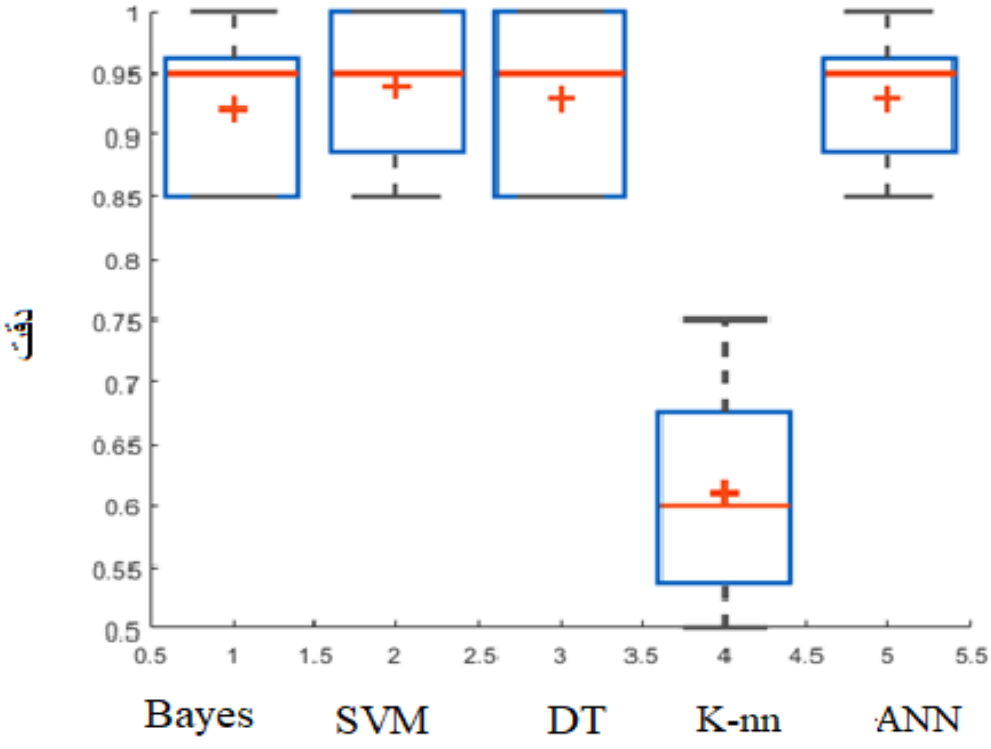
شکل ۳– نتیجه اعمال ۳ ویژگی و عملکرد طبقه‌بندهای مختلف؛ از چپ به راست، به ترتیب، کلاسیفایرهای بیز، اس وی ام، درخت تصمیم، کی ان ان و شبکه عصبی دینامیکی به روش وزن‌دهی با CV=10 و تکرار اول



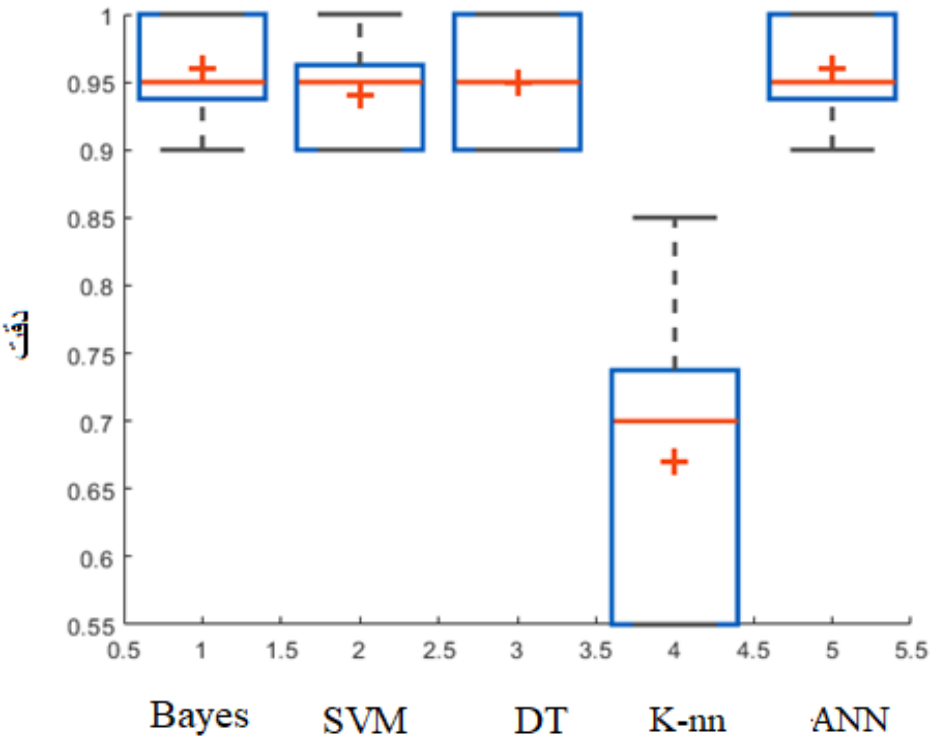
شکل ۴– نتیجه اعمال ۴ ویژگی و عملکرد طبقه‌بندهای مختلف؛ از چپ به راست به ترتیب، کلاسیفایرهای بیز، اس وی ام، درخت تصمیم، کی ان ان و شبکه عصبی دینامیکی به روش وزن‌دهی با CV=10 و تکرار دوم



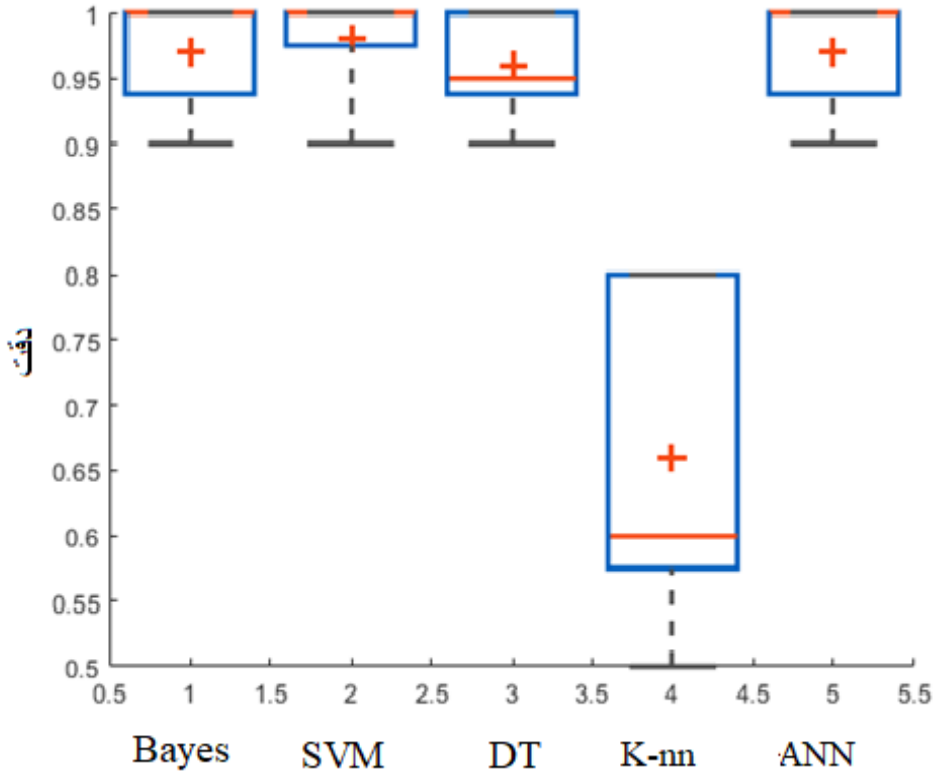
شکل ۵- نتیجه اعمال ۵ ویژگی و عملکرد طبقه‌بندهای مختلف؛ از چپ به راست، به ترتیب، کلاسیفایرهای بیز، اس وی ام، درخت تصمیم، کی ان ان و شبکه عصبی دینامیکی به روش وزن‌دهی با CV=10 و تکرار سوم



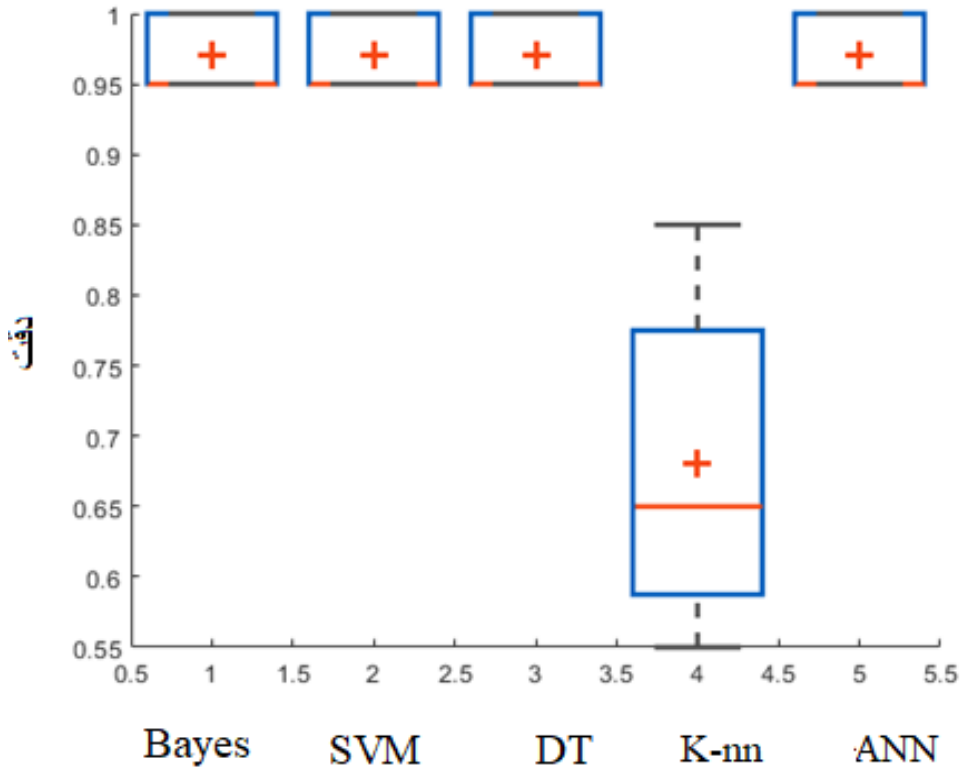
شکل ۶- نتیجه اعمال ۷ ویژگی و عملکرد طبقه‌بندهای مختلف؛ از چپ به راست، به ترتیب، کلاسیفایرهای بیز، اس وی ام، درخت تصمیم، کی ان ان و شبکه عصبی دینامیکی به روش وزن‌دهی با CV=10 و تکرار چهارم



شکل ۷– نتیجه اعمال ۱۲ ویژگی و عملکرد طبقه‌بندهای مختلف؛ از چپ به راست، به ترتیب، کلاسیفایرهای بیز، اس وی ام، درخت تصمیم، کی ان ان و شبکه عصبی دینامیکی به روش وزن‌دهی با CV=10 و تکرار پنجم



شکل ۸– نتیجه اعمال ۱۵ ویژگی و عملکرد طبقه‌بندهای مختلف؛ از چپ به راست، به ترتیب، کلاسیفایرهای بیز، اس وی ام، درخت تصمیم، کی ان ان و شبکه عصبی دینامیکی به روش وزن‌دهی با CV=10 و تکرار ششم



شکل ۹– نتیجه اعمال ۲۵ ویژگی و عملکرد طبقه‌بندهای مختلف؛ از چپ به راست، به ترتیب، کلاسیفایرهای بیز، اس وی ام، درخت تصمیم، کی ان ان و شبکه عصبی دینامیکی به روش وزن‌دهی با CV=10 و تکرار هفتم

نتایج مربوط به بیماری پروستات است که ما بر روی مجموعه ژن‌های آن پیاده‌سازی کردیم و به تعداد ۷ بار تکرار صورت پذیرفته است. شکل‌های (۳) تا (۹) منحنی مشخصه عملکرد[1] (ROC) را نشان می‌دهند. منحنی مشخصه عملکرد گیرنده (ROC) یک منحنی گرافیکی از طرح‌ریزی نرخ مثبت واقعی در مقابل نرخ مثبت کاذب برای یک مدل طبقه‌بندی باینری است. این منحنی برای زمانی است که آستانه تصمیم‌گیری مدل متفاوت است. محوطه زیر منحنی ROC (AUC)، معیاری است برای اینکه یک پارامتر چقدر می‌تواند در تمایز بین دو کلاس خوب باشد. هر چه AUC بزرگتر باشد، خطای طبقه‌بندی کمتر است. در برنامه رتبه‌بندی ویژگی، ویژگی‌های با بالاترین AUC انتخاب می‌شود.

در این شکل‌ها، روش یادگیری مبتنی بر شبکه عصبی دینامیک در طبقه‌بندی و نیز مدل انتخابگر ژن‌ها در یافتن بهترین زیر مجموعه از ویژگی‌های انتخاب شده از سایر روش‌ها عملکرد بهتری داشته است. نخست آنکه روش‌های دیگر دارای دامنه تغییرات ناحیه انتخاب ویژگی بالاتری هستند و قابل مقایسه است که نه تنها نقاط بیشینه و کمینه نمودارها از هم فاصله زیادی دارند، بلکه میانگین مقادیر دقت خروجی کمتری را عرضه نموده‌اند. به‌عبارت بهتر، روش‌های دیگر دارای ناحیه گسترده‌تر و بزرگتری هستند، زیرا قسمت جعبه آن‌ها بزرگتر از جعبه روش پیشنهادی است.

با مقایسه دامنه میان چارک یا ارتفاع جعبه نیز می‌توان فهمید که پراکندگی میانگین سطوح دقت طبقه‌بندی بسته به مرحله انتخاب ویژگی و طبقه‌بندی برای روش‌های دیگر در قیاس با روش پیشنهادی انتخاب ویژگی

[1]Receiver operating characteristic

بالاتر و نامناسب‌تر است؛ زیرا ارتفاع جعبه روش پیشنهادی انتخاب ویژگی و طبقه‌بندی به طور غالب در داده‌های بیان ژن مورد نظر بزرگتر از ارتفاع جعبه روش‌های دیگر است. حتی روش پیشنهادی دارای مقادیر پرت (Outlier AUC) کمتری در طبقه‌بندی است. هر چند برای هر میانگین سطح دقت روش‌ها، مقادیر پرت شناسایی شده ولی به نظر می‌رسد که برای یافتن بهترین طبقه‌بندی، مقادیر پرت روش‌های دیگر در قیاس با روش شبکه عصبی دینامیک بیشتر است.

جدول (۴) نیز نتایج مرحله آزمایش داده‌های لوسمی را نشان می‌دهد که شاخصه‌های طبقه‌بندی خوبی را ایجاد می‌کند.

جدول (۵) نیز نتایج مرحله آزمایش داده‌های DLBCL را نشان می‌دهد که شاخصه‌های طبقه‌بندی خوبی را ایجاد می‌کند.

جدول (۶) گزارش پارامترهای صحت، حساسیت و F1-Score سه داده سرطان پروستات، سینه و DLBCL برای تعداد ۲۵ ژن انتخابی را نشان می‌دهد.

با توجه به جدول (۶) مقادیر بدست آمده قابل قبول بوده و مدل بدست آمده برای سه پارامتر بدست آمده در داده‌های سرطان سینه با روش طبقه‌بندی شبکه عصبی دینامیکی از کارآمدی بالایی برخوردار است.

شکل (۱۰) نیز مقایسه دقت روش‌های بیان تئوری بیز، شبکه بردار پشتیبان، درخت تصمیم، شبکه عصبی دینامیکی و K نزدیک‌ترین همسایگی را نشان می‌دهد که شاخصه‌های طبقه‌بندی خوبی را ایجاد می‌کند.

در شکل (۱۰) مشاهده می‌شود که روش NN دارای بهترین دقت و انطباق است.

جدول ۴- دسته‌بندی مربوط به داده‌های بیماری لوسومی

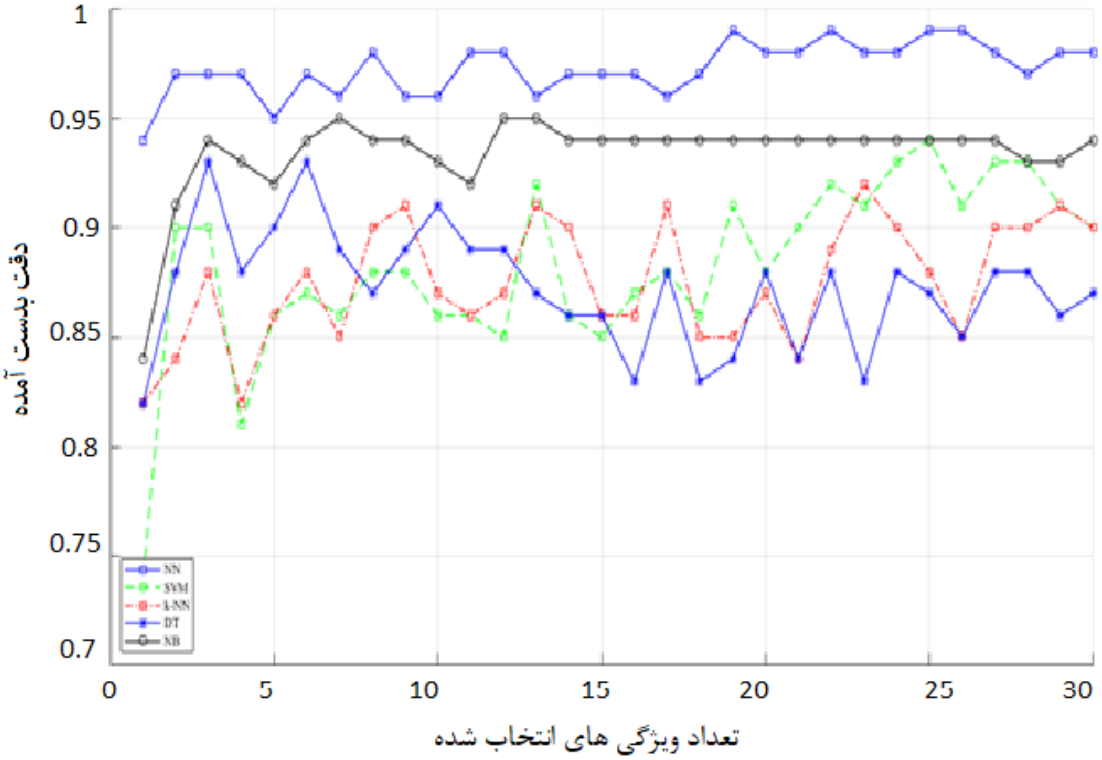
	تعداد ویژگی‌ها	۲	۳	۴	۵	۷	۱۲	۱۴	۲۱	۲۵
CV=5	دقت طبقه‌بندی در ۱۰ بار تکرار شبکه‌ها	٪۹۷	٪۹۸	٪۱۰۰	٪۱۰۰	٪۹۸/۵۷	٪۹۸	٪۹۷	٪۹۶	٪۹۷
	دقت طبقه‌بندی در ۲۰ بار تکرار شبکه‌ها	٪۹۹	٪۹۹	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۹۹	٪ ۹۹	٪۹۷	٪۹۸
CV=10	دقت طبقه‌بندی در ۱۰ بار تکرار شبکه‌ها	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۹۹	٪۹۸	٪۹۷	٪۹۸
	دقت طبقه‌بندی در ۲۰ بار تکرار شبکه‌ها	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۱۰۰	٪۹۸	٪۹۶	٪۹۸

جدول ۵- دسته‌بندی مربوط به داده‌های بیماری DLBCL

	تعداد ویژگی‌ها	۲	۳	۴	۵	۷	۱۲	۱۴	۲۱	۲۵
CV=5	دقت طبقه‌بندی در ۱۰ بار تکرار شبکه‌ها	٪۹۴	٪۹۴/۶۷	٪۹۳	٪۹۵	٪۹۶	٪۹۶	٪۹۵	٪۹۵	٪۹۴
	دقت طبقه‌بندی در ۲۰ بار تکرار شبکه‌ها	٪۹۵	٪۹۵	٪۹۴	٪۹۵	٪۹۵	٪۹۶	٪۹۴	٪۹۵	٪۹۲

جدول ۶– گزارش شاخص‌های آماری برای سه داده سرطان

نام بیماری	روش طبقه‌بندی	صحت	حساسیت	F1
سرطان سینه	بیز	۰/۸۲	۰/۷۹	۰/۸
DLBCL	بیز	۰/۸	۰/۸۲	۰/۸۱
سرطان پروستات	بیز	۰/۸۲	۰/۸۲	۰/۸۲
سرطان سینه	اس وی ام	۰/۷۱	۰/۷۶	۰/۷۹
DLBCL	اس وی ام	۰/۷۴	۰/۶۹	۰/۷۱
سرطان پروستات	اس وی ام	۰/۷۷	۰/۸۴	۰/۸
سرطان سینه	درخت تصمیم	۰/۸۱	۰/۸۱	۰/۸۱
DLBCL	درخت تصمیم	۰/۸۱	۰/۸۳	۰/۸۲
سرطان پروستات	درخت تصمیم	۰/۸۲	۰/۸۱	۰/۸۲
سرطان سینه	کی ان ان	۰/۷۴	۰/۶۹	۰/۷۱
DLBCL	کی ان ان	۰/۷۹	۰/۸۲	۰/۸۱
سرطان پروستات	کی ان ان	۰/۷۵	۰/۷۸	۰/۷۶
سرطان سینه	شبکه عصبی دینامیکی	۰/۹۳	۰/۸۷	۰/۹۱
DLBCL	شبکه عصبی دینامیکی	۰/۸۱	۰/۸۳	۰/۸۲
سرطان پروستات	شبکه عصبی دینامیکی	۰/۸۲	۰/۷۹	۰/۸



شکل ۱۰– مقایسه میان راهکارهای طبقه‌بندی

۴- نتیجه‌گیری

در این مقاله:

• طبقه‌بندی بیان ژن از سه نوع داده سرطان کولون، سرطان سینه، لوسمی، تومورهای پروستات و DLBCL استفاده می‌شود و هر کدام به‌طور جداگانه در چرخه انتخاب ویژگی و هم‌چنین طبقه‌بندی با تعداد ویژگی‌های متغیر وارد می‌شوند.

• در مرحله طبقه‌بندی از ۴ دسته خانواده عصبی شامل شبکه عصبی پویا، شبکه ماشین بردار پشتیبان، نظریه بیز، K نزدیکترین همسایه و در نهایت درخت تصمیم استفاده می‌شود.

• انتخاب ویژگی نیز بر اساس رأی اکثریت با استفاده از روش‌های Relife، PCC، F-Score و Term Variance است.

• برای اعتبارسنجی از مدل k-fold یا تقسیم داده‌ها به k قسمت مجزا از روش NN استفاده شد.

• روش NN بهترین دقت و عملکرد را در بین روش‌های طبقه‌بندی برای هر چهار نوع داده داشت.

مراجع

[1] A. D. Gordon, "Classification," 2nd Ed., CRC Press, USA, 1999, ISBN: 1584888539. https://books.google.com/books?id=_w5AJtbfEz4C.

[2] L. Boguslawski, "Influence of Pressure Fluctuations Distribution on Local Heat Transfer on Flat Surface Impinged by Turbulent Free Jet," in *Thermal Sciences 2004, Proceedings of the ASME-ZSIS International Thermal Science Seminar II*, 2004, doi: 10.1615/ICHMT.2004.IntThermSciSemin.

[3] R. G. Brereton, and G. R. Lloyd, "Support Vector Machines for Classification and Regression," *Analyst*, Vol. 135, No. 2, pp. 230-267, 2010, doi: 10.1039/B918972F.

[4] J. R. Quinlan, "Programs for Machine Learning," *Morgan Kaufmann Publishers*, Inc., *Elsevier, San Mateo, California*, 2014, ISBN: 0080500587, https://books.google.com/books?id=b3ujBQAAQBAJ.

[5] W.-Y. Loh, "Classification and Regression Trees," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 1, No. 1, pp. 14-23, 2011, First Edition, John Wiley & Sons, Inc., New Jersey, USA, doi: 10.1002/widm.8.

[6] K. P. Murphy, "Naive Bayes Classifiers," *University of British Columbia*, Vol. 18, No. 60, pp. 1-8, 2006, http://www.ic.unicamp.br/rocha/teaching/2011s1/mc906/aulas/naivebayes.pdf.

[7] L. Van, D. Maaten, E. Postma, and J. Van den Herik, "Dimensionality Reduction: A Comparative," *J Mach Learn Res*, Vol. 10, No. 66-71, p. 13, 2009. [Online] Available: https://members.loria.fr/moberger/Enseignement/AVR/Exposes/.

[8] L. Nanni, and A. Lumini, "Wavelet Selection for Disease Classification by DNA Microarray Data," *Expert Systems with Applications,* Vol. 38, No. 1, pp. 990-995, 2011, doi: 10.1016/j.eswa.2010.07.104.

[9] I. Porto-Diaz, V. Bolon-Canedo, A. Alonso-Betanzos, and O. Fontenla-Romero, "A Study of Performance on Microarray Data Sets for a Classifier Based on Information Theoretic Learning," *Neural Networks,* Vol. 24, No. 8, pp. 888-896, 2011, doi: 10.1016/j.neunet.2011.05.010.

[10] M. Shah, M. Marchand, and J. Corbeil, "Feature Selection with Conjunctions of Decision Stumps and Learning from Microarray Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 34, No. 1, pp. 174-186, 2011, doi: https://doi.org/10.1109/TPAMI.2011.82.

[11] F. V. Sharbaf, S. Mosafer, and M. H. Moattar, "A Hybrid Gene Selection Approach for Microarray Data Classification using Cellular Learning Automata and Ant Colony Optimization," *Genomics*, Vol. 107, No. 6, pp. 231-238, 2016, doi: 10.1016/j.ygeno.2016.05.001.

[12] Y. Leung, and Y. Hung, "A Multiple-filter-multiple-wrapper Approach to Gene Selection and Microarray Data Classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* Vol. 7, No. 1, pp. 108-117, 2008, doi: 10.1109/TCBB.2008.46.

[13] Y. Leung, and Y. Hung, "A Multiple-filter-multiple-wrapper Approach to Gene Selection and Microarray Data Classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* Vol. 7(1), pp. 108-117, 2008, doi: 10.1109/TCBB.2008.46.

[14] S. Maldonado, R. Weber, and F. Famili, "Feature Selection for High-dimensional Class-imbalanced Data Sets using Support Vector Machines," *Information Sciences,* Vol. 286, pp. 228-246, 2014, doi: 10.1016/j.ins.2014.07.015.

[15] A. Castaño, F. Fernández-Navarro, C. Hervás-Martínez, and P.A. Gutiérrez, "Neuro-logistic Models Based on Evolutionary Generalized Radial Basis Function for the Microarray Gene Expression Classification Problem," *Neural Processing Letters,* Vol. 34, pp. 117-131, 2011, doi: 10.1007/s11063-011-9187-8.

[16] J. Li, Y. Jia, and W. Li, "Adaptive Huberized Support Vector Machine and its Application to Microarray Classification," *Neural Computing and Applications,* Vol. 20, pp. 123-132, 2011, doi: 10.1007/s00521-010-0371-y.

[17] K.-H. Liu, Z.-H. Zeng, and V. T. Y. Ng, "A Hierarchical Ensemble of ECOC for Cancer Classification Based on Multi-class Microarray Data," *Information Sciences,* Vol. 349, pp. 102-118, 2016, doi: 10.1016/j.ins.2016.02.028.

[18] S. Karimi, and M. Farrokhnia, "Leukemia and Small Round Blue-cell Tumor Cancer Detection using Microarray Gene Expression Data Set: Combining Data Dimension Reduction and Variable Selection Technique," *Chemometrics and Intelligent Laboratory Systems,* Vol. 139, pp. 6-14, 2014, doi: 10.1016/j.chemolab.2014.09.003.

[19] H. Liu, L. Liu, and H. Zhang, "Ensemble Gene Selection by Grouping for Microarray Data Classification," *Journal of Biomedical Informatics,* Vol. 43, No. 1, pp. 81-87, 2010, doi: 10.1016/j.jbi.2009.08.010.

[20] P. Maji, "Fuzzy–rough Supervised Attribute Clustering Algorithm and Classification of Microarray Data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* Vol. 41, No. 1, pp. 222-233, 2010, doi: 10.1109/TSMCB.2010.2050684.

[21] C.-P. Lee, W.-S. Lin, Y.-M. Chen, and B.-J. Kuo, "Gene Selection and Sample Classification on Microarray Data Based on Adaptive Genetic Algorithm/k-nearest Neighbor Method," *Expert Systems with Applications,* Vol. 38, No. 5, pp. 4661-4667, 2011, doi: 10.1016/j.eswa.2010.07.053.

[22] D. Hernández-Lobato, J. M. Hernández-Lobato, and A. Suárez, "Expectation Propagation for Microarray Data Classification," *Pattern Recognition Letters,* Vol. 31, No. 12, pp. 1618-1626, 2010, doi: 10.1016/j.patrec.2010.05.007.

[23] P. Mohapatra, S. Chakravarty, and P. Dash, "Microarray Medical Data Classification using Kernel Ridge Regression and Modified Cat Swarm Optimization Based Gene Selection System," *Swarm and Evolutionary Computation,* Vol. 28, pp. 144-160, 2016, doi: 10.1016/j.swevo.2016.02.002.

[24] J. M. Cadenas, M. C. Garrido, and R. MartíNez, "Feature Subset Selection Filter–wrapper Based on Low Quality Data," *Expert Systems with Applications,* Vol. 40, No. 16, pp. 6241-6252, 2013, doi: 10.1016/j.eswa.2013.05.051.

[25] H. Deng, and G. Runger, "Gene Selection with Guided Regularized Random Forest," *Pattern Recognition,* Vol. 46, No. 12, pp. 3483-3489, 2013, doi: 10.1016/j.patcog.2013.05.018.

[26] F. Fernández-Navarro, C. Hervás-Martínez, R. Ruiz, and J.C. Riquelme, "Evolutionary Generalized Radial Basis Function Neural Networks for Improving Prediction Accuracy in Gene Classification using Feature Selection," *Applied Soft Computing,* Vol. 12, No. 6, pp. 1787-1800, 2012, doi: 10.1016/j.asoc.2012.01.008.

[27] S. Tabakhi, A. Najafi, R. Ranjbar, and P. Moradi, "Gene Selection for Microarray Data Classification using a Novel Ant Colony Optimization," *Neurocomputing,* Vol. 168, pp. 1024-1036, 2015, doi: 10.1016/j.neucom.2015.05.022.

[28] M. Czajkowski, M. Grześ, and M. Kretowski, "Multi-test Decision Tree and its Application to Microarray Data Classification," *Artificial Intelligence in Medicine,* Vol. 61, No. 1, pp. 35-44, 2014, doi: 10.1016/j.artmed.2014.01.005.

[29] H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue, and Z. Gao, "A Hybrid Feature Selection Algorithm for Gene Expression Data Classification," *Neurocomputing,* Vol. 256, pp. 56-62, 2017, doi: 10.1016/j.neucom.2016.07.080.

[30] K. Lee, Z. Man, D. Wang, and Z. Cao, "Classification of Bioinformatics Dataset using Finite Impulse Response Extreme Learning Machine for Cancer Diagnosis," *Neural Computing and Applications,* Vol. 22, pp. 457-468, 2013, doi: 10.1007/s00521-012-0847-z.

[31] Z. Liu, D. Tang, Y. Cai, R. Wang, and F. Chen, "A Hybrid Method Based on Ensemble WELM for Handling Multi Class Imbalance in Cancer Microarray Data," *Neurocomputing*, Vol. 266, pp. 641-650, 2017, doi: 10.1016/j.neucom.2017.05.066.

[32] L. Fan, K.-L. Poh, and P. Zhou, "Partition-conditional ICA for Bayesian Classification of Microarray Data," *Expert Systems with Applications*, Vol. 37, No. 12, pp. 8188-8192, 2010, doi: 10.1016/j.eswa.2010.05.068.

[33] A. Wang, N. An, G. Chen, L. Li, and G. Alterovitz, "Improving PLS–RFE Based Gene Selection for Microarray Data Classification," *Computers in Biology and Medicine*, Vol. 62, pp. 14-24, 2015, doi: 10.1016/j.compbiomed.2015.04.011.

[34] C. Bielza, V. Robles, and P. Larrañaga, "Regularized Logistic Regression without a Penalty Term: An Application to Cancer Classification with Microarray Data," *Expert Systems with Applications*, Vol. 38, No. 5, pp. 5110-5118, 2011, doi: 10.1016/j.eswa.2010.09.140.

[35] M.-Y. Wu, D.-Q. Dai, Y. Shi, H. Yan, and X.-F. Zhang, "Biomarker Identification and Cancer Classification Based on Microarray Data using Laplace Naive Bayes Model with Mean Shrinkage," IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 9, No. 6, pp. 1649-1662, 2012, doi: 10.1109/TCBB.2012.105.

[36] S. Monti, K. Savage, J. L. Kutok, F. Feuerhake, P. Kurtin, M. Mihm, B. Wu, L. Pasqualucci, D. Neuberg, R. C. T. Aguiar, P. Dal Cin, C. Ladd, G. S. Pinkus, G. Salles, N. Lee Harris, R. Dalla-Favera, T. M. Habermann, J. C. Aster, T. R. Golub, and M. A. Shipp, "Molecular Profiling of Diffuse Large B Cell Lymphoma Reveals a Novel Disease Subtype with Brisk Host Inflammatory Response and Distinct Genetic Features," *Blood*, Vol. 105, No. 5, pp. 1851-1861, 2005, https://doi.org/10.1182/blood-2004-07-2947.

[37] T. R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, Vol. 286, No. 5439, pp. 531-537, 1999, doi: 10.1126/science.286.5439.531.

[38] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P, W. Kantoff, T. R. Golub, and W. R. Sellers, "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, Vol. 1, No. 2, pp. 203-209, 2002, doi: 10.1016/S1535-6108(02)00030-2.

[39] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine, and A. J. Levine, "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proceedings of the National Academy of Sciences*, Vol. 96, No. 12, pp. 6745-6750, 1999, doi: 10.1073/pnas.96.12.6745.

[40] N. Matamala, M.T. Vargas, R. Gonzalez-Campora, R. Minambres, J. I. Arias, P. Menendez, E. Andrés-León, G. Gómez-López, K. Yanowsky, J. Calvete-Candenas, L. Inglada-Pérez, B. Martínez-Delgado, and J. Benítez, "Tumor MicroRNA Expression Profiling Identifies Circulating MicroRNAs for Early Breast Cancer Detection," *Clinical Chemistry*, Vol. 61, No. 8, pp. 1098-1106, 2015, doi: 10.1373/clinchem.2015.238691.

Classification of Five Types of Cancer Data Based on Neural Network Methods, and Analysis of Gene Expression using the Feature Selection Fusion Method

Farnoosh Turki
Ph.D. Candidate, Faculty of Mechanical Engineering, K. N. Toosi University of Technology, Tehran, Iran
farnoosh.turki@email.kntu.ac.ir

Atefeh Khadem
M.Sc. Graduate, Faculty of Mechanical Engineering, K. N. Toosi University of Technology, Tehran, Iran
atefeh.khadem@gmail.com

*Corresponding author: **Abdolhossein Jalali Aghchai**
Associate Professor, Faculty of Mechanical Engineering, K. N. Toosi University of Technology, Tehran, Iran
jalali@kntu.ac.ir

Abstract

In high-volume microarray data, the small number of samples and inherent variability in biological processes cause the problem of increasing computational cost and complexity of classifications. Also, the interpretation of disease-causing genes is complicated, because biologically, only a small set of genes can describe the disease more accurately. The first step in the analysis of microarray data is to significantly reduce the number of genes, or in other words, to select discriminating genes in the classification process. This step is called gene selection. In this article, gene expression classification of three types of data, colon cancer, breast cancer, leukemia, prostate tumors, and DLBCL are used, and each of them is separated in the feature selection cycle and also categorized with variable number of features. Results show the NN method had the best accuracy and performance among the classification methods for all four types of data.

Keywords: Data classification, Machine learning, Neural network, Cancer analysis