Published by: Iranian Society for Biomedical Engineering / www.isbme.ir

Iranian Journal of Biomedical Engineering

www.ijbme.org / P-ISSN: 2008-5869 / E-ISSN: 8006-9685

Volume 13, Issue 3, Autumn 2019, 223 - 234

Feature Selection Method based on Subspace Learning and Factorization of Basis Matrix for DNA Micro-Array Datasets

Dehtaghi Zadeh, Mahla [1] / Saberi-Movahed, Farid [2*] / Eftekhari, Mehdi [3]

[1] - M.Sc. Student, Department of Applied Mathematics, Faculty of Sciences and Modern Technologies, Graduate University of Advanced Technology, Kerman, Iran

[2] - Assistant Professor, Department of Applied Mathematics, Faculty of Sciences and Modern Technologies, Graduate University of Advanced Technology, Kerman, Iran

[3] - Associate Professor, Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

ARTICLE INFO

DOI: 10.22041/IJBME.2019.104143.1454

Received: 23 February 2019 Revised: 7/6/2019-29/9/2019 Accepted: 30 September 2019

KEYWORDS

Feature Selection

Subspace Learning

Matrix Factorization

DNA Micro-Array Datasets

ABSTRACT

DNA micro-array datasets play crucial role in machine learning and recognition of various kinds of cancer structures. Micro-array datasets are typically characterized by the high number of features and the small number of samples. Such problems may result in overfitting and low prediction accuracy of classifiers due to the irrelevant features, and therefore, they are considered as a challenging task in machine learning. The direct way to deal with such challenges is dimensionality reduction of data. In this regard, feature selection method acts as an effective solution for dimensinality reduction and increasing efficiency of learning algorithms. In this paper, by using the concept of "the basis for the DNA micro-array datasets", a new feature selection method is introduced. To be more specific, rather than utilizing the entire micro-array dataset for tackling the problem of feature selection, a basis that is a muchmore smaller subset of the micro-array dataset is used. This method is based on subspace learning and matrix factorization. Finally, by making use of the DNA micro-array datasets, the effectiveness of the proposed method is evaluated, and the obtained results are compared with some state-of-the-art supervised feature selection methods.

[*]Corresponding Author

Address	Department of Mathematics, Faculty of Sciences and Modern Technologies, Graduate University of Advanced Technology, Kerman, Iran		
Postal Code	76315-117	**Tel**	*+98-34-33776611*
E-Mail	*f.saberimovahed@kgut.ac.ir*	**Fax**	*+98-34-33776611*

ناشر: انجمن مهندسی پزشکی ایران / www.isbme.ir

مجله‌ی مهندسی پزشکی زیستی

شاپای چاپی: ۵۸۶۹-۲۰۰۸ / شاپای الکترونیکی: ۹۶۸۵-۸۰۰۶ / www.ijbme.org

دوره: ۱۳، شماره: ۳، پاییز ۱۳۹۸، ۲۲۳ - ۲۳۴

روش انتخاب ویژگی بر اساس یادگیری زیرفضا و تجزیه‌ی ماتریس پایه برای داده‌های میکرو-آرایه‌ای DNA

ده‌تقی‌زاده، مهلا [۱] / صابری موحد، فرید [۲*] / افتخاری، مهدی [۳]

[۱] – دانشجوی کارشناسی ارشد، گروه ریاضی کاربردی، دانشکده‌ی علوم و فناوری‌های نوین، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته، کرمان، ایران

[۲] – استادیار، گروه ریاضی کاربردی، دانشکده‌ی علوم و فناوری‌های نوین، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته، کرمان، ایران

[۳] – دانشیار، دانشکده‌ی مهندسی کامپیوتر، دانشگاه شهید باهنر کرمان، کرمان، ایران

مشخصات مقاله

شناسه‌ی دیجیتال: 10.22041/IJBME.2019.104143.1454

ثبت در سامانه: ۴ اسفند ۱۳۹۷ بازنگری: ۱۳۹۸/۳/۱۷-۱۳۹۸/۷/۷ پذیرش: ۸ مهر ۱۳۹۸

واژه‌های کلیدی

انتخاب ویژگی

یادگیری زیرفضا

تجزیه‌ی ماتریسی

داده‌های میکرو-آرایه‌ای DNA

چکیده

داده‌های میکرو-آرایه‌ای DNA در یادگیری ماشین و تشخیص انواع مختلف ساختارهای سرطانی نقش مهمی را ایفا می‌کنند. داده‌های میکرو-آرایه‌ای به طور معمول شامل تعداد زیادی ویژگی و تعداد اندکی نمونه هستند. هم‌چنین، این‌گونه داده‌ها به دلیل داشتن برخی ویژگی‌های نامرتبط می‌توانند موجب بیش‌برازش و کاهش دقت پیش‌بینی طبقه‌بند کننده‌ها شوند. بنابراین، آنالیز داده‌های میکرو-آرایه‌ای امری مهم و چالش برانگیز در یادگیری ماشین و فناوری ژنتیک مولکولی محسوب می‌شود. یک راه مستقیم برای مقابله با این چالش، کاهش بعد داده می‌باشد. روش انتخاب ویژگی به عنوان یک راه‌کار مهم برای کاهش ابعاد و افزایش کارایی الگوریتم‌های یادگیری عمل می‌کند. در این مقاله، با استفاده از مفهوم پایه برای مجموعه‌ی داده‌های میکرو-آرایه‌ای، یک روش جدید انتخاب ویژگی معرفی شده است. به عبارت دیگر، از یک پایه شامل یک زیرمجموعه‌ی بسیار کوچک از ژن‌ها، به جای کل مجموعه‌ی داده‌های میکرو-آرایه‌ای در تعریف مساله‌ی انتخاب ویژگی استفاده شده است. در این روش مساله‌ی انتخاب ویژگی بر اساس دیدگاه یادگیری زیرفضا و تجزیه‌ی ماتریس پایه فرمول‌بندی شده است. در نهایت، با استفاده از مجموعه‌ی داده‌های میکرو-آرایه‌ای DNA، کارایی روش پیشنهادی بررسی شده و نتایج به دست آمده با نتایج چند روش انتخاب ویژگی معتبر مقایسه شده است.

[*]نویسنده‌ی مسئول

نشانی گروه ریاضی کاربردی، دانشکده‌ی علوم و فناوری‌های نوین، دانشگاه تحصیلات تکمیلی صنعتی و فناوری پیشرفته، کرمان ، ایران

کد پستی ۱۱۷- ۷۶۳۱۵ **تلفن** ۳۳۷۷۶۶۱۱-۳۴-۹۸+

پست الکترونیک *f.saberimovahed@kgut.ac.ir* **دورنگار** ۳۳۷۷۶۶۱۱-۳۴-۹۸+

۱- مقدمه

با توجه به گسترش استفاده از داده‌های با ابعاد بالا در مسائل واقعی مرتبط با مهندسی پزشکی، اهمیت روش‌های یادگیری ماشین به منظور اکتشاف الگوهای موجود در داده‌ها، بیش از گذشته احساس می‌شود. از مجموعه‌ی داده‌های با ابعاد بالا، می‌توان به مجموعه‌ی داده‌های میکرو-آرایه‌ای[۱] DNA اشاره کرد، که یک شاخه‌ی تحقیقاتی معروف در یادگیری ماشین و فناوری ژنتیک مولکولی محسوب می‌شود [۱].

داده‌های میکرو-آرایه‌ای DNA معمولا شامل حجم عظیمی از داده‌های ژن مزاحم و غیرمرتبط بوده که حاوی اطلاعاتی نادرست و بی‌ربط نیز هستند. در تکنولوژی میکرو-آرایه‌ای، انتخاب ژن‌های موثر برای طبقه‌بندی نمونه‌ها، از اهمیت زیادی در مطالعات بیان ژن برخوردار است. شایان ذکر است که منظور از طبقه‌بندی داده‌های میکرو-آرایه‌ای باینری (دوکلاسه)، فرایندی برای جداسازی بیماران سالم از بیماران سرطانی بر اساس پروفایل بیان ژن آن‌ها می‌باشد. هم‌چنین داده‌های میکرو-آرایه‌ای غیرباینری (چندکلاسه) نیز وجود داشته که از آن‌ها برای تشخیص بین گونه‌های مختلف تومور استفاده شده و فرایند طبقه‌بندی این نوع از داده‌ها بسیار پیچیده‌تر از داده‌های باینری است [۱، ۲].

به همین دلیل، بهتر است که پیش از طبقه‌بندی داده‌های میکرو-آرایه‌ای، ژن‌های مزاحم حذف شوند. بنابراین، در بهبود کارایی طبقه‌بندی مربوط به داده‌های میکرو-آرایه‌ای، یک انتخاب ویژگی مناسب، به معنای یک انتخاب کوچک و موثر از میان حجم انبوهی از داده‌های بیان ژن با ابعاد بالا، نقشی حیاتی را ایفا می‌کند. در ادامه به سه مزیت اصلی انتخاب ژن‌های مناسب اشاره شده است.

۱- کاهش هزینه‌ی تشخیص بیماری، زیرا تمرکز روی یک زیرمجموعه‌ی کوچک از ژن‌ها نسبت به هزاران ژن، دارای هزینه‌ی کم‌تری در هنگام تشخیص بیماری است.

۲- کاهش هزینه‌ی محاسباتی به دلیل کاهش ابعاد

۳- آسان‌تر بودن تشخیص ژن‌های موثر در بیماری

داده‌های میکرو-آرایه‌ای معمولا به منظور انجام مراحل آموزش و امتحان، دارای تعداد کمی نمونه (معمولا کم‌تر از ۱۰۰ بیمار) بوده و به دلیل ساختار خود در اندازه‌گیری حالت ژن، دارای تعداد زیادی ویژگی (معمولا ده‌ها هزار ژن) می‌باشند. علاوه بر این همان‌طور که در ابتدای این بخش توضیح داده شد، در داده‌های میکرو-آرایه‌ای، تعداد زیادی از ویژگی‌ها حاوی اطلاعات مهمی نمی‌باشند. این‌گونه از ویژگی‌ها در ادبیات یادگیری ماشین با عنوان ویژگی‌های غیرمرتبط[۲] نامیده می‌شوند. در این حالت، در صورت بررسی تمام ویژگی‌ها، ممکن است روش طبقه‌بندی دچار بیش‌برازش[۳] شده و نتواند الگوهای جدید را به خوبی پیش‌بینی کند. به بیان دقیق‌تر، بیش‌برازش به این موضوع اشاره دارد که روش طبقه‌بندی، بسیار خوب آموزش دیده اما قدرت تعمیم‌پذیری را ندارد و فقط قادر است داده‌هایی که در مجموعه‌ی آموزشی وجود دارند را به درستی پیش‌بینی کند [۱]. بنابراین، به دلیل وجود تعداد کم نمونه‌ها در مقابل تعداد زیاد ویژگی‌ها و وجود ویژگی‌های غیرمرتبط، آنالیز داده‌های میکرو-آرایه‌ای موضوعی چالش برانگیز در یادگیری ماشین محسوب می‌شود.

روش انتخاب ویژگی یک شیوه‌ی رایج برای کاهش ابعاد داده بوده و به گونه‌ای طراحی می‌شود تا با پیدا کردن یک زیرمجموعه از ویژگی‌های موثر، ابعاد را کاهش داده و باعث بهبود کارایی الگوریتم‌های یادگیری شود. تا کنون مشهورترین روش‌های انتخاب ویژگی برای داده‌های میکرو-آرایه‌ای، روش‌های مبتنی بر فیلتر می‌باشند. اساس این روش‌ها، ارزیابی اهمیت یک مجموعه از ژن‌ها با توجه به مشخصات ذاتی داده و برخی از معیارهای آماری است. از معروف‌ترین روش‌های انتخاب ویژگی مبتنی بر فیلتر می‌توان به روش همبستگی سریع[۴] [۳]، روش اینتراکت[۵] [۴]، روش بهره‌وری اطلاعات[۶] [۵]، روش ReliefF [۶]، روش بیشینه‌ی ارتباط و کمینه‌ی افزونگی[۷] [۷] و روش SVM مبتنی بر حذف ویژگی‌های زائد[۸] [۸] اشاره کرد. اگر چه روش‌های مبتنی بر فیلتر از پیچیدگی محاسباتی پایینی برخوردار بوده و سرعت بالایی دارند، اما معمولا از دقت بالایی برخوردار نیستند.

یکی از تکنیک‌هایی که به تازگی در روش انتخاب ویژگی مورد توجه قرار گرفته است، روش تجزیه‌ی ماتریسی می‌باشد. هم‌چنین برخی از روش‌های انتخاب ویژگی، دیدگاه تجزیه‌ی ماتریسی را در کنار ایده‌ی یادگیری زیرفضا مورد بررسی قرار می‌دهند که از مزایای آن می‌توان به تبدیل یک فضای با ابعاد

---

[۱] Micro-Array
[۲] Irrelevant
[۳] Over Fitting
[۴] Fast Correlation-based Filter Method (FCBF)
[۵] Intract Method (INT)
[۶] Information Gain Method (IG)
[۷] Maximum Relevancy Minimum Redundancy Method (MRMR)
[۸] SVM Recursive Feature Elimination Method (SVM-REF)

بالا به یک زیرفضای با ابعاد پایین اشاره کرد. برای مثال، وانگ و هم‌کارانش [۹] یک روش جدید انتخاب ویژگی بدون نظارت ارائه کردند که این روش بر مبنای دیدگاه یادگیری زیرفضا توسعه یافته است و به عنوان یک مساله‌ی تجزیه‌ی ماتریسی بیان می‌شود. وانگ و هم‌کارانش [۱۰] یک روش انتخاب ویژگی بدون نظارت بر اساس بیشینه‌ی تصویر و کمینه‌ی افزونگی ارائه کردند. در این روش با استفاده از تکنیک تجزیه‌ی ماتریسی، ویژگی‌هایی انتخاب می‌شوند که از دیدگاه جبری کم‌ترین وابستگی را نسبت به یک‌دیگر داشته باشند.

ایده‌ی یادگیری زیرفضا در زمینه‌ی کاهش بعد در برخی از روش‌های استخراج ویژگی خطی مانند روش تحلیل مولفه‌های اصلی[۱] و تجزیه‌ی مقدار تکین[۲] نیز قابل مشاهده است [۱۱]. هنگامی که حجم داده بزرگ باشد، هر دو روش PCA و SVD می‌توانند عمل‌کرد بسیار خوبی از خود نشان دهند. با این حال هنگامی که مجموعه‌ی داده به شدت غیرخطی باشد، این دو روش معمولا کارایی مناسبی نخواهند داشت. به علاوه در روش PCA راستای نگاشت بر اساس واریانس انتخاب شده و بنابراین ممکن است که برخی از اطلاعات از بین بروند.

به عنوان مثالی دیگر از روش‌های کاهش بعد با استفاده از ایده‌ی یادگیری زیرفضا می‌توان به دو روش استخراج ویژگی غیرخطی تعبیه‌ی خطی محلی[۳] [۱۲] و نگاشت طول‌پا[۴] [۱۳] اشاره کرد که فرایند کاهش بعد را بر اساس ساختار منیفلد داده انجام می‌دهند. هنگامی که مجموعه‌ی داده شدیدا غیرخطی باشد، دو روش LEE و ISOMAP عمل‌کرد بسیار خوبی از خود نشان می‌دهند. از معایب این دو روش می‌توان به موارد زیر اشاره کرد.

۱- هنگامی که حجم داده بسیار بزرگ باشد، این دو روش معمولا کارایی مناسبی نخواهند داشت

۲- این دو روش، به خصوص روش LEE، در برخی از مسائل مصورسازی اطلاعات در مقایسه با مسائل کاهش بعد، عمل‌کرد بهتری از خود نشان می‌دهند [۱۴].

در نظریه‌ی فضای برداری، هرگاه اعضای یک مجموعه از بردارها مانند B مستقل خطی بوده و بتوان هر عضو از فضای برداری را بر اساس ترکیب خطی از اعضای مجموعه‌ی B بیان کرد، این مجموعه یک پایه نامیده می‌شود.

ابراهیم‌پور و هم‌کارانش [۲] با استفاده از مفهوم ریاضی مجموعه‌های مستقل خطی، یک روش انتخاب ویژگی را برای داده‌های میکرو-آرایه‌ای معرفی کردند که در آن یک زیرمجموعه‌ی مستقل خطی از ویژگی‌ها که میزان افزونگی[۵] آن کمینه باشد به عنوان مجموعه‌ای از ویژگی‌های متمایز و ویژه انتخاب می‌شود. ایده‌ی استقلال خطی در مقاله‌ی [۲]، این انگیزه را برای تحقیق جاری ایجاد کرده که مفهوم پایه در بحث انتخاب ویژگی برای داده‌های میکرو-آرایه‌ای مورد بررسی قرار داده شود. در واقع دلیل توجه به بحث پایه، کامل‌تر بودن این مفهوم در مقایسه با مفهوم استقلال خطی است. به بیان دقیق‌تر، باید توجه داشت که یک پایه به عنوان یک زیرمجموعه‌ی مستقل خطی از فضای اصلی، این ویژگی را دارد که کل فضا می‌تواند به طور منحصر به فرد با توجه به مجموعه‌ی پایه توصیف شود، در حالی‌که یک مجموعه‌ی مستقل خطی به تنهایی خاصیت توصیف فضا را ندارد. در نتیجه، با ارائه‌ی یک پایه برای مجموعه‌ی ژن‌های موجود در مجموعه‌ی داده‌های میکرو-آرایه‌ای، این بستر فراهم می‌شود که مجموعه‌ی داده‌های میکرو-آرایه‌ای با حجم عظیمی از داده‌های ژن توسط یک مجموعه‌ی بسیار کوچک از ژن‌ها توصیف شود.

در این مقاله با استفاده از ایده‌ی یادگیری زیرفضا در مقاله‌ی [۲]، فاصله‌ی بین دو زیرفضا تعریف شده و در مورد برخی از ویژگی‌های اساسی آن با جزئیات بحث شده است. سپس با استفاده از مفهوم پایه برای مجموعه‌ی داده‌های میکرو-آرایه‌ای، مساله‌ی انتخاب ویژگی به عنوان یک مساله‌ی تجزیه‌ی ماتریس بیان شده است. به علاوه برای منظم‌سازی پراکندگی و انتخاب ویژگی‌های با شباهت کم‌تر، از کمینه کردن نرم فروبنیوس ماتریس انتخاب ویژگی استفاده شده و در نهایت یک الگوریتم به‌روزرسانی برای حل مساله‌ی پیشنهادی معرفی شده است.

نوآوری اصلی و بحث کلیدی این مقاله، بررسی مفهوم پایه برای مجموعه‌ی داده‌های میکرو-آرایه‌ای در مساله‌ی انتخاب ویژگی است. به عبارت دیگر، به جای استفاده از کل مجموعه‌ی داده‌های میکرو-آرایه‌ای در تعریف مساله‌ی انتخاب ویژگی، از یک پایه که زیرمجموعه‌ای از کل ژن‌ها است استفاده می‌شود. در نتیجه، روش انتخاب ویژگی پیشنهادی تنها با بخش کوچکی از مجموعه‌ی داده در ارتباط بوده و بنابراین هنگامی که حجم داده بسیار بزرگ باشد، می‌تواند کارایی خوبی از خود نشان دهد. از آن‌جا که مجموعه‌ی پایه توانایی توصیف کل مجموعه‌ی داده را دارد، می‌تواند حاوی اطلاعات مهمی نیز باشد.

در ادامه، در بخش ۲ برخی از نمادهای مورد استفاده در مقاله معرفی شده است. در بخش ۳ به بررسی فاصله‌ی میان زیرفضاها

[۱] Principal Component Analysis (PCA)
[۲] Sigular-Value Decomposition (SVD)
[۳] Locally Linear Embedding (LEE)
[۴] Isometric Map (ISOMAP)
[۵] Redundancy

پرداخته شده و در بخش ۴ مساله‌ی انتخاب ویژگی به صورت تجزیه‌ی ماتریس پایه برای زیرفضای برداری تولید شده توسط داده‌ها فرمول‌بندی شده است. در بخش ۵ روش پیشنهادی و یک الگوریتم به‌روزرسانی تکراری به منظور حل آن پیشنهاد شده است. در بخش‌های ۶ تا ۱۱ نتایج آزمایش‌های عددی حاصل از مقایسه‌ی الگوریتم روش پیشنهادی با برخی از رایج‌ترین روش‌های انتخاب ویژگی با نظارت روی تعدادی از مجموعه‌ی داده‌های میکرو-آرایه‌ای ارائه شده، در بخش ۱۲ به بررسی هزینه‌ی محاسباتی روش پیشنهادی پرداخته شده و بحث و نتیجه‌گیری این تحقیق نیز در بخش ۱۳ بیان شده است.

۲- مقدمات و نمادها

در این مقاله اسکالرها با حروف کوچک و کج، بردارها با حروف کوچک و پررنگ و ماتریس‌ها و مجموعه‌ها با حروف بزرگ و پررنگ نشان داده شده است. ماتریس داده‌ی X یک ماتریس n×d بوده که n تعداد نمونه‌ها و d تعداد ویژگی‌ها است. ماتریس B هرگاه مجموعه‌ی ستون‌های آن یک مجموعه‌ی مستقل خطی بوده و هر عضو از X بر اساس اعضای آن بیان شود، یک ماتریس پایه برای X نامیده می‌شود. نرم فروبنیوس ماتریس X و ۲-نرم بردار x به ترتیب، با $\|\mathbf{X}\|_F$ و $\|\mathbf{x}\|_2$ نمایش داده شده و A^T و A^{-1} به ترتیب ترانهاده و معکوس ماتریس A هستند.

۳- فاصله‌ی بین دو زیرفضا

یادگیری زیرفضا یک روش مهم و پرکاربرد در ارتباط با بحث کاهش بعد است. در این روش سعی می‌شود تا یک زیرفضای با ابعاد پایین به گونه‌ای انتخاب شود که توصیف مناسبی از فضای اصلی با بعد بالا ارائه دهد. ارزیابی فاصله‌ی بین دو زیرفضا به ویژه دو زیرفضای دارای اشتراک، یک ابزار مهم در یادگیری زیرفضا محسوب می‌شود. در ادامه یک معیار مهم در زمینه‌ی ارزیابی فاصله‌ی بین دو زیرفضا بررسی شده است [۹].

۳-۱- تعریف ۱

فرض کنید S یک مجموعه‌ی m-تایی از بردارها در $\mathbb{R}^n$ باشد. در این صورت زیرفضای تولید شده توسط S که با span(S) نمایش داده می‌شود، به صورت مجموعه‌ی تمام ترکیبات خطی از بردارهای مجموعه‌ی S است. به علاوه فاصله‌ی بردار x از span(S) به صورت زیر محاسبه می‌شود.

$$\vec{d}(\mathbf{x},\mathrm{span}(\mathbf{S})) = \min_{\mathbf{s}\in\mathrm{span}(\mathbf{S})}\|\mathbf{x}-\mathbf{s}\|_2 = \min_{\boldsymbol{\alpha}}\|\mathbf{x}-\mathbf{S}\boldsymbol{\alpha}\|_2 \quad (۱)$$

در این رابطه S ماتریسی است که ستون‌های آن از بردارهای مجموعه‌ی S تشکیل شده است. با الهام گرفتن از تعریف (۱) می‌توان به بررسی فاصله‌ی بین دو زیرفضای دلخواه پرداخت.

۳-۲- تعریف ۲

فرض کنید S_1 و S_2 به ترتیب دو زیرمجموعه‌ی p و q-تایی از بردارها در $\mathbb{R}^n$ باشند. فاصله‌ی میان span(S_1) و span(S_2) به صورت زیر تعریف می‌شود.

$$\vec{d}(\mathrm{span}(\mathbf{S_1}),\mathrm{span}(\mathbf{S_2})) = \min_{\mathbf{H}\in\mathbb{R}^{q\times p}}\|\mathbf{S_1}-\mathbf{S_2}\mathbf{H}\|_F$$

در قضیه‌ی بعدی می‌توان نشان داد که اگر B یک ماتریس پایه برای مجموعه‌ی داده‌ی X باشد، آن‌گاه فاصله‌ی بین span(X) و span(B) برابر با صفر است.

۳-۳- قضیه‌ی ۱

فرض کنید B یک ماتریس پایه برای مجموعه‌ی داده‌ی X باشد. در این صورت رابطه‌ی زیر برقرار است.

$$\vec{d}(\mathrm{span}(\mathbf{X}),\mathrm{span}(\mathbf{B})) = 0$$

از آن‌جا که B یک ماتریس پایه برای مجموعه‌ی داده‌ی X است، ماتریسی مانند W وجود دارد که در رابطه‌ی X=BW صدق کرده و در نتیجه رابطه‌ی زیر برای دو زیرفضای span(X) و span(B) برقرار می‌باشد.

$$\vec{d}(\mathrm{span}(\mathbf{X}),\mathrm{span}(\mathbf{B})) = \min_{\mathbf{H}\in\mathbb{R}^{n\times d}}\|\mathbf{X}-\mathbf{B}\mathbf{H}\|_F = \min_{\mathbf{H}\in\mathbb{R}^{n\times d}}\|\mathbf{B}\mathbf{W}-\mathbf{B}\mathbf{H}\|_F$$

اکنون با فرض H=W، حکم قضیه به وضوح برقرار خواهد بود.

۴- معرفی مساله‌ی انتخاب ویژگی بر اساس استفاده از ماتریس پایه

در برخی از پژوهش‌هایی که اخیرا در ارتباط با بحث انتخاب ویژگی انجام شده است، رابطه‌ی (۱) نقشی اساسی ایفا می‌کند. برای مثال، در مقاله‌ی [۹]، مساله‌ی انتخاب ویژگی به صورت زیر بیان شده است.

$$\underset{\mathrm{I}}{\mathrm{argmin}}\ \vec{d}(\mathrm{span}(\mathrm{X}),\mathrm{span}(\mathrm{X_I})) \quad s.t. \quad |\mathbf{I}| = k \quad (۲)$$

در این رابطه، k تعداد ویژگی‌های انتخاب شده (k≤d)، I زیرمجموعه‌ای از مجموعه‌ی {1,…,d} (|I|=k) و X_I زیرمجموعه‌ای از ویژگی‌های انتخاب شده از X است.

از طرف دیگر، از دیدگاه فاصله‌ی بین دو زیرفضا، قضیه‌ی (۱) بیان‌گر این حقیقت است که زیرفضای تولید شده توسط B بر زیرفضای تولید شده توسط X منطبق خواهد شد. این نتیجه از آن جهت اهمیت دارد که در مساله‌ی انتخاب ویژگی (۲) می‌توان از ماتریس پایه‌ی B به جای ماتریس داده‌ی X استفاده کرد و مساله‌ی انتخاب ویژگی را به این صورت تعبیر کرد که از زیرفضای تولید شده توسط ماتریس پایه با تعداد ویژگی‌های کم‌تر به جای زیرفضای تولید شده توسط ماتریس داده‌ی X استفاده گردد. به عبارت دقیق‌تر، مساله‌ی انتخاب ویژگی به صورت زیر در نظر گرفته می‌شود.

$$\underset{\mathbf{I}}{\mathbf{argmin}}\ \vec{d}(\mathbf{span(B)}, \mathbf{span(B_I)}) \quad s.t. \quad |\mathbf{I}| = k \qquad (۳)$$

بر اساس آن‌چه در تعریف (۲) و قضیه‌ی (۱) بیان شد، می‌توان نتیجه گرفت که مساله‌ی (۳) با مساله‌ی زیر معادل است.

$$\underset{\mathbf{W}\epsilon\mathbb{R}^{n\times k},\mathbf{H}\epsilon\mathbb{R}^{k\times n}}{\operatorname{argmin}} \|\mathbf{B}-\mathbf{BWH}\|_{\mathrm{F}}^2$$

که W ماتریس ضرایب ویژگی‌های انتخاب شده و H ماتریس ضرایب فضای ویژگی پایه در فضای ویژگی‌های منتخب است.

۵- الگوریتم پیشنهادی

در این بخش یک روش جدید برای انتخاب ویژگی بر اساس تجزیه‌ی ماتریس پایه ارائه شده است. این روش از دیدگاه یادگیری زیرفضا، که منجر به مساله‌ی (۲) گردید، استفاده می‌کند. هم‌چنین به منظور منظم‌سازی پراکندگی و انتخاب ویژگی‌های با شباهت کم‌تر، از کمینه کردن نرم فروبنیوس ماتریس W استفاده شده است. بنابرین مساله‌ی انتخاب ویژگی به صورت رابطه‌ی زیر می‌باشد.

$$\underset{\mathbf{W},\mathbf{H}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{B}-\mathbf{BWH}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\|\mathbf{W}\|_{\mathrm{F}}^2 \qquad (۴)$$

که $\beta>0$ پارامتر منظم‌سازی متناظر با تنک‌سازی ماتریس ضرایب W است. در ادامه، تابع هدف زیر در نظر گرفته می‌شود.

$$\mathrm{f}(\mathbf{W},\mathbf{H}) = \frac{1}{2}\|\mathbf{B}-\mathbf{BWH}\|_{\mathrm{F}}^2 + \frac{\beta}{2}\|\mathbf{W}\|_{\mathrm{F}}^2$$

برای حل مساله‌ی بهینه‌سازی ۴ از دو گام زیر استفاده می‌شود.
۱- ماتریس H ثابت فرض شده، سپس مشتق جزئی تابع f بر حسب ماتریس W محاسبه و برابر با صفر قرار داده شده و در نهایت ماتریس W محاسبه می‌شود.
۲- ماتریس W ثابت فرض شده، سپس مشتق جزئی تابع f بر حسب ماتریس H محاسبه و برابر با صفر قرار داده شده و در نهایت ماتریس H محاسبه می‌شود.

مشتق تابع f نسبت به W و H به صورت زیر محاسبه می‌شود.

$$\frac{\partial \mathrm{f}}{\partial \mathbf{W}} = -\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{W}\mathbf{H}\mathbf{H}^{\mathrm{T}} + \beta\mathbf{W}$$
$$\frac{\partial \mathrm{f}}{\partial \mathbf{H}} = -\mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B} + \mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{W}\mathbf{H}$$

با صفر قرار دادن $\partial \mathrm{f}/\partial \mathrm{W}$ و $\partial \mathrm{f}/\partial \mathrm{H}$ رابطه‌ی زیر به دست می‌آید.

$$\mathbf{W} = \frac{\mathbf{1}}{\boldsymbol{\beta}}(\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{H}^{\mathrm{T}} - \mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{W}\mathbf{H}\mathbf{H}^{\mathrm{T}})$$
$$(\mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{W})\mathbf{H} = \mathbf{W}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{B}$$

با توجه به روابط به دست آمده، الگوریتم انتخاب ویژگی بر مبنای یادگیری زیرفضا، تجزیه‌ی ماتریس پایه و منظم‌سازی پراکندگی به صورت الگوریتم زیر ارائه می‌شود.

الگوریتم (۱)- الگوریتم انتخاب ویژگی بر مبنای یادگیری زیرفضا، تجزیه‌ی ماتریس پایه و منظم‌سازی پراکندگی

ورودی: ماتریس داده‌ی $X\epsilon\mathbb{R}^{n\times d}$، پارامتر منظم‌سازی β و پارامتر T
خروجی: مجموعه‌ی ویژگی‌های انتخاب شده‌ی I که |I|=k
۱- ماتریس پایه‌ی $\mathrm{B}\ \epsilon\mathbb{R}^{n\times n}$ را انتخاب کنید
۲- ماتریس‌های $\mathbf{H_{(0)}}\epsilon\mathbb{R}^{k\times n}$ و $\mathbf{W_{(0)}}\epsilon\mathbb{R}^{n\times k}$ را به صورت تصادفی انتخاب کنید
۳- برای t=0:T مراحل ۴ و ۵ را تکرار کنید
۴- ماتریس $\mathrm{H}_{(t)}$ را ثابت در نظر گرفته و قرار دهید:

$$\mathbf{W_{(t+1)}} = \frac{1}{\beta}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{H}_{(t)}^{T} - \mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{W}_{(t)}\mathbf{H}_{(t)}\mathbf{H}_{(t)}^{T}\right)$$

۵- ماتریس $\mathrm{W}_{(t)}$ را ثابت در نظر گرفته و قرار دهید:

$$\mathbf{H_{(t+1)}} = (\mathbf{W}_{(t)}^{T}\mathbf{B}^{\mathrm{T}}\mathbf{B}\mathbf{W}_{(t)})^{-1}\mathbf{W}_{(t)}^{T}\mathbf{B}^{\mathrm{T}}\mathbf{B}$$

۶- فرض کنید: W=[W₁;W₂;…;Wₙ]، مقادیر $\|\mathbf{w}_i\|_2$ $(i\epsilon\{1,\dots,n\})$ را به صورت نزولی مرتب کرده و مجموعه‌ی I را به عنوان k عضو اول آن انتخاب کنید

۵-۱- ساخت ماتریس پایه‌ی B

با توجه به گام اول الگوریتم (۱)، لازم است تا ماتریس پایه‌ی B ساخته شود. در این بخش نحوه‌ی ساخت یک پایه مانند B برای مجموعه‌ی داده‌ی میکرو-آرایه‌ای توضیح داده شده است.
ابراهیم‌پور و هم‌کارانش [۲] یک روش انتخاب ویژگی را برای داده‌های میکرو-آرایه‌ای با استفاده از روش بهره‌وری اطلاعات معرفی کردند. در روش پیشنهادی آن‌ها، امتیاز IG مربوط به هر یک از ویژگی‌ها محاسبه شده، و به هر ویژگی یک رتبه بر اساس امتیاز IG آن اختصاص داده می‌شود. سپس مجموعه‌ی امتیازهای IG به صورت نزولی مرتب شده و مجموعه‌ی ویژگی‌ها

بر اساس مجموعه‌ی امتیازهای IG به دست آمده مرتب می‌شود. بنابراین در مجموعه‌ی مرتب شده‌ی جدید، اولین ویژگی بالاترین امتیاز IG را داشته، دومین ویژگی در رتبه‌ی بعدی بالاترین امتیاز IG قرار داشته و همین روند تا آخرین ویژگی ادامه می‌یابد. جزئیات این فرایند به شرح زیر است.

۱- فرض کنید ماتریس $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_d}] \in \mathbb{R}^{n \times d}$ نشان دهنده‌ی یک مجموعه‌ی داده‌ی میکرو-آرایه‌ای باشد که شامل n نمونه و d ویژگی است.

۲- برای i=1,…,d انجام دهید:

۳- امتیاز IG ویژگی X_i-ام ← IG(i)

۴- مجموعه‌ی امتیازهای IG به صورت نزولی مرتب شود.

۵- بر اساس مجموعه‌ی مرتب شده‌ی IG، ویژگی‌های مجموعه‌ی داده‌ی X مرتب شوند. به عبارت دیگر:

۶- مجموعه‌ی داده‌ی مرتب شده‌ی X ← X_{new}

با استفاده از توضیحات فوق می‌توان یک پایه مانند B برای مجموعه‌ی داده‌ی میکرو-آرایه‌ای X ساخت. برای این منظور، فرض کنید مجموعه‌ی داده‌ی میکرو-آرایه‌ای X دارای n ستون مستقل خطی باشد، سپس:

۱- فرض کنید با توجه به مجموعه‌ی امتیازهای مرتب شده‌ی IG، مجموعه‌ی داده‌ی مرتب شده‌ی جدید X_{new} ساخته شده باشد و قرار دهید:

$$\mathbf{X_{new}} = [\mathbf{X_{new}}^{(1)}, \mathbf{X_{new}}^{(2)}, \dots, \mathbf{X_{new}}^{(d)}].$$

۲- اولین عضو X_{new} را به عنوان اولین عضو B انتخاب نمایید ($B=\{X_{new}^{(1)}\}$). اولین عضو B دارای بالاترین امتیاز IG است.

۳- $X_{new}^{(2)}$ را در نظر بگیرید. اگر مجموعه‌ی $\{X_{new}^{(1)}, X_{new}^{(2)}\}$ یک مجموعه‌ی مستقل خطی باشد، آن‌گاه قرار دهید:

$$\mathbf{B} = \{\mathbf{X_{new}}^{(1)}, \mathbf{X_{new}}^{(2)}\}$$

در غیر این صورت این روند را برای $X_{new}^{(3)}$ امتحان کنید. با ادامه‌ی این روند می‌توان یک مجموعه شامل n ویژگی مستقل خطی که یک پایه برای مجموعه‌ی داده‌ی میکرو-آرایه‌ای X است را تولید کرد.

یک مزیت قابل توجه پایه‌ی ساخته شده، بهره بردن از امتیاز IG ویژگی‌ها است. در واقع با توجه به رتبه‌بندی شدن ویژگی‌ها و کارایی روش IG در به دست آوردن اطلاعات مهم از ویژگی‌ها، اعضای مجموعه‌ی B حاوی اطلاعات خوبی از ویژگی‌ها هستند.

۶- توصیف مجموعه‌ی داده‌ها

تا کنون دو نوع از داده‌های میکرو-آرایه‌ای توسط محققان مورد بررسی قرار گرفته است. گونه‌ی اول که در فناوری ژنتیک مولکولی بسیار رایج است، مجموعه‌ی داده‌های میکرو-آرایه‌ای باینری (دوکلاسه) بوده و معمولا در ارتباط با جداسازی افراد سالم از بیماران سرطانی می‌باشد. گونه‌ی دوم که برای تشخیص انواع مختلف تومور استفاده می‌شود، مجموعه‌ی داده‌های میکرو-آرایه‌ای غیرباینری (چندکلاسه) بوده و فرایند طبقه‌بندی آن‌ها بسیار پیچیده‌تر از داده‌های باینری است [۱].

در این مقاله هفت مجموعه‌ی داده‌ی میکرو-آرایه‌ای باینری به کار گرفته شده که در مطالعات مربوط به بیماران سرطانی به صورت گسترده مورد استفاده قرار می‌گیرند. برای مثال، مجموعه‌ی داده‌ی سرطان کولون[۱] شامل ۶۲ نمونه و ۲۰۰۰ ژن بوده و از ۴۰ نمونه‌ی تومور و ۲۲ نمونه‌ی بافت سالم تشکیل شده است. جزئیات این داده‌ها در جدول (۱) ارائه شده است.

جدول (۱)- اطلاعاتی از هفت مجموعه‌ی داده‌ی میکرو-آرایه‌ای به کار گرفته شده در آزمایش‌ها

داده	تعداد ویژگی‌ها	تعداد نمونه‌ها	مرجع
Brain	۱۲۶۲۵	۲۱	[۱]
CNS	۷۱۲۹	۶۰	[۱]
Colon	۲۰۰۰	۶۲	[۱]
DLBCL	۴۰۲۶	۴۷	[۱]
GLI-85	۲۲۲۸۳	۸۵	[۱]
Ovarian	۱۵۱۵۴	۲۵۳	[۱]
SMK	۱۹۹۹۳	۱۸۷	[۱]

۷- روش‌های مقایسه

در زمینه‌ی مجموعه‌ی داده‌های میکرو-آرایه‌ای، کارایی یک روش انتخاب ویژگی بر اساس پیش‌بینی کلاس‌ها ارزیابی می‌شود. بنابراین لازم است تا از برخی روش‌های یادگیری بانظارت به منظور پیش‌بینی استفاده شود. در ادامه برخی از روش‌های بانظارت مشهور از جمله روش انتخاب ویژگی مبتنی بر فرم کاهش یافته‌ی سطری پلکانی[۲] [۲]، روش همبستگی سریع [۳]، روش اینتراکت [۴]، روش بهره‌وری اطلاعات [۵]، روش ReliefF [۶]، روش بیشینه‌ی ارتباط و کمینه‌ی افزونگی [۷] و روش SVM مبتنی بر حذف ویژگی‌های زائد [۸] برای مقایسه با روش پیشنهادی این مقاله در نظر گرفته شده است.

[۱] Colon

[۲] RRFES

۸- اعتبارسنجی متقابل

در برخی موارد مانند مجموعه‌ی داده‌های میکرو-آرایه‌ای با ابعاد بالا، می‌توان مشاهده کرد که نه تنها داده‌ها به مجموعه‌های آموزشی و امتحان دسته‌بندی نمی‌شوند، بلکه قابلیت تعادل در همه‌ی کلاس‌ها را نیز ندارند. با توجه به این نکته، تکنیک‌های اعتبارسنجی مهمی برای بهبود عمل‌کرد الگوریتم‌ها روی این مجموعه‌ی داده‌ها ایفا می‌شود تا مشخص شود که مدل مورد نظر تا چه اندازه در عمل مفید خواهد بود. اعتبارسنجی متقابل k-دسته‌ای[۱] یکی از مهم‌ترین تکنیک‌های اعتبارسنجی در مجموعه‌ی داده‌های میکرو-آرایه‌ای است که به صورت تصادفی داده‌های اصلی را به k قسمت مساوی تقسیم می‌کند. از این k زیرمجموعه، در هر تکرار یک زیرمجموعه برای اعتبارسنجی و k-1 مورد دیگر برای آموزش به کار می‌روند. این روال k مرتبه تکرار شده و همه‌ی داده‌ها دقیقا یک بار برای اعتبارسنجی به کار گرفته می‌شوند. در نهایت، میانگین نتیجه‌ی k مرتبه اعتبارسنجی به عنوان یک تخمین نهایی برگزیده می‌شود.

تا کنون مطالعات گسترده‌ای به منظور بررسی اثرات تکنیک‌های اعتبارسنجی متقابل روی مجموعه‌ی داده‌های میکرو-آرایه‌ای انجام شده است. نتایج این مطالعات نشان می‌دهد که روش توزیع بهینه‌ی متعادل طبقه‌طبقه شده‌ی اعتبارسنجی متقابل[۲] قادر است تا توزیع هر کلاس در هر دسته را تضمین کند. در این مقاله روش اعتبارسنجی k-دسته‌ای از نوع DOB-SCV به ازای k=۵ در همه‌ی آزمایش‌ها در نظر گرفته شده است.

۹- معیارهایی برای ارزیابی

برای ارزیابی کیفیت الگوریتم‌های انتخاب ویژگی روی داده‌های میکرو-آرایه‌ای، معمولا از چهار اندازه‌ی مشهور Ac، G-mean، Sp و Se استفاده شده که به شرح زیر می‌باشند.

$$Se = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$$
$$Sp = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$$
$$G-mean = \sqrt{\mathrm{Se}\times\mathrm{Sp}}$$
$$\mathrm{Ac} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

در واقع Se و Sp به ترتیب نشان دهنده‌ی این هستند که کدام یک از داده‌های آزمایش مثبت و منفی به درستی دسته‌بندی شده و G-mean میانگین هندسی Se و Sp است. هم‌چنین TP، TN، FN و FP به ترتیب بیان‌گر تعداد مثبت درست، تعداد منفی درست، تعداد منفی غلط و تعداد مثبت غلط هستند.

۱۰- نتایج و تجزیه و تحلیل

در این بخش نتایج کمی و کیفی عمل‌کرد روش پیشنهادی با سایر الگوریتم‌های انتخاب ویژگی مقایسه شده است. هدف اصلی الگوریتم‌های انتخاب ویژگی، انتخاب یک زیرمجموعه از ویژگی‌ها است که بیان‌گر تمام ویژگی‌ها باشد. در نتیجه تعداد ویژگی‌های انتخاب شده فرایند مهمی در این روش‌ها است. تعداد ویژگی‌های انتخاب شده توسط روش‌های متفاوت در جدول (۲) گزارش شده است. لازم به ذکر است که در الگوریتم (۱)، مقدار پارامتر T برابر با ۳۰ فرض شده و برای دست‌یابی به مقدار بهینه‌ی پارامتر β از استراتژی جست‌وجوی شبکه[۳] در محدوده‌ی $\{10^{-2}, 10^{-1}, 1, 10^{2}, 10^{4}, 10^{8}\}$ استفاده شده که در این مقاله نتایج مربوط به $\beta=10^{8}$ گزارش شده است.

جدول (۲)- تعداد ویژگی‌های انتخاب شده

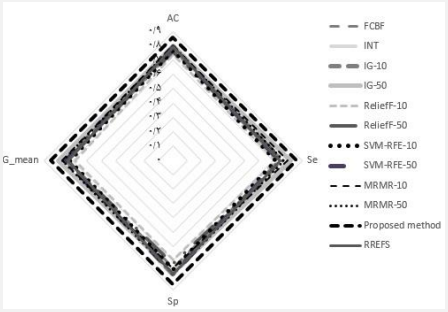
روش	Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk
FCBF	۱	۳۵	۱۵	۳۷	۱۱۸	۲۶	۵۵
INT	۴۹	۳۴	۱۶	۵۱	۱۲۳	۳۱	۵۱
IG-10	۱۰	۱۰	۱۰	۱۰	۱۰	۱۰	۱۰
IG-50	۵۰	۵۰	۵۰	۵۰	۵۰	۵۰	۵۰
ReliefF-10	۱۰	۱۰	۱۰	۱۰	۱۰	۱۰	۱۰
ReliefF-50	۵۰	۵۰	۵۰	۵۰	۵۰	۵۰	۵۰
RFE-10	۱۰	۱۰	۱۰	۱۰	۱۰	۱۰	۱۰
RFE-50	۵۰	۵۰	۵۰	۵۰	۵۰	۵۰	۵۰
MRMR-10	۱۰	۱۰	۱۰	۱۰	۱۰	۱۰	۱۰
MRMR-50	۵۰	۵۰	۵۰	۵۰	۵۰	۵۰	۵۰
RREFS	۵	۳۰	۴۹	۲۰	۲۰	۵۰	۲۰
Proposed method	۵	۳۰	۴۹	۲۰	۲۰	۵۰	۲۰

برای درک بهتر روش پیشنهادی، اندازه‌های Ac، G-mean، Sp و Se در جدول (۳) ارائه شده است. در این جدول نتایج تجربی به صورت میانگین عمل‌کرد سه طبقه‌بند C4.5، Naive-Bayes و SVM در هر مجموعه‌ی داده از طریق اجرای اعتبارسنجی متقابل ۵-دسته‌ای قابل مشاهده است. با توجه به میانگین اندازه‌های جدول (۳) می‌توان برتری روش پیشنهادی را بر سایر الگوریتم‌ها نتیجه گرفت. برای بررسی بهتر عمل‌کرد الگوریتم‌ها نمودار رادار میانگین اندازه‌های Ac، G-mean، Sp و Se که با اعمال سه طبقه‌بند C4.5، Naive-Bayes و SVM بر هر یک از الگوریتم‌ها به دست آمده، در شکل (۱) ارائه شده است.

[۱] K-Fold Cross Validation
[۲] DOB-SCV
[۳] Grid-Search

جدول (۳)- میانگین عمل‌کرد سه طبقه‌بند C4.5، Naive-Bayes و SVM

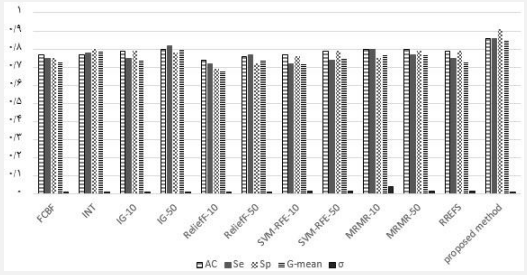
روش	معیارهای ارزیابی	Brain	CNS	Colon	DLBCL	Gli85	Ovarian	Smk	Avg
FCBF	Ac	۰/۷۰	۰/۵۶	۰/۸۰	۰/۸۷	۰/۸۵	۰/۹۹	۰/۶۴	۰/۷۷
	Se	۰/۵۰	۰/۶۵	۰/۷۱	۰/۹۰	۰/۸۸	۰/۹۸	۰/۶۴	۰/۷۵
	Sp	۰/۸۲	۰/۴۰	۰/۸۴	۰/۸۳	۰/۷۶	۰/۹۹	۰/۶۳	۰/۷۵
	G-mean	۰/۴۹	۰/۵۱	۰/۷۷	۰/۸۷	۰/۸۲	۰/۹۹	۰/۶۳	۰/۷۳
INT	Ac	۰/۸۸	۰/۵۸	۰/۸۱	۰/۸۶	۰/۸۱	۰/۹۹	۰/۴۶	۰/۷۷
	Se	۰/۷۷	۰/۶۲	۰/۷۳	۰/۸۷	۰/۸۴	۰/۹۸	۰/۶۹	۰/۷۸
	Sp	۰/۹۶	۰/۵۲	۰/۸۵	۰/۸۶	۰/۷۵	۰/۹۹	۰/۶۹	۰/۸۰
	G-mean	۰/۸۵	۰/۵۶	۰/۷۸	۰/۸۶	۰/۷۹	۰/۹۸	۰/۶۹	۰/۷۹
IG-10	Ac	۰/۷۵	۰/۶۲	۰/۷۸	۰/۸۷	۰/۸۷	۰/۹۶	۰/۶۶	۰/۷۹
	Se	۰/۵۳	۰/۶۹	۰/۶۳	۰/۸۹	۰/۸۹	۰/۹۴	۰/۶۵	۰/۷۵
	Sp	۰/۸۹	۰/۴۹	۰/۸۷	۰/۸۶	۰/۸۰	۰/۹۶	۰/۶۸	۰/۷۹
	G-mean	۰/۵۵	۰/۵۷	۰/۷۴	۰/۸۷	۰/۸۴	۰/۹۵	۰/۶۶	۰/۷۴
IG-50	Ac	۰/۷۸	۰/۶۲	۰/۸۳	۰/۹۱	۰/۸۵	۰/۹۹	۰/۶۵	۰/۸۰
	Se	۰/۸۰	۰/۶۸	۰/۷۹	۰/۹۳	۰/۸۷	۰/۹۸	۰/۶۵	۰/۸۲
	Sp	۰/۸۱	۰/۵۱	۰/۸۵	۰/۹۰	۰/۸۰	۰/۹۹	۰/۶۴	۰/۷۸
	G-mean	۰/۸۰	۰/۵۹	۰/۸۲	۰/۹۱	۰/۸۳	۰/۹۸	۰/۶۵	۰/۸۰
ReliefF-10	Ac	۰/۴۰	۰/۶۱	۰/۸۱	۰/۹۱	۰/۸۳	۰/۹۷	۰/۶۶	۰/۷۴
	Se	۰/۱۷	۰/۷۶	۰/۶۶	۰/۹۳	۰/۹۱	۰/۹۴	۰/۷۱	۰/۷۲
	Sp	۰/۴۸	۰/۳۲	۰/۸۹	۰/۹۰	۰/۶۴	۰/۹۸	۰/۶۰	۰/۶۹
	G-mean	۰/۲۴	۰/۴۷	۰/۷۶	۰/۹۱	۰/۷۶	۰/۹۵	۰/۶۵	۰/۶۸
ReliefF-50	Ac	۰/۴۶	۰/۶۲	۰/۸۱	۰/۹۰	۰/۸۶	۰/۹۸	۰/۶۹	۰/۷۶
	Se	۰/۴۰	۰/۷۱	۰/۷۲	۰/۹۳	۰/۸۹	۰/۹۷	۰/۷۸	۰/۷۷
	Sp	۰/۴۸	۰/۴۵	۰/۸۶	۰/۸۸	۰/۸۰	۰/۹۹	۰/۵۹	۰/۷۲
	G-mean	۰/۴۱	۰/۵۶	۰/۷۸	۰/۹۰	۰/۸۴	۰/۹۸	۰/۶۸	۰/۷۴
SVM-RFE-10	Ac	۰/۶۰	۰/۶۵	۰/۷۶	۰/۹۰	۰/۸۲	۰/۹۹	۰/۶۶	۰/۷۷
	Se	۰/۲۳	۰/۷۱	۰/۶۳	۰/۸۸	۰/۸۷	۰/۹۹	۰/۷۲	۰/۷۲
	Sp	۰/۷۶	۰/۵۳	۰/۸۳	۰/۹۲	۰/۷۰	۰/۹۹	۰/۵۹	۰/۷۶
	G-mean	۰/۴۰	۰/۶۱	۰/۷۲	۰/۹۰	۰/۷۷	۰/۹۹	۰/۶۵	۰/۷۲
SVM-RFE-50	Ac	۰/۶۹	۰/۶۵	۰/۷۸	۰/۸۷	۰/۸۳	۰/۹۹	۰/۶۸	۰/۷۹
	Se	۰/۳۷	۰/۷۱	۰/۶۸	۰/۸۸	۰/۸۵	۰/۹۸	۰/۷۲	۰/۷۴
	Sp	۰/۸۶	۰/۵۵	۰/۸۴	۰/۸۷	۰/۷۷	۰/۹۹	۰/۶۴	۰/۷۹
	G-mean	۰/۵۱	۰/۶۲	۰/۷۵	۰/۸۷	۰/۸۱	۰/۹۸	۰/۶۷	۰/۷۵
MRMR-10	Ac	۰/۶۷	۰/۶۴	۰/۸۱	۰/۹۱	۰/۸۵	۰/۹۹	۰/۷۰	۰/۸۰
	Se	۰/۶۰	۰/۷۷	۰/۷۱	۰/۹۳	۰/۸۹	۰/۹۹	۰/۷۴	۰/۸۰
	Sp	۰/۷۲	۰/۴۰	۰/۸۶	۰/۹۱	۰/۷۴	۰/۹۹	۰/۶۴	۰/۷۵
	G-mean	۰/۶۴	۰/۵۵	۰/۷۸	۰/۹۲	۰/۸۱	۰/۹۹	۰/۶۹	۰/۷۷
MRMR-50	Ac	۰/۷۳	۰/۶۶	۰/۸۲	۰/۸۹	۰/۸۳	۰/۹۹	۰/۶۸	۰/۸۰
	Se	۰/۴۷	۰/۷۰	۰/۷۶	۰/۹۰	۰/۸۶	۰/۹۸	۰/۷۱	۰/۷۷
	Sp	۰/۸۸	۰/۵۶	۰/۸۵	۰/۸۹	۰/۷۶	۰/۹۹	۰/۶۴	۰/۷۹
	G-mean	۰/۵۹	۰/۶۳	۰/۸۰	۰/۸۹	۰/۸۱	۰/۹۸	۰/۶۷	۰/۷۷
RREFS	Ac	۰/۷۴	۰/۶۵	۰/۸۲	۰/۸۵	۰/۸۷	۰/۹۷	۰/۶۶	۰/۷۹
	Se	۰/۷۰	۰/۵۰	۰/۷۵	۰/۸۳	۰/۸۵	۰/۹۸	۰/۶۵	۰/۷۵
	Sp	۰/۷۶	۰/۷۲	۰/۸۳	۰/۸۱	۰/۸۰	۰/۹۷	۰/۶۵	۰/۷۹
	G-mean	۰/۶۳	۰/۵۴	۰/۷۸	۰/۷۹	۰/۸۰	۰/۹۸	۰/۶۵	۰/۷۳
Proposed Method	Ac	۰/۹۲	۰/۷۸	۰/۸۴	۰/۹۱	۰/۹۱	۰/۹۷	۰/۷۱	۰/۸۶
	Se	۱	۰/۷۱	۰/۷۹	۰/۹۳	۰/۹۰	۰/۹۸	۰/۷۱	۰/۸۶
	Sp	۰/۸۸	۰/۸۱	۰/۹۲	۰/۸۸	۰/۸۷	۰/۹۷	۰/۷۰	۰/۸۶
	G-mean	۰/۹۳	۰/۷۵	۰/۸۱	۰/۹۰	۰/۸۹	۰/۹۷	۰/۷۱	۰/۸۵



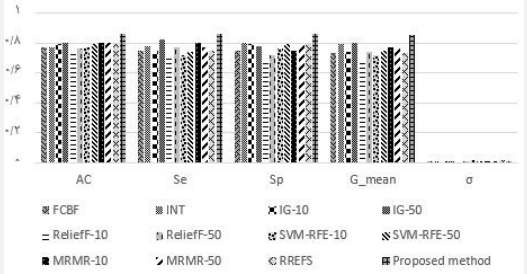
شکل (۱)- نمودار رادار میانگین اندازه‌های Ac، Sp، Se و G-mean برای سه طبقه‌بند C4.5، Naive-Bayes و SVM

در نمودار رادار شکل (۱)، یک چهارضلعی برای هر روش رسم شده است. به طوری که هر راس شکل مطابق با یک معیار می‌باشد. شکلی که دارای مساحت بیش‌تر و تقارن بالاتر باشد عمل‌کرد بهتری را نشان می‌دهد. طبق این توضیحات، مشخص است که روش پیشنهادی این مقاله از سایر روش‌های انتخاب ویژگی برتر می‌باشد.

میانگین عمل‌کرد سه طبقه‌بند برای روش‌های متفاوت از نظر ارزیابی اندازه‌ها (Ac، Se، Sp و G-mean)، واریانس خطا و روش‌های انتخاب ویژگی به ترتیب در شکل‌های (۲) و (۳) نمایش داده شده است. لازم به ذکر است که هر چه Ac، G-mean، Sp و Se بالاتر باشند نتایج طبقه‌بندی بهتر خواهد بود.

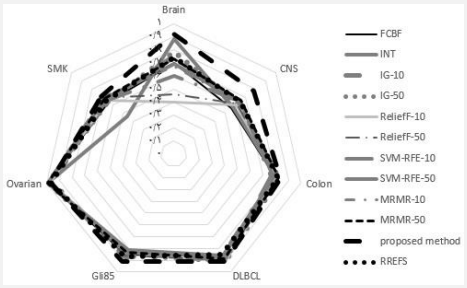


شکل (۲)- نمودار میله‌ای میانگین اندازه‌های Ac، Sp، Se و G-mean و واریانس خطا برای سه طبقه‌بند C4.5، Naive-Bayes و SVM

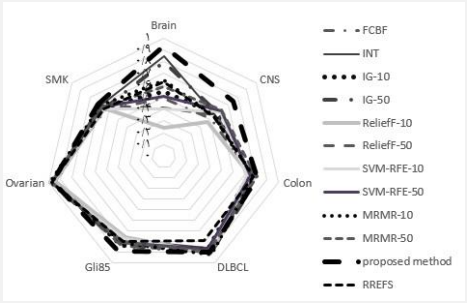


شکل (۳)- نمودار میله‌ای میانگین اندازه‌های Ac، Sp، Se و G-mean و واریانس خطا برای سه طبقه‌بند C4.5، Naive-Bayes و SVM

به منظور مقایسه‌ی دقیق روش پیشنهادی با سایر الگوریتم‌ها، میانگین طبقه‌بندی Ac و G-mean مربوط به هر یک از مجموعه‌ی داده‌ها نیز در نظر گرفته شده و در این راستا دو نمودار رادار در شکل‌های (۴) و (۵) نمایش داده شده است. به طوری که هر راس نمودار رادار نمایان‌گر یک مجموعه‌ی داده‌ی خاص می‌باشد. با توجه به شکل‌های (۴) و (۵) می‌توان برتری الگوریتم پیشنهادی را از نظر Ac و G-mean در مقایسه با سایر الگوریتم‌ها مشاهده کرد.



شکل (۴)- نمودار رادار مقادیر میانگین Ac برای سه طبقه‌بند C4.5، Naive-Bayes و SVM



شکل (۵)- نمودار رادار مقادیر میانگین اندازه‌ی G-mean برای سه طبقه‌بند C4.5، Naive-Bayes و SVM

۱۱- آزمون‌های آماری غیرپارامتری

هدف این بخش دست‌یابی به یک توجیه منصفانه‌ی آماری برای الگوریتم پیشنهادی در مقایسه با سایر الگوریتم‌ها است. بدین منظور آزمون رتبه‌بندی علامت‌دار ویلکاکسون غیرپارامتری[۱] روی میانگین دقت‌ها اجرا شده است. لازم به ذکر است که Exact p-value کوچک‌تر یا مساوی ۰/۰۵ به عنوان یک تفاوت قابل توجیه در نظر گرفته شده است. رتبه‌های مثبت و منفی آزمون رتبه‌بندی علامت‌دار ویلکاکسون برای الگوریتم‌های ذکر شده با توجه به معیار دقت در جدول (۴) ارائه شده است.

این آزمون روش‌هایی که Exact p-value آن‌ها کوچک‌تر از ۰/۰۵ باشد را رد می‌کند. بنابراین از جدول (۴) می‌توان نتیجه

[۱] Nonparametricwilcoxon Signed Rank Test

گرفت که روش پیشنهادی قادر است تا تمام الگوریتم‌ها را رد کند و این نشان می‌دهد که تفاوت قابل توجهی میان روش پیشنهادی و سایر الگوریتم‌های رد شده وجود دارد. در ادامه، مقایسه‌ی معناداری میان تمام الگوریتم‌ها در جدول (۵) ارائه شده است. در این جدول علامت ● نشان می‌دهد که الگوریتم در سطر، یکی از الگوریتم‌های موجود در ستون را رد می‌کند و علامت o نشان می‌دهد که الگوریتم موجود در ستون، الگوریتم موجود در سطر را رد می‌کند. همان‌گونه که در جدول (۵) مشاهده می‌شود، هیچ روشی قادر به رد الگوریتم پیشنهادی نبوده و الگوریتم پیشنهادی همه‌ی الگوریتم‌ها را رد می‌کند. لازم به ذکر است که از نرم افزار کیل (KEEL) برای اجرای این آزمون استفاده شده است [۱۵].

جدول (۴)- نتایج رتبه‌بندی ویلکاکسون برای روش پیشنهادی

روش	$\mathbb{R}^+$	$\mathbb{R}^-$	Exact P-value	Asymptotic P-value
FCBF	۲۷/۰	۱/۰	۰/۰۳۱۲۶	۰/۰۲۲۴۹۴
INT	۲۷/۰	۱/۰	۰/۰۳۱۲۶	۰/۰۲۲۴۹۴
IG-10	۲۸/۰	۰/۰	۰/۰۱۵۶۲۶	۰/۰۱۲۲۸۵
IG-50	۱۹/۰	۲/۰	۰/۰۹۳۷۶	۰/۰۵۹۱۷۲
ReliefF-10	۲۶/۵	۱/۵	۰/۰۳۹۰۷	۰/۰۲۷۹۹۲
ReliefF-50	۲۶/۵	۱/۵	۰/۰۳۹۰۷	۰/۰۲۴۷۷۹
RFE-10	۲۶/۰	۲/۰	۰/۰۴۶۸۸	۰/۰۳۴۶۱۱
RFE-50	۲۷/۰	۱/۰	۰/۰۳۱۲۶	۰/۰۲۲۴۹۴
MRMR-10	۱۹/۰	۲/۰	۰/۰۹۳۷۶	۰/۰۵۹۱۷۲
MRMR-50	۲۶/۰	۲/۰	۰/۰۴۶۸۸	۰/۰۲۱۰۰۲
RREFS	۲۱/۰	۰/۰	۰/۰۳۱۲۶	۰/۰۲۱۰۹۸

جدول (۵)- خلاصه‌ی آزمون رتبه‌بندی علامت‌دار ویلکاکسون برای الگوریتم‌ها

روش	FCBF	INT	IG 10	IG 50	ReliefF 10	ReliefF 50	RFE 10	RFE 50	MRMR 10	MRMR 50	RREFS	Proposed Method
FCBF	-			o								o
INT		-										o
IG-10			-									o
IG-50	●			-	●							o
ReliefF-10					-	o			o			o
ReliefF-50						-						o
RFE-10							-		o	o		o
RFE-50								-		o		o
MRMR-10					●				-			o
MRMR-50										-		o
RREFS											-	o
Proposed Method	●	●	●	●	●	●	●	●	●	●	●	-

۱۲- مقایسه‌ی هزینه‌ی محاسباتی روش پیشنهادی و روش RREFS

در این بخش هزینه‌ی محاسباتی روش پیشنهادی و روش RREFS مورد بررسی و مقایسه قرار گرفته است. فرض کنید که X یک ماتریس n×d (n تعداد نمونه‌ها و d تعداد ویژگی‌ها) و دارای n ستون مستقل خطی باشد (rank(X)=n). هم‌چنین فرض کنید که ماتریس B یک ماتریس پایه برای X باشد. طبق الگوریتم (۱)، روش پیشنهادی دارای سه مرحله‌ی اساسی شامل محاسبه‌ی ماتریس پایه‌ی B، ضرب‌های ماتریسی و به‌روزرسانی ماتریس‌های $\boldsymbol{H}\in\mathbb{R}^{k\times n}$ و $\boldsymbol{W}\in\mathbb{R}^{n\times k}$ است. در این راستا:

الف) فرض کنید که از فرایند توضیح داده شده در بخش ۵-۱ به منظور ساخت ماتریس پایه‌ی B استفاده شود. در این صورت از آن‌جا که این فرایند بر اساس استفاده از روش بهره‌وری اطلاعات (IG) بوده و سپس امتیاز IG مربوط به هر ویژگی محاسبه می‌شود، بنابراین با توجه به این‌که تعداد ویژگی‌ها برابر با d است، می‌توان نتیجه گرفت که هزینه‌ی محاسباتی مربوط به ساخت ماتریس پایه به صورت O(d) می‌باشد.

ب) محاسبه‌ی B^TBHH^T، W^TB^TBW و W^TB^TB به ترتیب دارای هزینه‌ی $O(kn^2)$، $O(kn^2+nk^2)$، $O(kn^2+nk^2)$ و $O(nk^2)$ است.

ج) برای به‌روزرسانی ماتریس H، هزینه‌ی محاسباتی جهت به دست آوردن معکوس ماتریس W^TB^TBW برابر با $O(k^3)$ است.

اکنون، با توجه به موارد اشاره شده، می‌توان نتیجه گرفت که هزینه‌ی محاسباتی روش پیشنهادی به صورت زیر است.

$$O(d+kn^2+nk^2+k^3) \quad (۵)$$

اگر فرض شود که تعداد ویژگی‌های انتخاب شده از تعداد کل نمونه‌ها و ویژگی‌ها کم‌تر باشد ($k\leq\{n,d\}$)، می‌توان استدلال کرد که هزینه‌ی محاسباتی روش پیشنهادی به صورت زیر است.

$$O(d+kn^2). \quad (۶)$$

در ارتباط با هزینه‌ی محاسباتی روش RREFS، با استناد به مقاله‌ی [۲]، این روش دارای دو مرحله‌ی اساسی است:

مرحله‌ی اول: استفاده از روش بهره‌وری اطلاعات برای مرتب کردن ویژگی‌ها بر اساس امتیاز IG مربوط به هر ویژگی که در نتیجه هزینه‌ی محاسباتی این مرحله به صورت O(d) می‌باشد.

مرحله‌ی دوم- استفاده از الگوریتم کاهش یافته‌ی سطری پلکانی به منظور یافتن ویژگی‌های مستقل خطی که در نتیجه هزینه‌ی محاسباتی این مرحله به صورت $O(n^3)$ می‌باشد.

در نتیجه با توجه به بحث فوق، هزینه‌ی محاسباتی روش RREFS به صورت زیر است.

$$O(d+n^3) \quad (۷)$$

بنابراین یک مقایسه بین هزینه‌ی محاسباتی روش پیشنهادی و روش RREFS نشان می‌دهد که روش پیشنهادی این مقاله از هزینه‌ی محاسباتی پایین‌تری برخوردار است.

۱۳- بحث و نتیجه‌گیری

هدف اصلی این مقاله معرفی یک روش جدید انتخاب ویژگی برای داده‌های میکرو-آرایه‌ای DNA است. برای این منظور با استفاده از ایده‌ی یادگیری زیرفضا و تجزیه‌ی ماتریس پایه برای زیرفضای برداری تولید شده توسط داده‌های میکرو-آرایه‌ای، مساله‌ی انتخاب ویژگی مورد بررسی قرار گرفته است. علاوه بر این به منظور منظم‌سازی پراکندگی و انتخاب ویژگی‌های با شباهت کم‌تر، از کمینه کردن نرم فروبنیوس ماتریس انتخاب ویژگی استفاده شده است.

کارایی روش پیشنهادی با استفاده از مجموعه‌ی داده‌های میکرو-آرایه‌ای DNA بررسی شده است که نتایج به دست آمده نشان دهنده‌ی برتری این روش بر چند روش انتخاب ویژگی مشهور بانظارت می‌باشد. هم‌چنین، نتایج مربوط به روش انتخاب ویژگی پیشنهاد شده در این مقاله، گویای این نکته است که استفاده از مفهوم پایه، از یک طرف بعد فضای ویژگی‌ها را تا حد مطلوبی کاهش داده و از طرف دیگر کارایی فرایند انتخاب ویژگی را به نحو محسوسی بهبود می‌بخشد.

در ادامه‌ی این تحقیق، بررسی ساختار هندسی نمونه‌ها و ویژگی‌ها و استفاده از مفهوم یادگیری منیفلد، می‌تواند موضوع قابل توجهی برای معرفی یک الگوریتم انتخاب ویژگی مبتنی بر ساخت پایه برای داده‌های میکرو-آرایه‌ای DNA به حساب آید.

۱۴- مراجع

[1] V. Bolon-Canedo, N. Sanchez-Marono, A. Alonso-Betanzos, J. M. Benıtez, F. Herrera, "A review of microarray datasets and applied feature selection methods," Inf. Sci., vol. 282, pp. 111-135, June, 2014.

[2] M. Ebrahimpour, M. Zare, M. Eftekhari, GH. Aghamolaei, "Occam's razor in dimension reduction: Using reduced row Echelon form for finding linear independent features in high dimensional microarray datasets," Eng. Appl. Artif. Intell., vol. 62, pp. 214-221, June, 2017.

[3] L. Yu, H. Liu, "Feature selection for high-dimensional data: A fast correlation-basedfilter solution," Proc. Proceedings of the 20thInternational Conference on Machine Learning, Washington D.C., USA, pp. 856-863, ICML, Aug., 2003.

[4] Z. Zhao, H. Liu, "Searching for interacting features in subset selection," Intell. Data. Anal., vol. 13, pp. 207-228, Apr., 2009.

[5] M. Hall, L. Smith,"Practical feature subset selection for machine learning," Proc. Proceedings of the 21st Australasian Computer Science Conference, Perth, Australia, pp. 181-191, ACSC, Feb., 1998.

[6] I. Kononenko, "Estimating attributes: analysisand extensions of RELIEF", in European conference on machine learning (ECML), Catania, Italy, 1994, pp. 171-182.

[7] H. Peng, F. Long, C. Ding, "Feature selectionbased on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach.Intell., vol. 27, pp. 1226-1238, Aug., 2005.

[8] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, "Gene selection for cancer classification using support vector machines," Mach. Learn., vol. 46, pp. 389-422, Jan., 2002.

[9] S. Wang, W. Pedrycz, Q. Zhu, W. Zhu, "Subspace learning for unsupervised feature selection via matrix factorization," Pattern Recognit., vol. 48, pp. 10-19, Aug., 2014.

[10] S. Wang, W. Pedrycz, Q. Zhu, W. Zhu, "Unsupervised feature selection via maximum projection and minimum redundancy," Knowl.-Based Syst., vol. 75, pp. 19-29, Nov., 2014.

[11] I. Jolliffe, Principal Component Analysis. feature Springer-Verlag, 1986.

[12] S. Roweis, L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, pp. 2323-2326, Dec., 2000.

[13] J. Tenenbaum, V. De Silva, J. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science, vol. 290, pp. 2319-2323, Dec., 2000.

[14] P. Glifani, H. Behnam, Z. Alizade Sani, "Analysis of Echocardiography images using manifold learning," Iran. Jour. Bio. Engin., vol. 4, pp. 149-160, Sep., 2010.

[15] J. Alcala-Fdez, L. Sanchez, S. Garcia, M. del Jesus, S. Ventura, J. Bacardit, V. Rivas, others, "KEEL: a software tool to assess evolutionary algorithms for data mining problems,"