DOI: 10.22041/ijbme.2014.13555

Two Stages Sparse Representation-based Classifier and its Application for Cancer Classification

M. Miri [1], M. T. Sadeghi [2*], V. Abootalebi[3]

[1] M.Sc., Signal Processing Research Lab, Electrical and Computer Engineering Department, Yazd University, Yazd, Iran, m.miri@stu.yazd.ac.ir
[2*] Assistant Professor, Signal processing Research Lab, Electrical and Computer Engineering Department, Yazd University, Yazd, Iran.
[3] Assistant Professor, Signal processing Research Lab, Electrical and Computer Engineering Department, Yazd University, Yazd, Iran, abootalebi@yazd.ac.ir

Abstract

Successful outcomes of Sparse Representation-based Classifier (SRC) and Sparse Subspace Clustering (SSC) in many applications motivated us to combine these methods and propose a hierarchical classifier. The main idea behind the SRC and SSC algorithms is to represent a data using a sparse linear combination of elementary signals so that those elementary signals which are similar to the data contribute mainly in the representation. In this paper, the performance of a sparse representation based classifier is improved by pre-clustering of training samples using the SSC algorithm. A two-stage SRC is then designed using the resulting clusters. A test data is classified by first determining the most similar cluster. The data label is subsequently found using the second stage classifier. The performance of the proposed method is evaluated considering cancer classification problem using the 14-Tumors microarray dataset. Due to low number of data samples per each class and high dimensionality of the data, this is a challenging problem. Curse of dimensionality, overfitting of the classifier to the training data and computational complexity are the possible related problems. Our experimental results show that the proposed method outperforms some other state of the art classifiers.

Key words: Sparse Subspace Clustering, Microarray data, Cancer classification, Hierarchical classifier, Sparse Representation-based Classification, Sparse representation.

* **Corresponding author**
Address: *Electrical and Computer Engineering Department, Yazd University, P.O.Box: 89195-741, Postal Code: 77141, Yazd, I.R. Iran*
Tel: *+98 351 8122389*
Fax: *+98 351 8200144*
E-mail: *m.sadeghi@yazd.ac.ir*

DOI: 10.22041/ijbme.2014.13555

طبقه‌بندی‌کنندهٔ دو مرحله‌ای مبتنی بر نمایش تنک و کاربرد آن در تشخیص سرطان

ملیحه میری[۱]، محمدتقی صادقی[۲*]، وحید ابوطالبی[۳]

[۱] دانشجوی کارشناسی ارشد،گروه مخابرات، دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد، یزد، ایران m.miri@stu.yazd.ac.ir

[۲] استادیار، گروه مخابرات، دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد، یزد، ایران.

[۳] استادیار، گروه مخابرات، دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد، یزد، ایران abootalebi@yazd.ac.ir

چکیده

با توجه به نتایج موفقیت‌آمیز طبقه‌بندی‌کننده مبتنی بر نمایش تنک (SRC) و خوشه‌بندی زیرفضای تنک (SSC) در کاربردهای مختلف، در این مقاله با ترکیب این دو روش، یک روش طبقه‌بندی سلسله مراتبی ارائه می‌شود. ایده اصلی در روش‌های طبقه‌بندی و خوشه‌بندی مبتنی بر نمایش تنک، نمایش هر داده به صورت ترکیب خطی تنک از سایر داده‌ها است به گونه‌ای که داده‌های مشابه با داده مورد نظر در این ترکیب خطی بیشترین وزن را به خود اختصاص دهند. در روش پیشنهادی، به منظور دست‌یابی به صحت طبقه‌بندی بیشتر، ابتدا داده‌های آموزشی با استفاده از روش خوشه‌بندی زیرفضای تنک بخش‌بندی می‌شوند. سپس با استفاده از شیوهٔ بکار گرفته شده در طبقه‌بندی‌کننده مبتنی بر نمایش تنک، طبقه‌بندی‌کننده‌ای دو مرحله‌ای طراحی می‌شود. در مرحلهٔ اول، خوشه‌ای که داده ورودی بیشترین شباهت را با آن دارد تعیین شده و در مرحله بعد طبقهٔ مربوطه (برچسب داده) تعیین می‌شود. برای ارزیابی روش پیشنهادی از دادگان ریزآرایه **14-Tumors** – که حاوی اطلاعات مربوط به ۱۴ نوع سرطان مختلف است– استفاده شده است. از جمله ویژگی‌های این دادگان تعداد زیاد بعد در مقابل تعداد کم نمونه در هر دسته است که عمل طبقه‌بندی آن‌ها را به مسأله‌ای چالش‌برانگیز تبدیل می‌کند. ابعاد زیاد داده‌ها نه تنها مشکلاتی از جمله نفرین ابعاد و بیش انطباق طبقه‌بندی‌کننده به داده‌های آموزشی را به دنبال دارد، بلکه باعث افزایش پیچیدگی محاسباتی شده؛ زمان لازم را برای اجرای الگوریتم‌ها افزایش می‌دهد. آزمایش‌های انجام شده بر این دادگان با استفاده از روش پیشنهادی نشان می‌دهد که در مقایسه با سایر روش‌های طبقه‌بندی، این روش به نتایج بهتری منجر می‌شود.

کلیدواژگان: خوشه‌بندی زیرفضای تنک، دادگان ریزآرایه، طبقه‌بندی سرطان، طبقه‌بندی‌کننده سلسله مراتبی، طبقه‌بندی‌کننده مبتنی بر نمایش تنک، نمایش تنک.

*عهده‌دار مکاتبات

نشانی: یزد، صفائیه، چهارراه پژوهش، دانشگاه یزد، دانشکدهٔ مهندسی برق و کامپیوتر، کد پستی: ۷۷۱۴۱

تلفن: ۸۱۲۲۳۸۹-۰۳۵۱، دورنگار: ۸۲۰۰۱۴۴-۰۳۵۱، پیام نگار: m.sadeghi@yazd.ac.ir

۱- مقدمه

در سال‌های اخیر شاهد توسعه روزافزون استفاده از مفاهیم نمونه‌برداری فشرده (CS)[1] و نمایش تنک (SR)[2] در کاربردهای مختلف پردازش سیگنال هستیم **[۲،۱]** و تحقیقات بسیاری در این زمینه شکل گرفته است. در بسیاری از کاربردهای شناسایی الگو و بینایی ماشین با داده‌هایی با ابعاد زیاد از قبیل تصاویر چهره و یا ریزآرایه‌ها سروکار داریم.

بعد زیاد داده‌ها نه تنها موجب افزایش حافظه مورد نیاز و پیچیدگی محاسباتی الگوریتم‌ها می‌شود، به دلیل اثر نویز و تعداد کم داده‌ها در مقایسه با تعداد ابعاد، موجب کاهش کارایی الگوریتم‌ها نیز می‌شود که از این موضوع به عنوان نفرین ابعاد[3] یاد می‌شود**[۳]**. داده‌های با ابعاد زیاد اغلب در زیرفضاهایی با بعد کمتر گسترده شده‌اند. روش‌های مختلف خوشه‌بندی[4] به عنوان ابزاری برای جداسازی داده‌های متناظر با هر یک از این زیرفضاها به‌کار می‌روند.

اخیراً به استفاده از روش‌های مبتنی بر نمایش تنک در خوشه‌بندی داده‌ها نیز توجه بسیار شده است**[۵،۴]**. از بین تمام نمایش‌های ممکن برای یک داده بر حسب سایر داده‌ها، یک نمایش تنک متناظر با انتخاب تعداد کمی از داده‌ها از همان زیرفضای داده مربوطه است. چنین نمایشی با حل مسأله بهینه‌سازی تنک بدست می‌آید. روش‌های متداول خوشه‌بندی از جمله روش‌های تکراری **[۶]**، روش‌های جبری **[۷]**، MPPCA[5] **[۸]**، RANSAC[6] **[۹]** و روش‌های مبتنی بر خوشه‌بندی طیفی[7] **[۱۰]** دارای محدودیت‌هایی هستند که استفاده از آنها را در عمل با مشکل مواجه می‌کند. برخی از این روش‌ها نیاز به داشتن اطلاعات اولیه‌ای از جمله تعداد و بعد خوشه‌ها دارند. همچنین بار محاسباتی زیاد - که بصورت نمایی با تعداد و بعد خوشه‌ها افزایش می‌یابد- و عدم مقاومت نسبت به نویز و خطای مدل‌سازی از دیگر معایب این روش‌ها است**[۱۱]**. در مقایسه با این روش‌ها، روش خوشه‌بندی زیرفضای تنک (SSC)[8] که از حل معادلات بهینه‌سازی محدب[9] برای یافتن نمایش‌های تنک استفاده می‌کند، از پیچیدگی محاسباتی کمتری برخوردار است. همچنین این روش به مقداردهی اولیه نیاز ندارد و نسبت به نویز نیز مقاوم است**[۴]**.

در سال ۲۰۰۹، رایت[10] استفاده از نمایش تنک سیگنال‌ها در طبقه‌بندی را پیشنهاد کرد، این روش - که طبقه‌بندی مبتنی بر نمایش تنک (SRC)[11] نامیده می‌شود- به نتایج رضایت‌بخشی بویژه در سیستم‌های تشخیص چهره منجر شده است**[۱۲]**. ایده اصلی این روش آن است که هر تصویر چهره به صورت ترکیب خطی تنک از سایر تصاویر چهره قابل بیان است و بیشترین ضرایب این ترکیب خطی متعلق به تصاویری است که در دسته‌ای یکسان با تصویر ورودی قرار دارند؛ لذا با صفر کردن سایر ضرایب و تنها استفاده از این ضرایب بزرگتر می‌توان به بازسازی قابل قبولی از تصویر ورودی دست یافت. با استفاده از معیار فاصله اقلیدسی میزان شباهت تصویر بازسازی شده با کمک تصاویر دسته‌های مختلف، با تصویر ورودی تعیین می‌شود و برچسب دسته‌ای که کمترین خطای بازسازی را تولید کند، به تصویر ورودی تعلق می‌گیرد.

پیشرفت‌های اخیر در زمینه زیست‌شناسی مولکولی و فناوری ریزآرایه[12]، امکان پایش سطح بیان[13] هزاران ژن (میزان فعال بودن ژن‌ها) را به صورت همزمان فراهم کرده است و مقادیر هنگفتی را از داده‌های ریزآرایه تولید کرده است. این داده‌ها در تشخیص و طبقه‌بندی انواع بافت‌های سرطانی نقش به‌سزایی دارند؛ اما مهمترین چالش در مورد این دادگان را می‌توان ابعاد بسیار بالای آنها در مقایسه با تعداد کم نمونه‌ها دانست که طراحی طبقه‌بندی‌کننده‌های مناسب را دچار مشکل می‌کند. اگرچه تاکنون تلاش‌های بسیاری با هدف یافتن طبقه‌بندی‌کننده‌های با صحت زیاد و پیچیدگی محاسباتی کم انجام شده است، بیشتر مطالعاتی که در این زمینه انجام شده توجه خود را به دادگانی با تعداد دسته‌های کم (دو یا سه نوع سرطان مختلف) معطوف کرده‌اند و از روش‌هایی مثل تحلیل جداکننده خطی (LDA)[14]، شبکه‌های عصبی، خوشه‌بندی، نزدیک‌ترین همسایگی (NN)[15] و ماشین بردار پشتیبان (SVM)[16] استفاده کرده‌اند**[۱۵،۱۴،۱۳]**. دادگان 14-Tumors –

[1]Compressive Sensing
[2]Sparse Representation
[3]Curse of Dimensionality
[4]Clustering
[5]Mixture of Probabilistic Principal Component Analysis
[6]Random Sample Consensus
[7]Spectral Clustering
[8]Sparse Subspace Clustering
[9]Convex Optimization
[10]Wright
[11]Sparse Representation-based Classification
[12]Microarray
[13]Expression
[14]Linear Discriminant Analysis
[15]Nearest Neighbor
[16]Support Vector Machine

که در این تحقیق بررسی شده‌اند- شامل اطلاعات مربوط به چهارده نوع سرطان مختلف است که تا کنون اغلب از دید انتخاب ویژگی[17] بررسی شده است[۱۶]. از جمله پژوهش‌های انجام شده در طبقه‌بندی این پایگاه داده می‌توان به روش ارائه شده در مرجع [۱۶] اشاره کرد. در این مقاله مسأله طبقه‌بندی چهارده دسته سرطان مختلف به چند مسأله طبقه‌بندی باینری تبدیل، و از طبقه‌بندی‌کننده‌های مختلفی نیز استفاده شده است که بهترین عملکرد مربوط به طبقه‌بندی‌کننده SVM بوده است. در طبقه‌بندهای باینری با روش‌های OVO[18] و OVA[19] عمل طبقه‌بندی انجام شده است که به ترتیب از $\frac{c(c-1)}{2}$ و c ابرصفحه برای جداسازی دسته‌ها استفاده می‌کنند که c تعداد دسته‌ها است. در منابع [۱۷،۱۸] با استفاده از روش‌های مبتنی بر شبکه عصبی چهارده نوع سرطان طبقه‌بندی شده‌اند؛ به عنوان مثال در مرجع [۱۷] از دو طبقه‌بندی‌کننده ANN[20] متوالی برای طبقه‌بندی استفاده شده است. در این روش ابتدا با استفاده از طبقه‌بندی‌کننده اول، دو دسته‌ای که بیشترین احتمال تعلق داده آزمون به آن‌ها وجود دارد انتخاب شده؛ سپس با اعمال طبقه‌بندی‌کننده دوم و انجام طبقه‌بندی باینری بین دو دسته انتخابی، برچسب داده آزمون تعیین می‌شود.

طبقه‌بندی‌کننده‌ای فازی نیز در مطالعهٔ [۱۹] ارائه شده است (CD-MFS)[21]، که با استفاده از قوانین فازی، طبقه‌بندی را انجام می‌دهد. مرجع [۲۰] نیز از روشی مبتنی بر نمایش تنک با عنوان SR برای جداسازی دسته‌های مختلف استفاده کرده است و به نتایج بهتری در مقایسه با تعمیم‌های مختلف روش SVM دست یافته است.

با وجود تحقیقاتی که در این زمینه شکل گرفته‌اند، همچنان دسته‌بندی این دادگان با مقادیر خطای قابل توجهی همراه است و به صورت بالینی قابل استفاده نیست.

در این مقاله با ترکیب دو روش خوشه‌بندی و طبقه‌بندی مبتنی بر نمایش تنک، روش طبقه‌بندی سلسله مراتبی ارائه شده است. بدین منظور در مرحله آموزش، داده‌های آموزشی خوشه‌بندی شده؛ سپس آزمایش سیستم در دو مرحله انجام می‌شود. در مرحله اول خوشه مربوط به داده ورودی مشخص می‌شود و در مرحله بعد با توجه به خوشه تعیین شده، عمل طبقه‌بندی انجام شده؛ برچسب داده آزمایشی تعیین می‌شود.

با توجه به این که یکی از روش‌های متداول در دسته‌بندی انواع سرطان‌ها خوشه‌بندی داده‌های مربوط به آنها است، انتظار می‌رود روش پیشنهادی - که در مرحله اول داده آزمون را خوشه‌بندی و سپس آن را طبقه‌بندی می‌کند- به نتایج مناسبی در این حوزه دست یابد. مقایسه نتایج به دست آمده از روش معرفی شده با نتایج حاصل از تحقیقات قبلی نشان دهنده عملکرد قابل قبول سیستم پیشنهادی است.

ساختار این مقاله به قرار زیر است: در بخش ۲ مقدمه‌ای بر نمایش تنک بیان شده است، سپس به توضیح اجمالی طبقه‌بندی‌کننده مبتنی بر نمایش تنک و خوشه‌بندی زیرفضای تنک پرداخته شده است. در بخش ۳ روش پیشنهادی و در بخش ۴ توضیح مختصری در مورد پایگاه داده استفاده شده و همچنین نتایج آزمایش‌های انجام شده آورده شده است. در نهایت در بخش ۵ جمع‌بندی و نتیجه‌گیری ارائه شده است.

۲- نمایش تنک (SR)

امروزه نمایش تنک به ابزاری قوی برای نمایش و فشرده‌سازی سیگنال‌ها تبدیل شده است. بر اساس این نمایش می‌توان سیگنالی را بصورت ترکیب خطی تنک از داده‌های آموزشی تقریب زد[۲]. در حالت کلی ترکیبی خطی از داده‌های آموزشی را می‌توان به صورت یک دستگاه معادلات خطی به شکل $\mathbf{y} = \mathbf{Ys}$ نشان داد. که در آن ماتریس $\mathbf{Y} = [\mathbf{Y}_1, ..., \mathbf{Y}_c]$ - که از کنار هم قرار گرفتن تمام داده‌های آموزشی مربوط به $\mathbf{c}$ دسته تشکیل شده است- ماتریس دیکشنری نامیده شده و ستون‌های آن اتم نامیده می‌شوند. اگر در این دستگاه تعداد معادلات از تعداد مجهولات کمتر باشد (تعداد ستون‌های $\mathbf{Y}$ بیشتر از تعداد سطرهای آن باشد)، دستگاه از نوع فرومعین[22] بوده و دارای بی‌شمار جواب است. با اعمال شرط تنکی بر جواب حاصل از دستگاه فوق می‌توان به پاسخی یکتا دست یافت. از آنجا که نرم صفر یک بردار معرف تعداد مؤلفه‌های غیر صفر آن است، کمینه‌سازی نرم صفر منجر به تحمیل شرط تنکی بر بردار خواهد شد؛ بنابراین

[17] Compressive Sensing
[18] One-Versus-One
[19] One-Versus-All
[20] Artificial Neural Networks
[21] Cancer Diagnosis with Memetic Fuzzy System
[22] Under-Determined

از رابطه زیر برای به دست آوردن نمایش تنک یک بردار بر نمونه‌های آموزشی استفاده می‌شود:

$$\min \|\mathbf{s}\|_0 \quad s.t. \quad \mathbf{y} = \mathbf{Y}\mathbf{s} \qquad (۱)$$

$\mathbf{y}$ سیگنال ورودی و $\mathbf{s}$ بردار ضرایب ترکیب خطی است و $\|\mathbf{s}\|_0$ نیز تعداد مؤلفه‌های غیر صفر آن را نشان می‌دهد. $\mathbf{s}$ را $\mathbf{k}$- تنک گویند اگر دارای حداکثر $\mathbf{k}$ مؤلفهٔ غیرصفر باشد.

همان‌گونه که قبلاً ذکر شد اگر ماتریس دیکشنری فراکامل[۲۳] باشد، دستگاه معادلات فوق پاسخی یکتا و تنک خواهد داشت. چون نرم صفر محدب و مشتق‌پذیر نیست، مسأله فوق NP-Complete[۲۴] بوده؛ حل آن منوط به جستجوی کامل است که در ابعاد زیاد غیر ممکن است. روشی رایج در حل سیستم‌های خطی با فرض تنک بودن جواب، تقریب نرم صفر با نرمی از مرتبه بالاتر است که کمینه کردن آن ساده‌تر باشد. بدین منظور برای تعیین نمایش تنک سیگنال از فرم زیر استفاده می‌شود که در آن شرط تنکی با حداقل‌سازی نرم l_1 تأمین می‌شود[۲]:

$$\min \|\mathbf{s}\|_1 \quad s.t. \quad \mathbf{y} = \mathbf{Y}\mathbf{s} \qquad (۲)$$

از آنجا که نرم یک محدب است می‌توان کمینه‌سازی آن را با استفاده از روش‌های بهینه‌سازی محدب انجام داد؛ همچنین می‌توان آن را به صورت مسأله خطی بیان کرد و با استفاده از روش‌های برنامه‌نویسی خطی مثل الگوریتم جستجوی پایه (BP)[۲۵] آن را حل کرد[۲۱]. از مهم‌ترین معایب این روش پیچیدگی زیاد محاسباتی و در نتیجه زمان‌بر بودن اجرای آن است؛ بنابراین اغلب با روش‌های دیگری جایگزین می‌شود.

یکی دیگر از روش‌های رایج در به دست آوردن نمایش تنک، مناسب‌ترین ارتباط یا MP[۲۶] نام دارد[۲۲]. این روش به صورت تکراری و حریصانه عمل می‌کند، به این صورت که در هر گام تنها یکی از ضرایب تنک را مشخص می‌کند. اگر $\mathbf{s}$ برداری $\mathbf{k}$- تنک باشد، $\mathbf{y}$ می‌تواند به صورت یک ترکیب خطی از $\mathbf{k}$ اتم دیکشنری نوشته شود. در این روش ابتدا ستون‌های مورد استفاده از ماتریس دیکشنری در ترکیب خطی آشکار شده؛ سپس ضرایب این ستون‌ها – که همان مقادیر غیرصفر بردار $\mathbf{s}$ هستند- با حل مسأله حداقل مربعات محاسبه می‌شوند. در این روش در هر مرحله یک اتم از دیکشنری که بیشترین شباهت را به داده آزمون دارد، به عنوان عضو فعال در ترکیب خطی در نظر گرفته شده؛ ضریب مربوط به آن محاسبه می‌شود. تفاضل حاصل‌ضرب این تقریب ۱- تنک در دیکشنری از داده آزمون را به عنوان باقی‌مانده در نظر گرفته؛ مراحل فوق تکرار می‌شوند. در هر مرحله جمع تقریب‌های ۱- تنک به دست آمده با تقریب‌های قبلی به عنوان تقریب جدید در نظر گرفته می‌شوند و این روند تا جایی ادامه می‌یابد که یا تعداد مراحل مشخصی طی شود و یا خطا از مقدار معینی کمتر شود. در روش OMP[۲۷] – که تعمیم‌یافته روش MP است- در هر مرحله ضرایب ستون‌های فعال از ماتریس دیکشنری به صورت مستقل از نتایج مراحل قبل انتخاب می‌شوند و از نتایج قبلی تنها در یافتن مکان مؤلفه‌های غیرصفر استفاده می‌شود. از جمله مهم‌ترین مزایای این روش‌ها می‌توان به سرعت زیاد آن‌ها اشاره کرد[۲۳]. از جمله دیگر الگوریتم‌هایی که در محاسبه نمایش تنک استفاده می‌شوند، می‌توان به Homotopy، PALM[۲۸] و DALM[۲۹] اشاره کرد[۲۴].

اگر N_i داده آموزشی متعلق به دستهٔ iام در ستون‌های ماتریس $\mathbf{Y}_i$ به صورت $\mathbf{Y}_i = [\mathbf{y}_{i,1}, \mathbf{y}_{i,2}, ..., \mathbf{y}_{i,N_i}] \in \mathbf{R}^{D \times N_i}$ قرار گرفته باشند؛ بر اساس نمایش تنک چنانچه $\mathbf{y} \in \mathbf{R}^D$ داده ورودی متعلق به دسته iام باشد، می‌توان آن را برحسب ستون‌های ماتریس $\mathbf{Y}_i$ نمایش داد:

$$\begin{aligned} \mathbf{y} &= s_{i,1}\mathbf{y}_{i,1} + s_{i,2}\mathbf{y}_{i,2} + ... + s_{i,N_i}\mathbf{y}_{i,N_i} \\ &= \sum_{j=1}^{N_i} s_{i,j}\mathbf{y}_{i,j} \end{aligned} \qquad (۳)$$

که در این رابطه $s_{i,j}$ ها ضرایب اسکالر هستند. از این ایده در طبقه‌بندی‌کننده مبتنی بر نمایش تنک استفاده می‌شود.

۳- ۲-۱- طبقه‌بندی مبتنی بر نمایش تنک (SRC)

طبقه‌بندی با استفاده از نمایش تنک سیگنال شامل دو مرحله است: ابتدا داده ورودی با استفاده از ماتریس دیکشنری بصورت تنک کد می‌شود؛ سپس با استفاده از این ضرایب تنک و با بازسازی داده آزمایشی، طبقه‌بندی انجام

[23] Over-Complete
[24] Non-Polynomial Time
[25] Basis Pursuit
[26] Matching Pursuit
[27] Orthogonal Matching Pursuit
[28] Primal Augmented Lagrangian Methods
[29] Dual Augmented Lagrangian Methods

می‌شود[۱۲].

از آنجا که ابتدا برچسب داده آزمون برای سیستم مجهول است، ماتریس $\mathbf{Y}$ به صورت ترکیبی از داده‌های آموزشی از تمام c دسته در نظر گرفته می‌شود:

$$\mathbf{Y}=[\mathbf{Y}_1,...,\mathbf{Y}_c]=[\mathbf{y}_{1,1},\mathbf{y}_{1,2},...,\mathbf{y}_{c,N_c}] \quad (۴)$$

در این رابطه هر کدام از $\mathbf{Y}_i$ ها زیر ماتریس‌هایی هستند که داده‌های دسته iام را در خود جای داده‌اند. حال می‌توان با استفاده از رابطه (۲) نمایش تنک دادهٔ ورودی $\mathbf{y}$ را برحسب ستون‌های ماتریس دیکشنری $\mathbf{Y}$ بدست آورد.

در SRC از میزان شباهت داده ورودی به داده‌های موجود در هر یک از دسته‌ها برای طبقه‌بندی استفاده می‌شود. بدین منظور تابع $\delta_i:\mathbf{R}^N\to\mathbf{R}^N$ بصورت زیر تعریف می‌شود و با اعمال آن بر بردار $\mathbf{s}$، مؤلفه بردار $\mathbf{s}$ متناظر با داده‌های دسته iام در دیکشنری حفظ شده؛ سایر مؤلفه‌ها صفر می‌شوند.

$$\delta_i(\mathbf{s})=[0,...,0,s_{i,1},...,s_{i,N_i},0,...,0]^T\in\mathbf{R}^N \quad (۵)$$

بنابراین $\hat{\mathbf{y}}_i=\mathbf{Y}\delta_i(\mathbf{s})$ دادهٔ $\mathbf{y}$ را تنها به صورت ترکیب خطی از داده‌های دسته iام بیان می‌کند. با اعمال این تابع بر نمایش تنک دادهٔ $\mathbf{y}$ و بازسازی آن با استفاده از داده‌های متعلق به هر یک از دسته‌ها در هر مرحله، در نهایت داده ورودی به دسته‌ای نسبت داده می‌شود که میزان باقی‌مانده زیر را حداقل کند:

$$\min_i r_i(\mathbf{y})=\left\|\mathbf{y}-\mathbf{Y}\delta_i(\hat{\mathbf{s}})\right\|_2 \quad (۶)$$

الگوریتم زیر به طور خلاصه این روش طبقه‌بندی را بیان می‌کند:

ورودی: ماتریس دیکشنری برای c طبقهٔ مختلف $\mathbf{Y}=[\mathbf{Y}_1,...,\mathbf{Y}_c]=[\mathbf{y}_{1,1},\mathbf{y}_{1,2},...,\mathbf{y}_{c,N_c}]$، دادهٔ ورودی $\mathbf{y}$

الف) بهنجار کردن ستون‌های ماتریس دیکشنری $\mathbf{Y}$ تا نرم l_2 واحد داشته باشند.

ب) حل مسألهٔ حداقل‌سازی نرم l_1 زیر:

$$\min\|\mathbf{s}\|_1 \quad s.t. \quad \mathbf{y}=\mathbf{Y}\mathbf{s}$$

ج) محاسبهٔ باقیماندهٔ $r_i(\mathbf{y})=\left\|\mathbf{y}-\mathbf{Y}\delta_i(\hat{\mathbf{s}})\right\|_2$ برای تمام طبقه‌ها $i=1,...,c$

خروجی: برچسب داده $\mathbf{y}$: $identity(\mathbf{y})=\arg\min_i r_i(\mathbf{y})$

۴- ۲-۲- خوشه‌بندی زیرفضای تنک (SSC)

فرض کنید $\{\mathbf{y}_j\in\mathbf{R}^D\}_{j=1}^N$ مجموعه داده‌های موجود روی n زیرفضای مستقل خطی $\{S_i\}_{i=1}^n$ با بعد $\{d_i\ll D\}_{i=1}^n$ باشند و $\mathbf{Y}_i\in\mathbf{R}^{D\times N_i}$ مجموعه N_i داده متعلق به زیرفضای iام باشد؛ از آنجا که نمی‌دانیم کدام داده‌ها به کدام زیرفضا تعلق دارد، ماتریس داده را بصورت:

$$\mathbf{Y}=[\mathbf{y}_1,\mathbf{y}_2,...,\mathbf{y}_N]=[\mathbf{Y}_1,\mathbf{Y}_2,...,\mathbf{Y}_n]\in\mathbf{R}^{D\times N} \quad (۷)$$

تشکیل می‌دهیم که $N=\sum_{i=1}^n N_i$ تعداد کل داده‌هاست. اگر $\mathbf{y}_i$ داده‌ای از زیرفضای S_i باشد، می‌تواند به‌صورت ترکیبی خطی از d_i نقطه از این زیرفضا نوشته شود. به‌عبارتی $\mathbf{y}_i$ نمایشی d_i-تنک خواهد داشت که به‌صورت پاسخ تنک معادله (۲) قابل بازیابی است و در آن برای $j\neq i$ ، $\mathbf{s}_i\neq 0$ و $\mathbf{s}_j=0$ خواهد بود.

حال برای تعیین خوشه‌بندی داده‌ها به کمک نمایش‌های تنک بدست آمده، از روش خوشه‌بندی طیفی استفاده می‌شود. در این روش با استفاده از اطلاعات محلی حول هر داده، شباهت‌هایی بین جفت داده‌ها تعریف می‌شود. سپس عمل خوشه‌بندی با استفاده از ماتریس شباهت[۳۰] داده‌ها و تقسیم داده‌ها به چندین گروه به گونه‌ای که داده‌های هر گروه با هم مشابه و داده‌های گروه‌های مختلف با یکدیگر متفاوت باشند، انجام می‌شود[۱۰].

اگر $\mathbf{Y}_{\hat{i}}\in\mathbf{R}^{D\times N-1}$ ماتریسی باشد که از حذف ستون iام $\mathbf{Y}$ به دست آید، $\mathbf{y}_i$ نمایشی تنک بر آن خواهد داشت که می‌تواند از رابطه زیر محاسبه شود:

$$\min\|\mathbf{c}_i\|_1 \quad s.t \quad \mathbf{y}_i=\mathbf{Y}_{\hat{i}}\mathbf{c}_i \quad (۸)$$

پاسخ بهینه $\mathbf{c}_i\in\mathbf{R}^{N-1}$ برداری است که مؤلفه‌های غیرصفر آن متناظر با ستون‌هایی از $\mathbf{Y}_{\hat{i}}$ هستند که در همان زیرفضای $\mathbf{y}_i$ قرار دارند. پس از حل معادله فوق برای تمام نقاط $i=1,...,N$ ماتریس ضرایب $\mathbf{C}=[\hat{\mathbf{c}}_1,\hat{\mathbf{c}}_2,...,\hat{\mathbf{c}}_N]$ به دست می‌آید که در آن $\hat{\mathbf{c}}_i$ با صفر قرار دادن در سطر iام $\mathbf{c}_i$ بدست آمده است. با استفاده از این ماتریس، گراف $G=(V,E)$ به گونه‌ای تعریف می‌شود که گوشه‌های V در آن متناظر با N نقطه داده باشند. یال $(v_i,v_j)\in E$ در صورتی وجود خواهد

[30] Similarity Matrix

داشت که داده $\mathbf{y}_j$ درنمایش تنک $\mathbf{y}_i$ حاضر باشد. از آنجا که اگر $\mathbf{y}_i$ بتواند بصورت ترکیب خطی از تعدادی نقاط یک زیرفضا شامل $\mathbf{y}_j$ نوشته شود، آنگاه $\mathbf{y}_j$ نیز می‌تواند بصورت ترکیب خطی از نقاط همان زیرفضا شامل $\mathbf{y}_i$ نوشته شود؛ لذا ماتریس مجاورت[۳۱] گراف فوق بصورت $\tilde{\mathbf{C}} = |\mathbf{C}| + |\mathbf{C}|^T$ خواهد بود[۳].

پس از تشکیل گراف G انتظار می‌رود تمام گوشه‌هایی که مربوط به داده‌های یک زیرفضا هستند مجموعه‌ای متصل به یکدیگر را در گراف تشکیل دهند، در حالی‌که گوشه‌های نمایش‌دهنده داده‌های متعلق به زیرفضاهای متفاوت هیچ یال مشترکی نداشته باشند. لذا $\tilde{\mathbf{C}}$ نمایشی قطری بلوکی به شکل زیر پیدا می‌کند:

$$\tilde{\mathbf{C}} = \begin{bmatrix} \tilde{\mathbf{C}}_1 & 0 & \cdots & 0 \\ 0 & \tilde{\mathbf{C}}_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \tilde{\mathbf{C}}_n \end{bmatrix} \qquad (۹)$$

سپس ماتریس لاپلاسین گراف شباهت به‌صورت $\mathbf{L} = \mathbf{D} - \tilde{\mathbf{C}}$ تشکیل می‌شود که $\mathbf{D} \in \mathbf{R}^{N \times N}$ ماتریسی قطری است که بصورت $\mathbf{D}_{ii} = \sum_j \tilde{c}_{ij}$ محاسبه می‌شود.

تعداد مقادیر ویژه صفر ماتریس لاپلاسین متناظر با گراف G برابر با تعداد مؤلفه‌های متصل گراف است. در حالت ایده‌ال، داده‌های متعلق به n زیرفضای مستقل خطی دارای n مؤلفه متصل در گراف شباهت خواهند بود. بنابراین، وقتی تعداد زیرفضاها معلوم نباشد می‌توان آن را به‌صورت تعداد مقادیر ویژه صفر ماتریس $\mathbf{L}$ تخمین زد. در نهایت با اعمال الگوریتم K-means به n بردار ویژه متناظر با n مقدار ویژه کوچکتر ماتریس لاپلاسین، خوشه‌بندی داده‌ها بدست می‌آید[۴].

۵- روش پیشنهادی

در ادامه با ترکیب دو روش خوشه‌بندی و طبقه‌بندی مبتنی بر نمایش تنک، روش طبقه‌بندی سلسله مراتبی[۳۲] پیشنهاد شده است. ایده اصلی این روش، خوشه‌بندی داده‌ها و سپس انجام طبقه‌بندی در هر یک از خوشه‌های بدست آمده است. بدین صورت که پس از بدست آوردن نمایش تنک داده ورودی بر داده‌های آموزشی، به جای بازسازی سیگنال روی تمام دسته‌ها، این عمل تنها روی شبیه‌ترین داده‌های آموزشی به داده ورودی انجام شود که این داده‌ها با انجام خوشه‌بندی مشخص می‌شوند.

در این روش، داده‌های آموزشی در مرحله آموزش با استفاده از الگوریتم SSC خوشه‌بندی می‌شوند. بدین منظور ابتدا با استفاده از رابطه (۸) نمایش تنک هر یک از داده‌های آموزشی روی سایر داده‌های آموزشی به دست می‌آید. پس از تعیین ماتریس مجاورت گراف به صورت رابطه (۹) و ماتریس لاپلاسین، به صورتی که در بخش ۲-۲ توضیح داده شد خوشه‌ها مشخص می‌شوند.

در مرحله آزمایش با ورود هر داده آزمون به سیستم، ابتدا نمایش تنک آن بر داده‌های آموزشی با استفاده از رابطه (۲) به دست می‌آید. سپس در مرحله بازسازی، در هر گام ضرایب نمایش تنک متناظر با یک خوشه نگه داشته شده؛ مابقی ضرایب نمایش تنک برابر با صفر قرار داده می‌شوند. با یافتن حداقل خطای بازسازی، خوشه‌ای که داده به آن تعلق دارد شناسایی می‌شود.

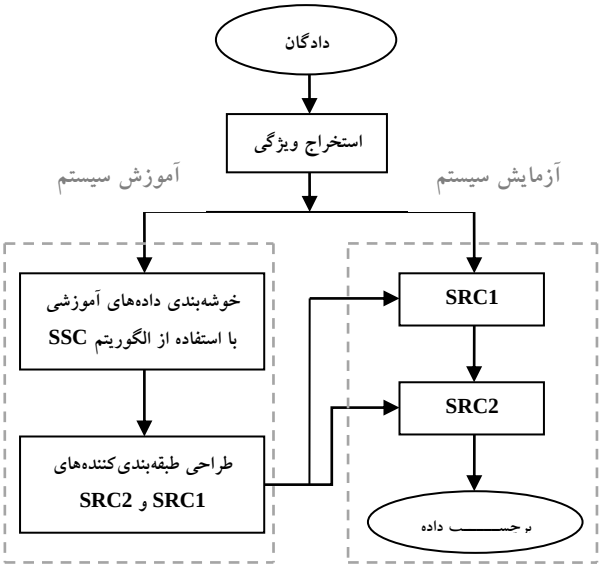
در مرحله بعد برای تعیین برچسب داده مورد نظر، با استفاده مجدد از SRC، نمایش تنک داده تنها بر حسب داده‌های خوشه مربوط به آن که در مرحله قبل تعیین شده است، پیدا می‌شود. بنابراین تصمیم‌گیری نهایی در محدوده‌ای کوچکتر و تنها روی داده‌های یک خوشه انجام می‌گیرد. در شکل (۱) مراحل انجام طبقه‌بندی در روش پیشنهادی به طور خلاصه نشان داده شده است.

۶- نتایج

در این بخش ابتدا پایگاه داده استفاده شده بررسی شده؛ سپس نتایج شبیه‌سازی و تحلیل آن‌ها ارائه می‌شود.

[31] Adjacency Matrix

[32] Hierarchical Classification



شکل (۱)- روندنمای روش پیشنهادی

۷- ۴-۱- دادگان مورد استفاده

برای ارزیابی عملکرد روش طبقه‌بندی معرفی شده از پایگاه داده 14-Tumors استفاده شده است[۱۶]. این مجموعه شامل ۱۹۸ نمونه حاوی ۱۶۰۶۳ ژن (ویژگی) از ۱۴ نوع سرطان است که از این میان ۱۴۴ نمونه برای آموزش سیستم و ۵۴ نمونه برای آزمایش بکار رفته‌اند. مقدار داده شده برای هر ژن در هر نمونه، میزان فعال بودن ژن در آن نمونه یا به عبارتی بیان ژن را نشان می‌دهد. مشخصات کلی این دادگان در جدول (۱) آورده شده است. همان‌گونه که از این جدول نیز ملاحظه می‌شود، تعداد نمونه در هر دسته از این داده‌ها در مقایسه با ابعاد آن‌ها بسیار کم است.

۸- ۴-۲- نتایج شبیه‌سازی

از آنجا که پایگاه داده مورد استفاده در این پژوهش، از جمله پایگاه داده‌های چالش‌برانگیز در امر طبقه‌بندی است، اعمال روش‌های مختلف طبقه‌بندی بر آن اغلب نتایج ضعیفی از خود نشان داده‌اند. تا کنون از تعمیم‌های مختلفی از روش SVM شامل OVR[۳۳]، OVO، DAG[۳۴]، WW[۳۵] و CS[۳۶] برای طبقه‌بندی این داده‌ها استفاده شده است. در تمام این روش‌ها از کرنل‌های چندجمله‌ای و RBF[۳۷] استفاده شده است و انتخاب پارامترهای این طبقه‌بندی‌کننده‌ها با استفاده از روش ارزیابی متقابل[۳۸] به صورت دقیق انجام شده است. با این حال کمترین خطای به دست آمده در این روش‌ها حدود ۲۴ درصد است[۲۰]. در ادامه نتایج حاصل از اعمال روش پیشنهادی به این پایگاه داده ارائه شده است.

جدول (۱)- مشخصات دادگان مورد آزمایش

تعداد نمونه‌های آزمون	تعداد نمونه‌های آموزشی	نوع سرطان
۴	۸	سینه
۶	۸	پروستات
۴	۸	ریه
۴	۸	کولورکتال[۳۹]
۶	۱۶	لنفوم
۳	۸	مثانه
۲	۸	ملانوم[۴۰]
۲	۸	رحم
۶	۲۴	خون
۳	۸	کلیه
۳	۸	لوزالمعده
۴	۸	تخمدان
۳	۸	مزوتلیوما[۴۱]
۴	۱۶	سیستم عصبی مرکزی

همان‌طور که در بخش ۲ نیز ذکر شد، روش‌های متعددی برای محاسبه نمایش تنک ارائه شده است که هر یک مزایا و معایبی دارند. در جدول (۲) نتایج حاصل از استفاده از تعدادی از این روش‌ها برای محاسبه نمایش تنک در الگوریتم SRC آورده شده است. چنانچه ملاحظه می‌شود دو روش OMP و BP در این میان بهترین عملکرد را دارند و کمترین میزان خطا را در طبقه‌بندی این دادگان متحمل می‌شوند؛ اما با توجه به زمان‌بر بودن اجرای الگوریتم BP، در این کاربرد از الگوریتم OMP استفاده کرده‌ایم و نتایجی که در ادامه اعلام می‌شوند با استفاده از این الگوریتم به دست آمده‌اند.

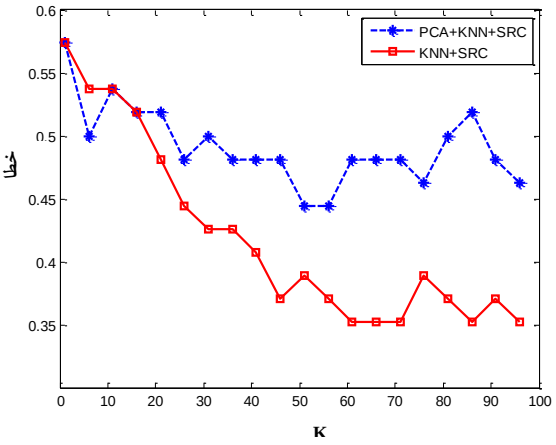
در مرجع [۲۵] با ترکیب دو روش KNN و SRC، طبقه‌بندی انجام شده است. روش کار به این صورت است که ابتدا با اعمال طبقه‌بندی‌کننده KNN، K نزدیک‌ترین همسایه به داده آزمون از میان داده‌های آموزشی تعیین می‌شود. سپس

[33] One-Versus-Rest (All)
[34] Directed Acyclic Graph
[35] all-at-once method by Weston and Watkins
[36] all-at-once method by Crammer and Singer
[37] Radial Basis Function
[38] Fold Cross Validation
[39] Colorectal
[40] Melanoma
[41] Mesothelioma

با استفاده از SRC، نمایش تنک داده آزمون تنها روی این K داده آموزشی محاسبه شده؛ طبقه‌بندی نهایی انجام می‌شود. از آنجا که ایده اصلی این روش مشابه روش پیشنهادی ما است، لذا نتایج حاصل از اعمال این روش بر دادگان مورد بررسی را نیز به منظور مقایسه با نتایج روش معرفی شده بررسی کرده‌ایم. نتایج حاصل از این بررسی در شکل (۳) ارائه شده است.

جدول (۲)- درصد خطای طبقه‌بندی با روش SRC و با استفاده از روش‌های مختلف محاسبه نمایش تنک

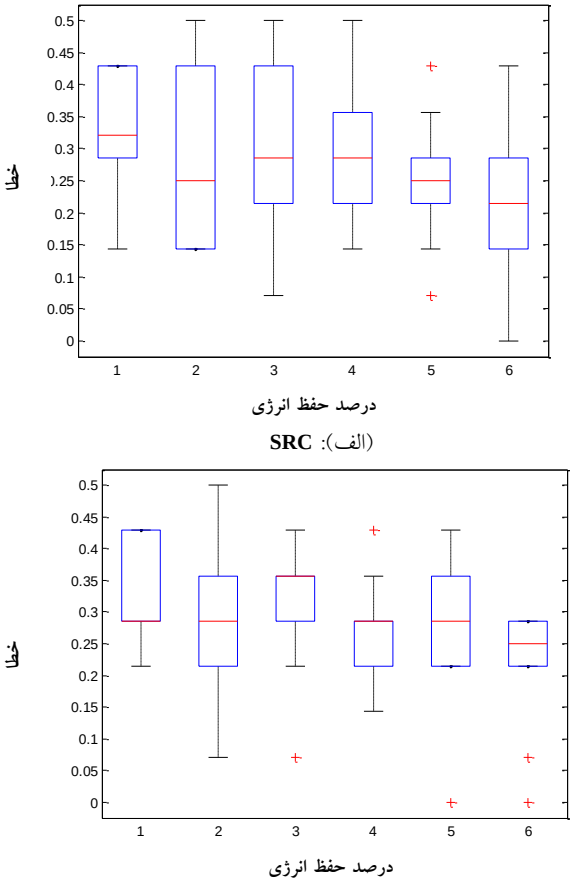
الگوریتم محاسبه پاسخ تنک	**BP**	**OMP**	**Homotopy**	**DALM**	**PALM**
درصد خطای طبقه‌بندی با **SRC**	۲۴/۰۷	۲۴/۰۷	۳۳/۳۳	۳۳/۳۳	۲۵/۹۳



شکل (۳)- خطای طبقه‌بندی در روش KNN+SRC

در این شکل مقدار خطای حاصل از طبقه‌بندی با روش KNN+SRC به ازای مقادیر مختلف K و در دو حالت بدون استخراج ویژگی و استفاده از روش استخراج ویژگی PCA آورده شده است. همان‌طور که در این شکل ملاحظه می‌شود، کمترین مقادیر خطا در حالت استفاده از تمام ویژگی‌ها و به ازای مقادیر بزرگ K بدست آمده که در حدود ۳۵ درصد است. همچنین ملاحظه می‌شود کاهش بعد داده‌ها با استفاده از PCA منجر به افزایش خطای طبقه‌بندی می‌شود.

هم در روش SRC و هم در روش پیشنهادی (SSC+SRC)، نیز به منظور کاهش بعد داده‌ها از روش استخراج ویژگی PCA استفاده شده است. به منظور تعیین بعد بهینه داده‌ها، از روش ارزیابی متقابل ۱۰-قسمتی برای داده‌های آموزشی استفاده شده است. به این ترتیب که ابتدا داده‌های آموزشی به ۱۰ قسمت مساوی تقسیم شده‌اند؛ سپس در هر مرحله یکی از قسمت‌ها به عنوان داده آزمون و ۹ قسمت باقی‌مانده به عنوان داده آموزشی برای سیستم در نظر گرفته شده‌اند. سپس عمل طبقه‌بندی به ازای شش مقدار مختلف بعد داده ورودی که به ترتیب ۹۶، ۹۷، ۹۸، ۹۹، ۹۹/۵ و ۹۹/۹ درصد انرژی داده‌های ورودی را حفظ کنند، در هر مرحله تکرار شده است. نتایج حاصل از این آزمایش‌ها در شکل (۴) نشان داده شده است.



(الف): **SRC**

(ب): روش پیشنهادی

شکل (۴)- میانه و انحراف معیار خطاهای بدست آمده در روش ارزیابی متقابل برای SRC (الف) و روش پیشنهادی (ب)

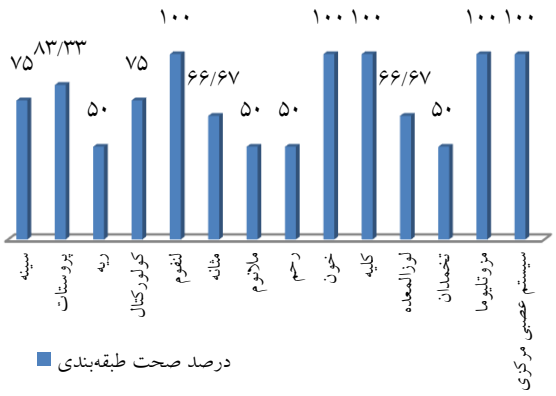
در این شکل‌ها مقدار میانه و چارک‌های اول و سوم برای ده مقدار خطای بدست آمده در هر بعد، نشان داده شده است. همان‌طور که در شکل‌های بالا دیده می‌شود در هر دو روش، کمترین میزان میانه در حالت ششم و به ازای حفظ ۹۹/۹ درصد انرژی بدست آمده است. در بخش (ب) شکل (۴) حداقل تغییرات خطا نیز در این بعد بدست آمده است، ولی در قسمت (الف) این شکل در حالت پنجم کمترین تغییرات خطا رخ داده است. در مقایسه بین دو حالت پنجم و ششم برای روش SRC، از آنجایی که مقدار چارک سوم در این دو حالت برابر و میزان میانه در حالت ششم کمتر است، بعد معادل با حالت ششم را انتخاب کرده‌ایم. بنابراین بعد بهینه بمنظور اعمال الگوریتم PCA در مرحله آزمایش سیستم برای هر دو روش SRC و SSC+SRC، در ازای حفظ ۹۹/۹ درصد انرژی – که برابر با ۱۳۰ است– در نظر گرفته شده است.

نتایج طبقه‌بندی با استفاده از روش پیشنهادی به ازای داده ورودی با بعد ۱۳۰ و تعداد خوشه‌های متفاوت در جدول (۳) خلاصه شده است. هر چند در روش خوشه‌بندی SSC نیازی به دانستن تعداد خوشه‌ها نیست، در عمل به علت وجود نویز ممکن است بین داده‌های خوشه‌های مختلف نیز یال‌هایی وجود داشته باشد که تعیین تعداد خوشه‌ها را با مشکل مواجه کند؛ لذا آزمایش‌ها به ازای مقادیر مختلف تعداد زیرفضاها تکرار شده‌اند و نتایج اعلام شده است. بر اساس نتایج بدست آمده در این بررسی، در ادامه آزمایش‌ها تعداد خوشه‌ها برابر با پنج در نظر گرفته شده است.

نتایج حاصل از روش طبقه‌بندی ارائه شده به تفکیک هر یک از دسته‌ها نیز در شکل (۵) آورده شده است. همان‌طور که ملاحظه می‌شود پنج نوع سرطان لنفوم، خون، کلیه، مزوتلیوما و سیستم عصبی مرکزی به صورت کاملاً صحیح و با صحت ۱۰۰ درصد طبقه‌بندی شده‌اند و کمترین میزان صحت طبقه‌بندی مربوط به سرطان‌های ریه، ملانوم، رحم و تخمدان و به میزان ۵۰ درصد است.

جدول (۳)– خطای طبقه‌بندی روش SRC+SSC به ازای تعداد خوشه‌های مختلف

خطای طبقه‌بندی(٪)	تعداد خوشه‌ها
۲۴/۰۷	۲
۲۲/۲۲	۳
۲۲/۲۲	۴
۲۰/۳۷	۵
۲۰/۳۷	۶
۲۲/۲۲	۷
۲۴/۰۷	۸
۲۲/۲۲	۹
۲۲/۲۲	۱۰
۲۰/۳۷	۱۱
۲۰/۳۷	۱۲
۲۷/۷۸	۱۳



شکل (۵)– صحت طبقه‌بندی روش پیشنهادی به تفکیک دسته‌ها

در جدول (۴) نتایج حاصل از اعمال طبقه‌بندی‌کننده‌های مختلف به پایگاه داده مورد نظر آورده شده‌اند. نتایج تعمیم‌های مختلف روش SVM و همچنین روش SR از مرجع **[۲۰]** آورده شده‌اند. در هر یک از این روش‌ها، بهترین نتایج اعلام شده، گزارش شده است.

با توجه به نتایج ذکر شده می‌توان دید که روش طبقه‌بندی SRC خود به تنهایی در مقایسه با روش‌های پیشین از عملکرد بهتری برخوردار است. همچنین عدم نیاز به انتخاب مدل و یا پارامترهای اولیه در مرحله آموزش را می‌توان از جمله مزایای دیگر این طبقه‌بندی‌کننده برشمرد.

نتایج بدست آمده همچنین حاکی از عملکرد قابل قبول روش پیشنهادی است که منجر به کمترین مقدار خطای طبقه‌بندی در مقایسه با سایر روش‌ها شده است. اگرچه هر دو روش KNN+SRC و SSC+SRC، با رویکردی یکسان نسبت به طبقه‌بندی این دادگان اقدام می‌کنند، اختلاف قابل توجهی در نتایج حاصل از آن‌ها مشاهده می‌شود که نشان‌دهنده قابلیت روش معرفی شده است. همچنین باید به این نکته نیز اشاره کرد که نتیجه روش KNN+SRC با استفاده از تمام ویژگی‌ها و به ازای مقادیر بزرگ K بدست آمده است که بار محاسباتی زیادی را به سیستم تحمیل کرده؛ زمان اجرای الگوریتم نیز طولانی می‌شود.

جدول (۴)- مقایسه نتایج حاصل از طبقه‌بندی‌کننده‌های مختلف

روش طبقه‌بندی	خطای طبقه‌بندی(٪)
SVM (OVR)	۲۴/۷۱
SVM (OVO)	۵۳/۶۱
SVM (DAG)	۵۴/۹۰
SVM (WW)	۳۴/۱۶
SVM (CS)	۲۴/۶۲
SR	۲۵/۹۶
KNN+SRC	۳۵
SRC	۲۴/۰۷
SSC+SRC	۲۰/۳۷

۵- جمع‌بندی و نتیجه‌گیری

در این مقاله روش طبقه‌بندی سلسله مراتبی با ترکیب دو روش خوشه‌بندی زیرفضای تنک و طبقه‌بندی مبتنی بر نمایش تنک ارائه شد. نتایج آزمایش برای داده‌های 14-Tumors حاکی از این است که روش پیشنهادی در مقایسه با روش‌های طبقه‌بندی دیگر به صحت طبقه‌بندی بیشتری دست یافته است. بر اساس نتایج بدست آمده در جدول (۴)، نه تنها استفاده از روش SRC خطای طبقه‌بندی کمتری را در مقایسه با روش‌های پیشین در پی داشته است، بلکه ترکیب آن با روش SSC منجر به بهبود قابل توجهی در صحت طبقه‌بندی داده‌ها شده است.

از آنجایی که یکی از روش‌های متداول و مفید در دسته‌بندی دادگان با ابعاد زیاد و از جمله دادگان مورد بررسی، خوشه‌بندی است، لذا روش معرفی شده - که در مرحله اول خوشه مربوط به داده آزمون را تعیین کرده؛ سپس در این خوشه عمل طبقه‌بندی را انجام می‌دهد- توانسته به عملکرد قابل قبولی دست یابد. بدین ترتیب، در این روش تنها نمونه‌هایی از مجموعه آموزشی وارد فرایند طبقه‌بندی نهایی می‌شوند که بیشترین شباهت را به داده آزمون داشته باشند. بنابراین احتمال این که نمونه‌های نامربوط در نمایش تنک مربوط به داده آزمون نقش مؤثری را ایفا کنند، به شدت کاهش می‌یابد که این امر افزایش صحت طبقه‌بندی‌کننده پیشنهادی را در مقایسه با روش SRC به دنبال داشته است.

اغلب روش‌های طبقه‌بندی اعمال شده به این پایگاه داده، از تمام ۱۶۰۶۳ ویژگی برای طبقه‌بندی استفاده کرده‌اند، اما نتایج حاصل از اعمال دو طبقه‌بندی‌کننده SRC و SSC+SRC عملکرد خوبی را در ابعاد داده بسیار کمتر از بعد اولیه ویژگی‌ها نشان می‌دهند. در این روش‌ها ضمن حفظ تمام ویژگی‌ها، با انتقال آنها به فضایی با بعد کمتر علاوه بر کاهش بار محاسباتی به نتایج بهتری دست یافته‌ایم.

همچنین از آنجایی که روش ارائه شده روشی غیرپارامتریک است و نیازی به انتخاب مدل و تعیین پارامترهای اولیه ندارد، از این منظر نیز عملکرد آن در مقایسه با روش‌های مبتنی بر SVM قابل توجه است.

بدیهی است به منظور دست‌یابی به نتایج بهتر در تفکیک انواع سرطان‌ها، علاوه بر بهبود روش‌های پردازش سیگنال به استفاده بیشتر از اطلاعات بالینی مربوط به ژن‌ها و دانش زیست‌شناسی نیز نیاز است. امید است در آینده‌ای نزدیک با انجام مطالعات گسترده‌تر در این زمینه، به نتایج قابل قبولی بمنظور استفاده در تشخیص‌های بالینی دست یابیم.

مراجع

[1] Donoho D., Compressed sensing; IEEE Trans. Information Theory, 2006; 52(4): 1289–1306.
[2] Donoho D., For Most Large Underdetermined Systems of Linear Equations the Minimal l_1 -Norm

Solution Is Also the Sparsest Solution; Comm. Pure and Applied Math., 2006; 59(6): 797-829.

[3] Elhamifar E., Vidal R., Sparse subspace clustering: Algorithm, theory, and applications; to appear in IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012; arXiv preprint arXiv: 1203.1005.

[4] Elhamifar E., Vidal R., Sparse subspace clustering; Proc. Int. Conf. on Computer Vision and Pattern Recognition, 2009; 2790–2797.

[5] Liu G., Lin Z., Yu Y., Robust subspace segmentation by low-rank representation; Proc. Int. Conf. on International Conference on Machine Learning, 2010.

[6] Ho J., Yang M.H., Lim J., Lee K.C., Kriegman D., Clustering appearances of objects under varying illumination conditions; Proc. Int. Conf. on Computer Vision and Pattern Recognition, 2003.

[7] Vidal R., Ma Y., Sastry S., Generalized principal component analysis (gpca); IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005; 27(12): 1945–1959.

[8] Tipping M.E., Bishop C.M., Mixtures of probabilistic principal component analyzers; Neural Computation, 1999; 11(2): 443–482.

[9] Fischler M.A., Bolles R.C., Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography; Communications of the ACM, 1981; 24(6): 381–395.

[10] Von Luxburg U., A tutorial on spectral clustering; Statistics and Computing, 2007; 17.

[11] Saha B., Pham D., Phung D., Venkatesh S., Sparse Subspace Clustering via Group Sparse Coding; 2013.

[12] Wright J., Yang A.Y., Ganesh A., Sastry S.S., Ma Y., Robust face recognition via sparse representation; IEEE Trans. on Pattern Analysis and Machine Intelligence, 2009; 31(2): 210–227.

[13] Sahu S., Panda G., Barik R., A Hybrid Method of Feature Extraction for Tumor Classification Using Microarray Gene Expression Data; International Journal of Computer Science & Informatics, 2011;1(1).

[14] Sharma A., Paliwal K., Cancer classification by gradient LDA technique using microarray gene expression data; Data & Knowledge Engineering, 2008; 66: 338–347.

[15] Guyon I., Weston J., Barnhill S., Gene Selection for Cancer Classification using Support Vector Machines; Machine Learning, 2002; 46: 389–422.

[16] Ramaswamy S., Tamayo P., Rifkin R., Multiclass cancer diagnosis using tumor gene expression signatures; Proc. National Academy of Sciences of the United States of America, 2001; 98(26): 15149–15154.

[17] Linder R., Dew D., Sudhoff H., Theegarten D., Remberger K., Poppel S. J., Wagner M., The subsequent artificial neural network (SANN) approach might bring more classificatory power to ANN based DNA microarray analyses; Bioinformatics, 2004; 20(18): 3544–3552.

[18] Zhang R., Huang G. B., Sundararajan N., Saratchandran P., Multi-category classification using an extreme learning machine for microarray gene expression cancer diagnosis; IEEE/ACM Trans. on Computational Biology and Bioinformatics, 2007;4(3): 485– 495.

[19] Shabgahi A. Z., Abadeh M. S., A fuzzy classification system based on memetic algorithm for cancer disease diagnosis; in Proc. 18th IEEE Iranian Conference of Biomedical Engineering (ICBME), 2011.

[20] Hang X. Wu F., Sparse Representation for Classification of Tumors Using Gene Expression Data; Journal of Biomedicine and Biotechnology, 2009.

[21] Chen S., Donoho D., Saunders M. A., Atomic decomposition by basis pursuit; SIAM journal on scientific computing, 1998; 20: 33-61.

[22] Mallat S. G., Zhifeng Z., Matching pursuit with time-frequency dictionaries; IEEE Trans. On Signal Processing, 1993; 41: 3397-3415.

[23] Pati Y. C., Rezaiifar R., Krishnaprasad P. S., orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition; 1993; rec. 27 Asilomar cnf. Signals, syst. Comput, 41-44.

[24] Yang A. Y., Zhou Z., Ganesh A., Sastry, S. S., Ma Y., Fast l1 -Minimization Algorithms For Robust Face Recognition; IEEE Transactions on Image Processing, 2013; 99:1-1- 0.

[25] Nan Z., Jian Y., K nearest neighbor based local sparse represen-tation classifier; CCPR, IEEE, 2010; 1–5.