دوره دوازدهم، شماره اول، بهار وتابستان ۱۳۹۸

مجله فناوری اطلاعات در طراحی مهندسی

Information Technology in Engineering Design

http://ited.sinaweb.net

خوشه‌بندی مبتنی بر ماشین بردار پشتیبان دوقلو به‌منظور انتخاب ویژگی در مسئله دسته‌بندی داده‌های ریزآرایه

نفیسه سلیمانی * (۱) **محمدحسین معطر** (۱)

(۱) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران.*

(۲) گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران.

(تاریخ دریافت : ۱۳۹۷/۱/۲۷ تاریخ پذیرش : ۱۳۹۸/۹/۱۹)

چکیده

طبقه‌بندی سرطان، به‌عنوان مسئله ای مهم در تشخیص و درمان سرطان به شمار می‌رود. یکی از مؤثرترین روش‌ها در طبقه‌بندی سرطان، شناسایی ژن‌هایی مرتبط و تبعیض‌آمیز برای طبقه‌بندی نمونه‌ها در آنالیز بیانی ژن می‌باشد. درروش پیشنهادی در این مقاله، با خوشه‌بندی ویژگی‌ها و اعمال انتخاب ویژگی درون خوشه‌ها، انتظار می‌رود که متمایزکننده‌ترین و مهم‌ترین ویژگی‌ها استخراج شوند. درروش پیشنهادی، به‌منظور کاهش ابعاد مجموعه داده، فن انتخاب ویژگی مبتنی بر اهمیت ویژگی‌ها به کار گرفته می‌شود، ویژگی‌های رتبه بالا استخراج‌شده و جهت خوشه‌بندی به ماشین بردار پشتیبان دوقلو برای خوشه‌بندی ارائه می‌شوند. پس از خوشه‌بندی، با به کار گرفتن فن انتخاب ویژگی مبتنی بر همبستگی، قابل‌اعتمادترین ویژگی‌ها انتخاب‌شده و توسط طبقه بند پرسپترون چندلایه، طبقه‌بندی می‌شوند. جهت ارزیابی روش پیشنهادی، از چهار مجموعه داده‌ی SRBCT، Leukemia، DLBCL و Prostate استفاده‌شده است. نتایج آزمایش‌ها بیانگر بهبود عملکرد دقت طبقه‌بندی می‌باشد.

واژه‌های کلیدی : انتخاب ویژگی، خوشه‌بندی، ماشین بردار پشتیبان دوقلو جهت خوشه‌بندی (TWSVC)، طبقه‌بندی، پرسپترون چندلایه (MLP)

* عهده‌دار مکاتبات:

نشانی: گروه مهندسی کامپیوتر، واحد مشهد، دانشگاه آزاد اسلامی، مشهد، ایران.

تلفن: ۰۹۱۵۵۰۳۰۵۴۰ پست الکترونیکی: *moattar@mshdiau.ac.ir*

۱. مقدمه

بهبود مداوم فنّاوری بیانی ژن، توانایی اندازه‌گیری سطوح بیانی هزاران ژن را به‌طور موازی فراهم می‌کند[۱]. یکی از اهداف مهم تحقیقات ریزآرایه‌ی DNA توسعه‌ی ابزارهایی برای تشخیص سرطان با دقّت بیشتر و مبتنی بر پروفایل ژنتیکی یک تومور می‌باشد. پیش‌بینی دقیق انواع مختلف تومور درمانی بهتر و کاهش سمّی بودن را برای بیماران ممکن می‌سازد. انتخاب ژن موفق، به طبقه‌بندی انواع مختلف سرطان کمک کرده و منجر به درک بهتری از مشخصات ژنتیکی سرطان‌ها و استراتژی‌های درمانی می‌شود [۲].

روش‌های طبقه‌بندی مولکولی مبتنی بر الگوریتم‌های یادگیری ماشین درداده مای ریزآرایه‌ی DNA بکار گرفته‌شده است تا نمایش‌دهنده‌ی ارتباط کلینیکی و آماری انواع مختلف تومورها باشند [۳]. یک مجموعه داده‌ی ریزآرایه‌ی معمولی بسیار پراکنده است، پراکندگی شدید و حجم کم نمونه، مانند گلوگاهی جهت طبقه‌بندی دقیق و قدرتمند می‌باشد [۲]. اساساً انتخاب ویژگی باهدف انتخاب اطلاعات مرتبط و داده مای حاوی اطلاعات مفید مورداستفاده قرار می‌گیرد. با کاهش داده‌ها و هرس داده مای زائد و بی‌ربط، انتخاب ویژگی می‌تواند عملکرد و سرعت الگوریتم یادگیری ماشین را افزایش داده، هم‌چنین کاهش نیاز جهت ذخیره‌ی اطلاعات، ارتقاء فهم بهتر داده‌ها و ساده‌سازی تجسّم داده‌ها را در پی دارد [٤]. شناسایی یک زیرمجموعه‌ی کوچک از ژن‌ها برپایه‌ی تشخیص، نه‌تنها دقت طبقه‌بندی را بهبود می-بخشد بلکه دید قابل‌توجهی درباره‌ی طبیعت بیماری و مکانیزم مای ژنتیکی مسئول در مقابل بیماری می‌یابد [۱].

وظیفه‌ی طبقه‌بندی سرطان با استفاده از داده مای ریزآرایه، طبقه‌بندی بافت مای نمونه در کلاس مای مرتبط از فنوتایپ ها، به‌عنوان‌مثال سرطان در برابر نرمال می‌باشد [۵]. مهم‌ترین چالش‌ها درزمینه ی طبقه‌بندی سرطان عبارت‌اند از:

۱. با فرض اینکه M حالات ژن و N نمونه‌های بافتی باشد: M>>N. M محدوده‌ای۲۰۰۰-۲۰۰۰۰ داشته درحالی‌که N محدوده‌ای ۳۰-۲۰۰ دارد.

۲. اغلب ژن‌ها برای طبقه‌بندی انواع مختلف بافت‌ها مرتبط نیستند.

۳. داده‌ها طبیعتی نویزی دارند.

جهت غلبه بر این چالش‌ها، رویکرد معمولاً انتخاب ژن جهت انتخاب زیرمجموعه‌ی کوچکی از ژن مای آگاهی‌بخش که موجب به حداکثر رساندن دقت طبقه‌بندی توسط طبقه-بندها می‌شود، مورداستفاده قرار می‌گیرد. این رویکرد مزایایی هم چون:

۱. شناسایی ویژگی‌های تفسیری بیشتر

۲. کاهش ابعاد داده به‌منظور کاهش هزینه‌ی محاسباتی

۳. کاهش نویز به‌منظور افزایش عملکرد طبقه‌بندی را دارد [۵].

Jain و همکارانش در [٦] مدل هیبریدی دومرحله‌ای برای طبقه‌بندی سرطان ارائه دادند که ادغامی از انتخاب ویژگی مبتنی بر همبستگی (CFS) و بهینه‌سازی ازدحام ذرات بهبودیافته‌ی باینری (iBPSO) است. مسئله‌ی همگرایی به بهینه‌ی محلی در BPSO سنتی، توسط iBPSO کنترل می‌شود. Mahajan و همکارانش در [۷] جهت انتخاب ویژگی و کاهش ابعاد، روشی به نام فرآیند تحلیل سلسله مراتبی تطبیقی (A2HP) ارائه دادند. فرآیند تحلیل سلسله مراتبی یک روش تحلیل تصمیم مبتنی بر معیارهای چندگانه است که نتایج آن به دانش متخصص یا تصمیم گیرندان وابسته است. ابتدا، پیش‌پردازش مجموعه بیانی ژن انجام می‌شود و سپس ژن مای کاهش‌یافته‌ی به‌دست‌آمده، به‌عنوان ورودی برای

A2HP استفاده می‌شوند. A2HP از برگ خریدهای کمّی و کیفی برای انتخاب ژن مای حاوی اطلاعات مفید استفاده می‌کند. Aziz و همکارانش در [۸] روشی ارائه داده‌اند که ترکیبی از رویکرد انتخاب/استخراج ویژگی برای طبقه‌بندی شبکه مای عصبی مصنوعی (ANNs) درداده مای ریزآرایه با ابعاد بالا است که از آنالیز مؤلفه مای مستقل (ICA) به‌عنوان فن استخراج و کلونی زنبورعسل مصنوعی (ABC) به‌عنوان فن بهینه-سازی بهینه‌سازی استفاده می‌کند. Zhang و همکارانش در [۹] از ماشین بردار پشتیبان ۱-norm به همراه مجذور اتلاف (۱-norm SVMSL) جهت اجرا سریع انتخاب ویژگی برای طبقه-بندی سرطان استفاده می‌کنند. ۱- norm SVMSL، نوعی ماشین بردار پشتیبان ۱-norm است که در این مقاله ارائه‌شده که توانایی انتخاب ژن و طبقه‌بندی به‌طور هم‌زمان دارد. جهت بهبود تفسیرپذیری ژن مای منتخب برای دقت پیش‌بینی، روش انتخاب ژن بهبودیافته مبتنی بر بهینه‌سازی ازدحام ذرات باینری BPSO و اطلاعات پیشین، توسط Han و همکارانش در [۱۰] ارائه‌شده است. به‌منظور انتخاب ژن، BPSO اطلاعات حساسیت ژن به کلاس GCS را رمزگذاری می‌کند. اطلاعات حساسیت ژن به کلاس، که توسط ماشین یادگیری مفرط ELM از نمونه‌ها استخراج‌شده‌اند، در فرآیند انتخاب، طی چهار مرحله کدگذاری می‌شوند: آماده‌سازی ذرات، به‌روزرسانی ذرات، اصلاح حداکثر سرعت و اتخاذ عملیات جهش انطباقی.

در این مقاله برای ویژگی‌های مناسب و بااهمیت از خوشه‌بندی ویژگی‌ها به کمک ماشین بردار پشتیبان دوقلو استفاده گردید. جهت انتخاب آموزنده‌ترین ویژگی‌های هر خوشه فن انتخاب ویژگی مبتنی بر همبستگی استفاده شد. آنگاه با کنار هم قرار دادن ویژگی‌های مستخرج از خوشه‌های متعدد، مجموعه داده‌ی موردنظر نهایی ایجاد گردید. جهت طبقه‌بندی و به دست آوردن دقت حاصل از طبقه‌بندی برای ویژگی‌های مستخرج و منتخب از خوشه‌های متعدد طبقه بند پرسپترون چندلایه استفاده شد.

مهم‌ترین نوآوری گروه مقاله عبارت است از خوشه‌بندی ویژگی‌های مرتبط با بیشترین مشابهت با یکدیگر و یافتن آموزنده‌ترین ویژگی‌ها در میان ویژگی‌هایی با خصیصه‌هایی مشابه از هر خوشه. نتایج حاصل از آزمایش‌ها بیانگر این امر است که انجام مراحل مذکور نه‌تنها منجر به کاهش چشمگیر ویژگی‌ها و پالایش مجموعه داده‌ها می‌شود، بلکه انتخاب آموزنده‌ترین ویژگی‌ها از میان ویژگی‌هایی با خصایص مشابه باعث بهبود دقت طبقه بند می‌شود.

ساختار مقاله به شرح زیر است: در بخش دوم به توصیف انتخاب ویژگی، انتخاب ویژگی مبتنی بر اهمیت ویژگی‌ها و انتخاب ویژگی مبتنی بر همبستگی؛ در بخش سوم، خوشه-بندی و ماشین بردار پشتیبانی دوقلو برای خوشه‌بندی، و در بخش چهارم طبقه بند پرسپترون چندلایه که درروش پیشنهادی مورداستفاده قرار می‌گیرند، می‌پردازیم. در بخش پنجم، به بیان روش پیشنهادی پرداخته خواهد شد؛ ارزیابی روش پیشنهادی در بخش ششم و نتایج در بخش هفتم ارائه می‌شوند.

۲. انتخاب ویژگی

انتخاب ویژگی، زیرمجموعه‌ای از ویژگی‌های مرتبط را انتخاب می-کند و ویژگی‌های زائد و بی‌ارتباط را از داده‌ها به‌منظور ساخت مدل مای یادگیری قدرتمند حذف می‌کند [۱۱]. مهم‌ترین هدف انتخاب ویژگی، کاهش ابعاد با حذف زوائد و انتخاب ویژگی‌های مرتبط می‌باشد. این امر موجب افزایش دقّت یادگیری می‌شود که قابل‌فهم بودن نتایج را افزایش داده و زمان یادگیری را کاهش می‌دهد [۱۲]. در مفهوم طبقه‌بندی، فن مای انتخاب ویژگی در سه دسته با توجه به چگونگی ترکیب

جستجو انتخاب ویژگی با ساخت مدل طبقه‌بندی سازمان‌دهی می‌شوند: روش‌های پالایه، روش‌های پوشش‌دهنده و روش‌های تعبیه‌شده [۱۳].

۲-۱. انتخاب ویژگی مبتنی بر اهمیت ویژگی‌ها

در فن انتخاب ویژگی مبتنی بر اهمیت ویژگی‌ها به هر ویژگی، مقادیر اهمیت اختصاص داده می‌شود [۱٤]. بر این -اساس که اگر یک ویژگی مهم باشد، آنگاه احتمال قوی وجود دارد که عناصری با مقادیر مجموعه مای مکمل برای این ویژگی، متعلق به مجموعه‌ی مکمل از کلاس‌ها باشند. با توجه به این‌که تصمیمات کلاس برای دو مجموعه از عناصر متفاوت است، انتظار می‌رود که مقادیر اهمیت ویژگی برای این دو مجموعه از عناصر متفاوت باشد. اهمیت یک ویژگی به‌عنوان یک عملکرد دوطرفه از ارتباط با تصمیم کلاس محاسبه می‌شود. یک ویژگی در صورتی مهم است که ارتباط ویژگی باکلاس و ارتباط کلاس با ویژگی برای ویژگی بالا باشد.

۲-۲. انتخاب ویژگی مبتنی بر همبستگی

عوامل زیادی جهت موفقیت وظیفه‌ای محوّل شده در یادگیری ماشین تأثیرگذار هستند. اولین و مهم‌ترین عامل، بازنمایی و کیفیت نمونه داده‌ها می‌باشد [۱۵].

در قلب الگوریتم CFS، مکاشفه‌ای برای ارزیابی ارزش یا شایستگی یک زیرمجموعه از ویژگی‌ها وجود دارد. این مکاشفه میزان مفید بودن ویژگی‌های فردی برای پیش‌بینی برچسب کلاس به همراه سطح همبستگی میان آن‌ها را در نظر می‌گیرد. فرضیه‌ی مکاشفه به این صورت است که:

زیرمجموعه ویژگی‌های خوب شامل ویژگی‌هایی است که همبستگی بالایی باکلاس دارند درحالی‌که به همدیگر وابسته نیستند. در نظریه‌ی آزمون، همین اصل برای طراحی یک آزمون مرکب (مجموع یا میانگین آزمون مای انفرادی) برای پیش‌بینی یک متغیر خارجی موردعلاقه استفاده‌شده است. در این شرایط، ویژگی‌ها آزمون مای انفرادی هستند که ویژگی‌های مرتبط با متغیر موردعلاقه (کلاس) را اندازه‌گیری می‌کنند. فرمول معادله‌ی مکاشفه، به‌صورت زیر می‌باشد:

$$Merit_s = \frac{k\,\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad (۱)$$

Merits مکاشفه‌ی "شایستگی" زیرمجموعه ویژگی S که حاوی k ویژگی است، میانگین همبستگی ویژگی کلاس و میانگین همبستگی ویژگی است. صورت کسر می‌تواند به‌عنوان نشانی داده‌شده از چگونگی پیش‌بینی یک گروه از ویژگی‌ها در نظر گرفته شوند؛ مخرج کسر میزان افزونگی که در میان آن‌ها وجود دارد را نشان می-دهد. مکاشفه ویژگی‌های بی‌ارتباط را به‌عنوان این‌که در پیش‌گویی مای کلاس ضعیف عمل می‌کنند، دسته‌بندی می کند. ویژگی‌های اضافی در برابر آن‌هایی که با یک یا بیش از یک ویژگی همبستگی بالایی دارند، متمایز می‌شوند. به دلیل این‌که ویژگی‌ها به‌طور مستقل رفتار می‌کنند، CFS نمی‌تواند باقدرت اثرات متقابل ویژگی‌ها را شناسایی کند. بااین‌حال می‌تواند ویژگی‌های مفید را تحت سطوح معتدلی از تعاملات شناسایی کند [۱٦].

۳. خوشه‌بندی

خوشه‌بندی به معنای انجام تقسیم‌بندی یک مجموعه داده‌ی بدون برچسب به گروه‌هایی با اشیاء مشابه می‌باشد [۱۷]. خوشه‌بندی به-عنوان یک فن پردازش داده در حوزه مای مختلف بسیاری، همانند هوش مصنوعی، بیوانفورماتیک، بیولوژی، بینایی کامپیوتر، برنامه‌ریزی شهری، داده‌کاوی، فشرده‌سازی داده‌ها، تجزیه‌وتحلیل تصاویر و ... استفاده می‌شود. هر خوشه باید دو ویژگی مهم را دارا باشد؛ تشابه کم مابین

کلاسی و تشابه زیاد درون کلاسی. خوشه‌بندی یک یادگیری بدون نظارت می‌باشد. هیچ برچسب کلاس از پیش تعریف‌شده‌ای برای نقاط داده وجود ندارد. خوشه‌بندی موجب دستیابی به توزیع کلی الگوها و همبستگی میان اشیاء داده‌ها می‌شود [۱۸].

۳-۱. ماشین بردار پشتیبانی دوقلو برای خوشه‌بندی

روش خوشه‌بندی جدیدی مبتنی بر صفحه و بر پایه‌ی ماشین بردار پشتیبانی دوقلو (TWSVM) به نام بردار پشتیبانی دوقلو برای خوشه‌بندی (TWSVC) ارائه‌شده است. TWSVM نقطه‌ی عطفی در طبقه‌بندی مبتنی بر صفحه می‌باشد.

مزایای این روش عبارت‌اند از:

۱. با استفاده از کالبد TWSVM، TWSVC از اطلاعات درون و مابین خوشه‌ای بهره‌برداری می‌کند.

۲. متفاوت از TWSVM که یک صفحه‌ی کلاس برای حفظ نمونه‌های کلاس مای مختلف که به‌دوراز فقط یک صفحه هستند، موردنیاز است؛ در TWSVC، نیازمندی از فقط یک بخش به‌طور معقول‌تری از دو بخش ابر صفحه‌ی مرکز خوشه جایگزین شده است [۱۹].

۳-۱-۱. الگوریتم ماشین بردار پشتیبانی دوقلو برای خوشه‌بندی

برای مسئله‌ی خوشه‌بندی، TWSVC ارائه‌شده، k صفحه‌ی مرکز خوشه را جستجو می‌کند:

$$Center - plane_i = w_i^T x + b_i = 0, \quad i = 1,...,k \qquad (۲)$$

k کلاس نمونه‌های X1,X2,…,Xk به‌عنوان مجموعه‌ی آموزش X می‌باشد. با در نظر گرفتن مسئله‌ی زیر با :

که c>0 یک پارامتر پنالتی و یک بردار ضعیف

$$\min_{w_i, b_i, x_i, X_i} \quad \frac{1}{2} \| X_i w_i + b_i e \|^2 + c e^T x_i \qquad (۳)$$
$$s.t. \ | \hat{X}_i w_i + b_i e |^3 \ e - x_i, \ x_i \ ^3 \ 0$$

می‌باشد. به‌طور دقیق‌تر، مرکز صفحه‌ی i امین خوشه در TWSVC تا حد امکان به iامین خوشه‌ی Xi نزدیک و از دیگر خوشه‌های از هر دو سمت با i=1,..,k دور باشد.

با برچسب مای اولیه‌ی خوشه‌ی X، TWSVC مراکز صفحات تمام خوشه‌ها را به‌روزرسانی می‌کند و برچسب مای نمونه‌ی جایگزین تا رسیدن به شرایط راضی‌کننده به-روزرسانی می‌شوند [۱۹].

٤. طبقه‌بندی

طبقه‌بندی یکی از وظایف مهم شناسایی الگو می‌باشد [۲۰]. پرسپترون چندلایه (MLP) نوعی از شبکه‌ی عصبی است که داده مای ورودی را به داده مای هدف مورد انتظار نگاشت می‌کند. MLP متشکل از چندین لایه گره در یک گراف جهت‌دار است که هر لایه به‌طور کامل به لایه‌ی بعدی متصل است. به‌طورکلی (شکل ۱)، فرض می‌شود که [(x(n),t(n)] معرّف n امین نمونه آموزشی است و x(n)=[x1(n), x2(n),…, xd(n)]T که (n=1,…,N) بردار ورودی با ابعاد d و t(n)=[t1(n), t2(n),…, tc(n)]T هدف با ابعاد c را نشان می-دهند. آموزش MLP، مسئله‌ی بهینه‌سازی به حداقل رساندن مجموع میانگین مربعات خطا (E (MLP بین هدف tk(n) و خروجی واقعی yk(n) می‌باشد.

$$E = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{c} (y_k(n) - t_k(n))^2 \qquad (٤)$$

با فرض این‌که g تابع فعال‌سازی در لایه پنهان، k ابعاد هدف، h تابع فعال‌سازی در لایه ورودی، A وزن مای

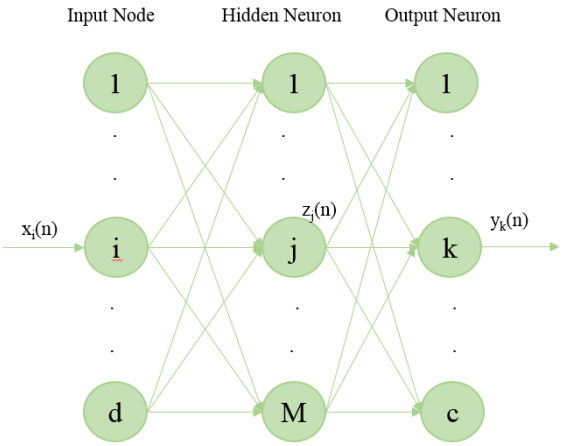
اتصال لایه مای ورودی و پنهان و B وزن مای اتصال لایه مای پنهان و خروجی باشد، آنگاه داریم:

$$y_k(n) = h\left(\sum_{j=0}^{M} B_{kj} z_j(n)\right) \quad (۵)$$

که zj(n) خروجی j امین نرون در لایه پنهان را نشان می‌دهد و j=1,2,…,M ، با تعریف

$$z_j(n) = g\left(\sum_{i=0}^{d} A_{ji} x_i(n)\right) \quad (۶)$$

MLP از دو نقص رنج می‌برد: ۱) تعیین تعداد نرون مای بهینه-ی پنهان دشوار است و ۲) ممکن است وزن آموزشی در نقاط بهینه‌ی محلی به دام بیفتد[۲۱].



شکل ۱. ساختار MLP با یک‌لایه پنهان [۲۱]

۵. روش پیشنهادی

گام مای روش پیشنهادی به شرح ذیل می‌باشد:

۱- انتخاب ویژگی به‌منظور کاهش حجم ویژگی‌ها: به‌منظور کاهش حجم ویژگی‌ها، در اولین مرحله فن انتخاب ویژگی مبتنی بر اهمیت ویژگی‌ها در مجموعه داده‌ها اعمال می‌شود. تعداد بسیار زیادی از ویژگی‌های هر مجموعه داده به علت کسب رتبه پایین و مؤثر نبودن حذف می‌شوند. در پایان این مرحله، تعداد ویژگی‌های مهم هر مجموعه داده تا حد ممکن حفظ‌شده و حجم هر مجموعه داده به‌طور چشم‌گیری کاهش می‌یابد.

۲- خوشه‌بندی ویژگی‌ها: درروش پیشنهادی با انجام عملیات خوشه‌بندی، ویژگی‌ها در گروه مای سازمان-یافته قرار می‌گیرند. ویژگی‌هایی درون خوشه‌های واحد قرار می‌گیرند که دارای حداکثر شباهت با یکدیگر و حداقل شباهت با دیگر خوشه‌ها هستند. جهت خوشه‌بندی مهم‌ترین ویژگی‌های استخراج‌شده از مرحله‌ی اول، از الگوریتم ماشین بردار پشتیبانی دوقلو برای خوشه‌بندی ((TWSVC استفاده می‌شود.

۳- انتخاب ویژگی درون خوشه‌ها، به‌منظور استخراج مهم‌ترین ویژگی‌ها: خوشه‌های متعدد ایجادشده با برخورداری از خصایص مشابه درون هر خوشه دارای ارتباط ساختاری با یکدیگر هستند، ازاین‌رو با به‌کارگیری فن انتخاب ویژگی مبتنی بر همبستگی درون خوشه‌ها، می‌توان ویژگی‌های آموزنده که بیشترین همبستگی رادارند، را از میان ویژگی‌هایی با خصیصه مای مشابه انتخاب نمود؛ قابل‌ذکر است که پس از به‌کارگیری فن انتخاب ویژگی مبتنی بر همبستگی درون هر خوشه، ویژگی-هایی که مقدار همبستگی صفر دارند، به‌عنوان ویژگی زائد در نظر گرفته‌شده و حذف می‌شوند. سپس می-توان تضمین نمود که ویژگی‌های منتخب و مستخرج، آموزنده‌ترین و قابل‌اعتمادترین ویژگی‌ها درون هر خوشه هستند، با کنار هم قرار گرفتن ویژگی‌های مستخرج، مجموعه داده‌ی قابل‌اعتماد نهایی تشکیل می‌شود. قابل‌ذکر است که با تغییر تعداد خوشه‌ها و افزایش آن، زمان لازم جهت خوشه‌بندی افزایش می-یابد.

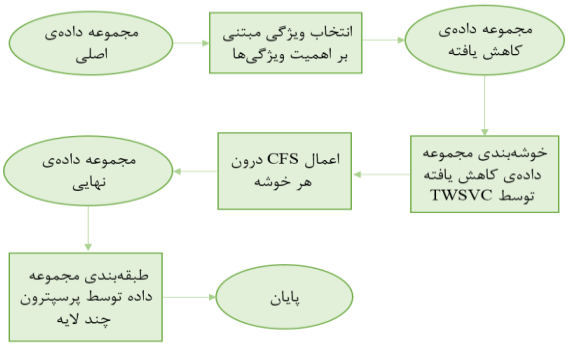
۴- طبقه‌بندی: مرحله خوشه‌بندی داده‌ها و انتخاب زیرمجموعه مناسب از ویژگی‌هایی گام میانی است و اهمیت این گام و صحت آن در گام طبقه‌بندی تعیین

می‌شود. به‌منظور طبقه‌بندی مجموعه داده‌ی پالایش‌شده و قابل‌اعتماد نهایی و به دست آوردن دقت و صحت الگوریتم از طبقه بند پرسپترون چندلایه استفاده می‌شود.

٦. ارزیابی روش پیشنهادی

٦-١- مجموعه داده مای ریزآرایه

جهت بررسی روش پیشنهادی، چهار مجموعه داده از وب‌سایت [٢٢] انتخاب‌شده است. خلاصه‌ای از مجموعه داده‌ها در جدول (١) ارائه‌شده است.



شکل ٢. فلوچارت روش پیشنهادی

ویژگی‌ها	نمونه‌ها	کلاس‌ها	مجموعه داده
٢٣٠٨	٨٣	٤	SRBCT
٥١٤٧	٧٢	٢	Leukemia
٧٠٧٠	٧٧	٢	DLBCL
١٢٥٣٢	١٠٢	٢	Prostate

جدول ١. مشخصات مجموعه داده مای مورداستفاده

ویژگی‌ها نمونه‌ها کلاس هامجموعه داده

نتایج به‌دست‌آمده پس از اجرای مراحل روش پیشنهادی، برای هر مجموعه داده در جدول مای ٥، ٤، ٣، ٢ ارائه‌شده است.

* دومین مرحله درروش پیشنهادی، خوشه‌بندی ویژگی‌ها می‌باشد؛ در این گام، تعداد ویژگی‌ها و نمونه‌ها تغییر نمی‌کند.

جدول ٢. نتایج به‌دست‌آمده برای مجموعه داده SRBCT

ویژگی‌ها	نمونه‌ها	کلاس‌ها	
٢٣٠٨	٨٣	٤	مجموعه داده‌ی اولیه
٦٦٩	٨٣	٤	مرحله‌ی اول
١٦٨	٨٣	٤	مرحله‌ی سوم

جدول ٣. نتایج به‌دست‌آمده برای مجموعه داده Leukemia

ویژگی‌ها	نمونه‌ها	کلاس‌ها	
٥١٤٧	٧٢	٢	مجموعه داده‌ی اولیه
٩١٩	٧٢	٢	مرحله‌ی اول
٢٥٢	٧٢	٢	مرحله‌ی سوم

جدول ٤ نتایج به‌دست‌آمده برای مجموعه داده DLBCL

ویژگی‌ها	نمونه‌ها	کلاس‌ها	
٧٠٧٠	٧٧	٢	مجموعه داده‌ی اولیه
٩٠٠	٧٧	٢	مرحله‌ی اول
١٧٢	٧٧	٢	مرحله‌ی سوم

جدول٥. نتایج به‌دست‌آمده برای مجموعه داده Prostate

ویژگی‌ها	نمونه‌ها	کلاس‌ها	
١٢٥٣٢	١٠٢	٤	مجموعه داده‌ی اولیه
٢٢٢١	١٠٢	٤	مرحله‌ی اول
٢١٢	١٠٢	٤	مرحله‌ی سوم

با توجه به نتایج ثبت‌شده در جدول مای ٥، ٤، ٣، ٢ کاهش ابعاد هر مجموعه داده ، کاملاً مشهود می‌باشد.

٦-٢- معیار ارزیابی عملکرد

در این مقاله، هدف ارزیابی تأثیر انتخاب ویژگی خوشه-بندی ویژگی‌ها بر دقت طبقه‌بندی است، ازاین‌رو دقت طبقه-بندی به‌عنوان مقیاس ارزیابی استفاده می‌شود.

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (٧)$$

در رابطه‌ی (٧) TP تعداد نمونه‌های مثبت صحیح، FP تعداد نمونه‌های مثبت کاذب، TN تعداد نمونه‌های منفی صحیح و FN تعداد نمونه‌های منفی کاذب است.

٦-٣- نتایج

به‌منظور ارزیابی روش پیشنهادی، دقت طبقه بند طی مراحل متفاوتی از اجرای روش پیشنهادی به‌دست‌آمده

است، نتایج در جدول ٦ ارائه‌شده‌اند.

ستون اول: اعمال مرحله‌ی اول از روش پیشنهادی – (Significiance+MultiLayerPerceptron)

ستون دوم : اعمال مرحله‌ی اول و سوم از روش پیشنهادی

(Significiance + CFS+ MultiLayerPerceptron) –

ستون سوم : روش پیشنهادی (به‌اختصار TWCMP)– (Significiance+TWSVC+CFS+ MultiLayerPerceptron)

با توجه به جدول ٦، می‌توان مشاهده کرد که افزودن هر یک از گام مای روش پیشنهادی، موجب بهبود دقت طبقه بندی می‌شود. دقت به‌دست‌آمده برای مجموعه داده مای SRBCT و Leukemia بدون تغییر باقی می‌ماند، امّا روش پیشنهادی در مجموعه داده‌ی DLBCL، به نتیجه‌ی ۱۰۰٪ دست‌یافته است

جدول ٦: دقت طبقه‌بندی پیش و پس از اجرای روش پیشنهادی

Dataset	Sig+MLP	Sig+CFS+MLP	Proposed method
SRBCT	۱۰۰	۱۰۰	۱۰۰
Leukemia	۹۸.٦۱	۹۸.٦۱	۹۸.٦۱
DLBCL	۹۷.٤۰	۹۸.۷۰	۱۰۰
Prostate	۷۱.٥٦	۹٤.۱۱	۹٥.۰۹
Average	۹۱.۸۹	۹۷.۸٥	۹۸.٤۲

که نسبت به ستون اول ۲.٦٪ و نسبت به ستون دوم ۱.۳٪ بهبود داشته است. پیشرفت چشمگیر بهبود دقت طبقه‌بندی در مجموعه داده Prostate کاملاً مشهود است، روش پیشنهادی به‌دقت ۹٥.۰۹٪ دست‌یافته است، که به معنای بهبود دقت طبقه‌بندی به میزان ۲۳.٥۳٪ نسبت به ستون اول و ۰.۹۸٪ نسبت به ستون دوم می‌باشد. روش پیشنهادی به میانگین دقت۹۸.٤۲٪ دست‌یافته است که مقدار بالاتری نسبت به مقادیر میانگین اعلام‌شده برای ستون مای دوم و سوم را دارا می‌باشد.

از سوی دیگر، به‌منظور ارزیابی روش پیشنهادی با دیگر رویکردها، دقت به‌دست‌آمده توسط روش پیشنهادی یا به‌اختصار TWCMP، با سه روش FPRS+ISVM، CBFS+ISVM و SNR+kNN مقایسه شده است؛ نتایج حاصل در جدول ۷ و توضیح مختصری در مورد رویکردهای مذکور در ادامه ارائه‌شده‌اند.

اولویت فازی مبتنی بر مجموعه خشن (FPRS): آنالیز اولویت، وظیفه‌ی مهمی در تصمیم‌گیری چندمعیاری است. تئوری مجموعه خشن با جایگزینی روابط معادل با روابط غالب به حل مسئله‌ی آنالیز اولویت رسیدگی می‌کند. اولویت فازی مبتنی بر مجموعه خشن، روابط اولویت فازی را از نمونه‌هایی که با معیارهای عددی مشخص‌شده‌اند، استخراج می‌کند. درواقع، روابط اولویت فازی در مدل مجموعه‌ی خشن فازی گنجانده می‌شود. جهت محاسبه‌ی ارتباط میان معیار و تصمیمات، مقیاس وابستگی ویژگی مدل مجموعه فازی پاولک تعمیم داده‌شده است [۲۳]. یادگیری SVM استاندارد یا SVM استقرایی (ISVM)، در تلاش برای تفکیک داده‌ها در فضای ورودی با داده مای آموزشی جدید و توسط یک ابر صفحه است؛ در مواردی که داده مای آموزشی خطی جدایی‌ناپذیر باشند، تابع هدفِ یادگیری ISVM با حل مسئله‌ی بهینه–سازی درجه دوم، ابر صفحه را به‌طوری‌که فاصله‌ی میان نزدیک‌ترین بردارها با ابر صفحه حداکثر باشد، مشخص می–کند [۲٤]. جستجو مبتنی بر همسازی در انتخاب ویژگی (CBFS)، بر مقیاس ناهمسازی تمرکز دارد؛ به‌طوری‌که یک زیرمجموعه ویژگی با حداقل دو نمونه با مقادیر ویژگی یکسان اما برچسب طبقات متفاوت ناسازگار محسوب می‌شود [۲٥]. نسبت سیگنال به نویز (SNR) الگوهای بیانی را با حداکثر اختلاف در میانگین حالت میان دو گروه و حداقل انحراف حالت درون هر گروه شناسایی می‌کند. نسبت سیگنال به نویز به‌صورت معادله‌ی زیر تعریف می‌شود:

$$SNR = \frac{m_1 - m_2}{s_1 + s_2} \qquad (۸)$$

به ترتیب و بیانگر میانگین و انحراف معیار از طبقه i با در نظر گرفتن ویژگی‌های متناظر می‌باشند [۲۶]. الگوریتم K همسایه نزدیک (KNN) الگوریتم ساده‌ای است که تمام موارد موجود را ذخیره کرده و موارد جدید را برمبنای مقیاس شباهت طبقه‌بندی می‌کند. K همسایه‌ی نزدیک، در تلاش برای یافتن k نمونه با بیشترین شباهت به‌عنوان نزدیک‌ترین همسایگان برای نمونه‌ی داده‌شده و پیش‌بینی طبقه‌ی نمونه بر اساس اطلاعاتِ همسایگان انتخاب‌شده می‌باشد. برای نمونه داده‌شده، ابتدا فاصله‌ی اقلیدسیِ نمونه از دیگر نقاط مجموعه داده محاسبه می‌شود. فاصله‌ها از کمترین تا بیشترین مقدار، جهت شناسایی KNN مرتب می‌شوند. با بررسی هویت کلاس در میان نقاط KNN، نظرسنجی انجام می‌شود. بر اساس کلاسِ اکثریتِ KNN، نمونه به کلاسی در مجموعه داده منتسب می-شود. اگر کلاس منتسب یافته و کلاس واقعی نمونه تطابق داشته باشند، آزمون موفقیت‌آمیز محسوب می‌شود. بدیهی است که اکثریت آراء KNN ها، زمانی رخ می‌دهد که اکثریت متغیرهای اندازه‌گیری هم رای باشند [۲۷].

جدول ۷: مقایسه‌ی دقت طبقه‌بندی چهار روش مختلف

Dataset	TWCMP	FPRS+ISVM	CBFS+ISVM	SNR+kNN
SRBCT	۱۰۰	۹۸.۸۹	۹۵.۱٤	۸۱.۸۰
Leukemia	۹۸.٦۱	۹۸.۷۵	۱۰۰	۹٤.٦۲
DLBCL	۱۰۰	۹۵.۷۱	۹۱.٤۲	۸۸.٤۵
Prostate	۹۵.۰۹	۹۳.۲۷	۹٤.۰۹	۹۲.۲۷
Average	۹۸.٤۲	۹٦.٦۵	۹۵.۱٦	۸۹.۲۸

با توجه به جدول ۷، می‌توان مشاهده کرد که در سه مجموعه داده‌ی سرطانی SRBCT، DLBCL و Prostate، روش پیشنهادی دقت بالاتری نسبت به دیگر روش‌ها داراست. در مجموعه داده‌ی سرطان خون، رویکرد CBFS+ISVM به‌دقت ۱۰۰٪ دست‌یافته است که به میزان ۱.۳۹٪ عملکرد بالاتری نسبت به روش پیشنهادی دارد. به‌طور میانگین، در چهار مجموعه داده‌ی سرطانی مورداستفاده، نتایج حاصل از آزمایش‌ها، بیانگر بهبود دقت طبقه‌بندی توسط روش پیشنهادی می‌باشند؛ این بهبود به میزان ۱.۷۷٪ نسبت به روش FPRS+ISVM، 3.26% نسبت به روش CBFS+ISVM و ۹.۱٤٪ نسبت به روش SNR+kNN می‌باشد.

۷- نتیجه‌گیری

اخیراً، فن مای زیادی به‌منظور انتخاب ویژگی و طبقه-بندی داده مای ریزآرایه ارائه‌شده‌اند. به‌منظور بالا بردن دقت طبقه‌بندی، رویکرد پیشنهادی به این صورت می‌باشد: ابتدا با استفاده از روش انتخاب ویژگی مبتنی بر اهمیت ویژگی‌ها، ویژگی‌ها بااهمیت بالا تعیین می‌شوند؛ این ویژگی‌ها در مجموعه داده حفظ و ویژگی‌هایی بااهمیت کم به‌منظور کاهش حجم مجموعه داده حذف می‌شوند. سپس ویژگی‌های استخراج‌شده از مرحله‌ی اول، با استفاده از ماشین بردار پشتیبانی دوقلو برای خوشه‌بندی، خوشه‌بندی می‌شوند. در گام بعدی به‌منظور انتخاب مهم‌ترین ویژگی‌های درون هر خوشه، الگوریتم انتخاب ویژگی مبتنی بر همبستگی اعمال می‌شود. ویژگی‌های منتخب از تمامی خوشه‌ها در کنار یکدیگر قرارگرفته و مجموعه داده‌ی جدید، قابل‌اعتماد و نهایی جهت طبقه‌بندی را تشکیل می‌دهند. در مرحله‌ی آخر مجموعه داده ی جدید، توسط طبقه بند پرسپترون چندلایه، طبقه‌بندی می شود. نتایج به‌دست‌آمده بر روی مجموعه داده مای مورداستفاده، بیانگر کارایی روش پیشنهادی و بهبود دقت طبقه‌بندی می‌باشد.

۷. مراجع

[1] Mohamad, Mohd Saberi, Safaai Deris, and Rosli Md Illias. "A hybrid of genetic algorithm and support vector machine for features selection and classification of gene expression microarray." International Journal of Computational Intelligence and Applications 5.01 (2005): 91-107.

[2] Maulik, Ujjwal, and Debasis Chakraborty. "Fuzzy preference based feature selection and semisupervised SVM for cancer classification." IEEE transactions on nanobioscience 13.2 (2014): 152-160.

[3] Yu, Hualong, and Sen Xu. "Simple rule-based ensemble classifiers for cancer DNA microarray data classification." Computer Science and Service System (CSSS), 2011 International Conference on. IEEE, (2011).

[4] Xu, Qifeng, and Xuegong Zhang. "Multiclass feature selection algorithms base on R-SVM." Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit & International Conference on. IEEE, (2014).

[5] Mohamad, Mohd Saberi, et al. "Selecting informative genes from microarray data by using hybrid methods for cancer classification." Artificial Life and Robotics 13.2 (2009): 414-417.

[6] Jain, Indu, Vinod Kumar Jain, and Renu Jain. "Correlation feature selection based improved-Binary Particle Swarm Optimization for gene selection and cancer classification." Applied Soft Computing 62 (2018): 203-215.

[7] Mahajan, Shafa, and Shailendra Singh. "Informative Gene Selection Using Adaptive Analytic Hierarchy Process (A2HP)." Future Computing and Informatics Journal (2017).

[8] Aziz, Rabia, et al. "Artificial neural network classification of microarray data using new hybrid gene selection method." International Journal of Data Mining and Bioinformatics 17.1 (2017): 42-65.

[9] Zhang, Li, et al. "Applying 1-norm SVM with squared loss to gene selection for cancer classification." Applied Intelligence (2017): 1-13.

[10] Han, Fei, et al. "A Gene Selection Method for Microarray Data Based on Binary PSO Encoding Gene-to-Class Sensitivity Information." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 14.1 (2017): 85-96.

[11] H. Liu and H. Motoda. Computational Methods of Feature Selection. Chapman & Hall, Boca Raton, FL, (2008).

[12] Park, Chan Hee, and Seoung Bum Kim. "Sequential random k-nearest neighbor feature selection for high-dimensional data." Expert Systems with Applications 42.5 (2015): 2336-2342.

[13] Saeys, Yvan, Iñaki Inza, and Pedro Larrañaga. "A review of feature selection techniques in bioinformatics." bioinformatics 23.19 (2007): 2507-2517.

[14] Ahmad, Amir, and Lipika Dey. "A feature selection technique for classificatory analysis." Pattern Recognition Letters 26.1 (2005): 43-56.

[15] Hall, Mark A. "Correlation-based feature selection of discrete and numeric class machine learning." In Proceedings of the Seventeenth International Conference on Machine Learning, (2000): 359–366.

[16] Hall, Mark Andrew. "Correlation-based feature selection for machine learning." Doctoral dissertation, University of Waikato, Dept. of Computer Science, (1999).

[17] P. Berkhin. Survey of clustering data mining techniques. In J. Kogan, C. K. Nicholas, and M. Teboulle, editors, Grouping Multidimensional Data: Recent Advances in Clustering. Springer, ,(2006): pp. 25-71.

[18] Pratima, Depa, and Nivedita Nimmakanti. "Pattern Recognition Algorithms for Cluster Identification Problem." Special Issue of International Journal of Computer Science & Informatics (IJCSI), Vol.-II, No.-1, (2012).

[19] Wang, Zhen, et al. "Twin support vector machine for clustering." IEEE transactions on neural networks and learning systems 26.10 (2015): 2583-2588.

[20] Sun, Yanmin, Andrew KC Wong, and Mohamed S. Kamel. "Classification of imbalanced data: A review." International Journal of Pattern Recognition and Artificial Intelligence 23.04 (2009): 687-719.

[21] Zhang, Yudong, et al. "A multilayer perceptron based smart pathological brain detection system by fractional Fourier entropy." Journal of medical systems 40.7 (2016): 173.

[22] "Cancer gene expression data sets and their visualizations" [Online]. Available: http://www.biolab.si/supp/bi-cancer/projections/

[23] Hu, Qinghua, Daren Yu, and Maozu Guo. "Fuzzy preference based rough sets." Information Sciences 180.10 (2010): 2003-2022.

[24] Singla, Anshu, Patra, Swarnajyoti, and Bruzzone, Lorenzo." A novel classification technique based on progressive transductive SVM learning." Pattern Recognition Letters, 42 (2014): 101-106.
[25] Dash, Manoranjan, and Huan Liu. " Consistency-based search in feature selection." Artificial Intelligence 151.1-2 (2003): 155-176.
[26] Begum, S., Bera, S. P., Chakraborty, D., & Sarkar, R. "Breast cancer detection using feature selection and active learning." Proceedings of the International Conference on Advancement of Computer Communication and Electrical Technology (ACCET 2016). Computer, Communication and Electrical Technolog 42 (2017): 54-59.
[27] Mejdoub, Mahmoud, and Amar Chokri Ben. " Classification improvement of local feature vectors over the KNN algorithm." Multimedia Tools and Applications 64.1 (2013): 197-218.