

روش های آماری در تحقیقات پزشکی 2016.

جلد 25 (5) 1823-1804



ابویسنده(گان) ۲۰۱۳

تجدید چاپ ۲۰۱۴

sagepub.com

DOI: 10.1177/0962280213502437

/journalsPermissions.nav

sagepub.co.uk

©SAGE

مقایسه روش های یادگیری ماشینی برای طبقه بندی با استفاده از شبیه سازی با چندین داده واقعی نمونه هایی از ذهنی مطالعات سلامت

میزانورخوندوکر^{۱،۲}، اریچارد دابسون^{۲،۳}، کارولین اسکیرو^۴
اندروسیمونز^{۲،۳} و دنیل استال ابرای طرح ابتکاری تصویربرداری عصبی بیماری
آلزایمر*

چکیده

پیشینه: ادبیات اخیر در مورد مقایسه روش های یادگیری ماشین، سوالاتی را در مورد بنی طرفی، بنی طرف و سودمندی بسیاری از مطالعات تطبیقی مطرح کرده است. گزارش نتایج روی مجموعه داده های مطلوب و خطای نمونه گیری در معیارهای عملکرد تخمینی بر اساس نمونه های واحد، منابع اصلی سوگیری در چنین مقایسه هایی تلقی می شوند. عملکرد بهتر در یک یا چند مورد لزوماً به معنای عملکرد بهتر در سطح میانگین یا در سطح جمعیت نیست و مطالعات شبیه سازی ممکن است جایگزین بهتری برای مقایسه عینی عملکرد الگوریتم های یادگیری ماشین باشند.

روش ها: ما عملکرد طبقه بندی تعدادی از الگوریتم های یادگیری ماشین مهم و پرکاربرد، یعنی جنگل های تصادفی (RF)، ماشین های بردار پشتیبان (SVM)، تحلیل تفکیک خطی (LDA) و k-نزدیک ترین همسایه (kNN) را مقایسه من کنیم. با استفاده از پردازش موازی گسترشده بر روی ابررایانه های با کارایی بالا، خطاهای تعمیم را در ترکیبات مختلف سطوح چندین عامل مقایسه می کنیم: تعداد ویژگی ها، اندازه نمونه آموزشی، تنوع بیولوژیکی، تنوع تجربی، اندازه اثر، تکرار و همبستگی بین ویژگی ها.

۱) کالج کینگ لندن، موسسه روانپژوهی، گروه آمار زیستی، لندن، انگلستان
۲) کالج کینگ لندن، موسسه روانپژوهی، مرکز تحقیقات زیست پزشکی NIHR برای سلامت روان در جنوب لندن و بنیاد NHS مادرلی، لندن، انگلستان

۳) کالج کینگ لندن، موسسه روانپژوهی، واحد تحقیقات زیست پزشکی NIHR برای زوال عقل در جنوب لندن و بنیاد خدمات درمانی ملی مادرلی، لندن، انگلستان

۴) کالج کینگ لندن، موسسه روانپژوهی، مرکز روانپژوهی اجتماعی، ژنتیک و رشدی MRC، انگلستان

* یکی از مجموعه داده های نمونه مورد استفاده در تهیه این مقاله از یاگاه داده ابتكار تصویربرداری عصبی بیماری آلزایمر (ADNI) (adni.loni.ucla.edu) به دست آمده است. به همین ترتیب، محققان ADNI در طراحی و بیان سازی ADNI مشارکت داشته و/یا داده ها را ارائه داده اند. اما در تجزیه و تحلیل با نوشتن این گزارش مشارکت نداشته اند. فهرست کامل از محققان ADNI را می توانید در آدرس زیر بباید: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

نویسنده مسئول:

میزانورخوندوکر، گروه آمار زیستی، موسسه روانپژوهی، کالج کینگ لندن، لندن، انگلستان. ایمیل:
mizanur.khondoker@kcl.ac.uk

نتایج: برای تعداد کمتری از ویژگی‌های هم سته، تعداد ویژگی‌هایی که تقریباً از نصف اندازه نمونه تجاوز نمی‌کنند، LDA از نظر میانگین خطاهای تعمیم و همچنین پایداری (دقت) تخمین خطا، روش انتخابی بود. SVM (با هسته RBF) با افزایش مجموعه ویژگی‌ها، به شرط که اندازه نمونه خیلی کوچک نباشد (حداقل 20)، با اختلاف زیادی از LDA و RF بهتر عمل می‌کند. عملکرد KNN همچنین با افزایش تعداد ویژگی‌ها بهبود می‌یابد و از LDA و RF همچنین kNN بهتر عمل می‌کند. مگر اینکه تغییرپذیری داده‌ها خیلی زیاد باشد و/یا اندازه های اثر خیلی کوچک باشند. مشخص شد که kNN فقط در برخی موارد که داده‌ها متغیرتر و اندازه های اثر کوچکتری دارند، از kNN بهتر عمل می‌کند، که در این موارد نیز تخمین های خطای پایدارتری نسبت به kNN و LDA ارائه می‌دهد. کاربردها در تعدادی از مجموعه داده‌های واقعی، یافته‌های مطالعه شیوه سازی را پشتیبانی می‌کنند.

کلمات کلیدی
پادگیری ماشین، اعتبارسنجی متقابل، خطای تعمیم، توزیع ناقص، ریزآرایه‌ها، الکتروانسفالوگرام (EEG)، تصویربرداری تشیددمغناطیسی (MRI)

اپیشینه

پیشرفت‌های اخیر در فناوری‌های ژئوم، پروتئومیک، تصویربرداری عصبی و سایر فناوری‌های با توان عملیاتی بالا منجر به افزایش داده‌های با ابعاد بالا شده است که نیازمند توسعه روش‌های جدید یا اصلاح تکنیک‌های آماری و پادگیری ماشین موجود برای بدائل رساندن بهره اطلاعاتی از چنین داده‌هایی است. افزایش تعداد روش‌های موجود، مقایسه روش‌ها را به طور منطقی ضروری کرده است تا بهترین روش در یک موقعیت خاص پیدا شود که منجر به انتشار مقالات متعددی با تمکن بر مطالعات تطبیقی در ادبیات بیوانفورماتیک و زیست‌شناسی محاسباتی اخیرشده است. بخش بزرگی از این مطالعات، روش‌های آماری و پادگیری ماشینی تحت نظرارت را برای طبقه‌بندی موضوعات، عمدها بر اساس بیان ژن ریزآرایه یا داده‌های طیف سنجی جرمی با ابعاد بالا، مقایسه کرده اند.^{۱۰-۱۱}

ادبیات اخیر^{۱۲-۱۴} ادر مورد این موضوع، سوالاتی در مورد بی طرفی، بی سودمندی و روش‌های انجام اکثر این مقایسه‌های مطرح شده است، زیرا بین یافته‌های چنین مطالعاتی اجماع کمی وجود دارد. مروری توسط بولستیکس و همکاران^{۱۳} نشان دهنده تمایلی در برخی مطالعات تطبیقی برای نشان دادن برتری یک روش خاص با استفاده از مجموعه داده‌هایی است که به نفع روش انتخاب شده عمل می‌کنند. نگرانی‌های مشابهی در یک مطالعه اخیر نیز تکرارشده بیوانفورماتیک مقاله یوسفی و همکاران^{۱۴} اپیشنهاد (i) گزارش نتایج روی مجموعه داده‌های مطلوب و (ii) به اصطلاح سوگیری چند قاعده‌ای که در آن چندین قانون طبقه‌بندی روی مجموعه داده‌ها مقایسه می‌شوند تا مزیت یکروش خاص را نشان دهند، زیرا منابع اصلی سوگیری در چنین مقایسه‌هایی هستند.

مطالعات تطبیقی وجود دارند که هدف آنها نشان دادن یک روش خاص به عنوان بهتر از بقیه نیست. چندین مطالعه‌ای این دست، به اصطلاح ختنی مقایسه‌ها، توسط بولستیکس و همکاران ذکر شده اند. یکی از محدودیت‌های این مطالعات این است که مقایسه‌ها عمدهاً مبتنی بر مجموعه داده‌های واقعی هستند و مشکلی که در مقایسه عملکرد طبقه‌بندی تخمین زده شده روی مجموعه داده‌های واقعی وجود دارد، خطای نمونه گیری یا نویز در معیارهای عملکرد تخمین زده است. با توجه به این واقعیت که تخمین‌های عملکرد تابع تغییرپذیری نمونه گیری هستند، بهترین عملکرد در یک یا چند مورد لزوماً به معنای آن نیست که به طور متوسط یا در سطح جمعیت نیز همینطور باشد.

سه مسیر جایگزین را می‌توان به طور بالقوه برای مقایسه ای قوی تر و عینی تر بررسی کرد:
(۱) استفاده از آزمون آماری برای در نظر گرفتن تغییرپذیری نمونه برداری یا نویز در تخمین‌های عملکرد، (۲) مقایسه تحلیلی بر اساس نظریه توزیع تخمین‌های عملکرد و (۳) تخمین مکرر معیار عملکرد روی تعداد زیادی از داده‌های شبیه سازی شده/مصنوعی برای میانگین گیری تغییرپذیری نمونه برداری یا نویز از تخمین

معیار عملکرد. هانزار و دوگرتی^{۲۴} امکان استفاده از آزمون های آماری برای مقایسه عملکرد را برسی کرد و نتیجه گرفت که مقایسه مستقیم بر اساس آزمون آماری غیرقابل اعتماد است و اغلب منجر به نتیجه گیری های اشتباه شود. مقایسه تحلیلی یک رویکرد زیبا خواهد بود اما نیاز به یافتن توزیع نمونه گیری از تخمین های عملکرد بر اساس توزیع مشترک متغیرهای ویژگی و کلاس دارد. با توجه به این واقعیت که بسیاری از الگوریتم های یادگیری ماسین مدرن پیچیده هستند (به اصطلاح جعبه سیاه) و اغلب مبنی بر هیچ مدل آماری زیربنایی نیستند، تعیین ویژگی های توزیعی تحلیلی تخمین های عملکرد به جز برعی روشن های کلاسیک تبعیض آماری مانند تحلیل تبعیض خطی (LDA) امکان پذیرنیست.^{۲۵} همانطور که توسط هوآ و همکاران نشان داده شده است،^{۲۶} این امر، مسیر شبیه سازی را به عنوان ابزاری عملی و پایدار برای مطالعه عینی ویژگی های معیارهای عملکرد مبتنى بر الگوریتم های یادگیری با طیف وسیعی از پیچیدگی ها و اشکال، باز من گذارد. این امر در مطالعات دیگر در ادبیات مانند هانزار و دوگرتی نیز محقق شده است.^{۲۷} که نتیجه گرفت که «مقایسه ای قواعد طبقه بنده در داده های واقعی بدتر از آزمایش های داده های مصنوعی است... قویاً پیشنهاد من کند که محققان مطالعات تطبیقی خود را روی مجموعه داده های مصنوعی انجام دهند».

ماهمچینین فکر من کنیم که مطالعات شبیه سازی روشنی مناسب و کاربردی برای فهمیدن این است که «کدام روشن در چه شرایط عملکرد بهتری دارد». حقیقت همیشه در داده های شبیه سازی شده مشخص است و بنابراین بررسی سوگیری، یعنی نزدیکی یک تخمین به حقیقت، که با داده های واقعی آسان نیست، آسان است. داده های مصنوعی همچنین امکان مطالعه خواص یک تخمین گر را در سطوح مختلف ویژگی های مختلف داده ها مانند تغییرپذیری، اندازه اثر، همبستگی و غیره فراهم می کنند. با این حال، نقش داده های واقعی نیز مهم است و باید برای تکمیل تحقیقات مبتنی بر شبیه سازی استفاده شود، زیرا الگوها و ساختارها در داده های واقعی عموماً بسیار پیچیده تر هستند و هیچ مدل شبیه سازی نمی تواند الگوها، ابعاد و منابع تغییرپذیری در داده های تولید شده از یک سیستم بیولوژیکی واقعی را به طور کامل ثبت کند.

در این مطالعه، ما یک آزمایش شبیه سازی گسترده را برای مقایسه عملکرد طبقه بنده تعدادی از الگوریتم های یادگیری ماسین مهم و پرکاربرد، از کلاسیک ترین LDA^{۲۸-۳۰} گرفته تا ...، انجام من دهیم. ابه روش های مدرن مانند ماسین های بردار پشتیبان (SVM) اگرچه مقایسه ها در مقالات عمدتاً بر روی داده های واقعی انجام شده است، اما داده های مصنوعی نیز تا حدودی قبلاً مورد توجه قرار گرفته اند.^{۳۱-۳۴} با این حال، مطالعات شبیه سازی که قبلاً مورد استفاده قرار گرفته بودند، از نظر تعداد ویژگی های داده ها و پوشش آنها (فضای پارامتر) محدود بودند. کاربردهای طرح های چند عاملی برای آزمایش های شبیه سازی توسط اسکرونداال موربد بحث قرار گرفت.^{۳۵} در زمینه آزمایش های مونت کارلو، بررسی همزمان چندین عامل در سطوح مختلف، به بهبود اعتبار خارجی، یعنی میزان تعمیم نتایج به موقعیت های دیگر و به جمعیت، از نتایج مطالعه شبیه سازی کمک من کند. با استفاده از پردازش موازی گسترده بر روی ابرایانه های با کارایی بالا (مرکز محاسبات و داده های ادینبورگ، ECDF و مرکز تحقیقات زیست پژوهشی NIH) برای خوشه لینوکس سلامت روان)، خطاهای تعمیم (خطاهای اعتبارسنجی متقاطع (CV) را برای تعداد زیادی از ترکیبات هفت عامل زیر ارزیابی و مقایسه من کنیم: تعداد متغیرها (ص)، حجم نمونه آموزشی (ن)، تنواع بیولوژیکی (با بین آزمودنی ها) (-ب)، تغییرات درون آزمودنی (-ای)، اندازه اثر (تغییر برابر، -)، تکرار (رو) همبستگی () بین متغیرها. ما معتقدیم که هیچ تک سایز مناسب برای همه روش نوع در یادگیری ماسین، که مدت ها پیش توسط و پرلت محقق شد^{۳۶} و در اطلاعیه اخیر تکرار شدی بیانفورماتیک مقاله سرمهای از همه روش های دهد که هیچ روش طبقه بنده وجود ندارد که در همه شرایط از همه روش های دیگر بهتر عمل کند. انگیزه در نظر گرفتن چنین طیف وسیعی از عوامل، ارائه راهنمایی در مورد اینکه «کدام روش در چه شرایط بهتر عمل من کند» است. ما همچنین یافته های خود را در مورد داده های شبیه سازی شده با ارزیابی عملکرد روی تعدادی از مجموعه داده های تجربی واقعی تولید شده از طیف وسیعی از پلتفرم های با توان عملیاتی بالا مانند داده های بیان ژن از ... تکمیل من کنیم.

ریزآرایه های DNA، داده های تصویربرداری عصبی از سیستم تصویربرداری تشخیص مغناطیسی (MRI) باوضوح بالا، و داده های پتانسیل مرتبه با رویداد (ERP) که فعالیت مغز را از سیستم الکتروانسفالوگرام (EEG) اندازه گیری می کند. برنامه شبیه سازی (simData) و برنامه تخمین عملکرد (classificationError) به عنوان بخشی از بسته R به نام optBiomarker ارائه شده اند که از شبکه جامع بایگانی (<http://www.cran.r-project.org/web/packages/optBiomarker>) قابل دسترس است.

۲.۱ روش های طبقه بندی

معاملکرد طبقه بندی تعدادی از روش های طبقه بندی پرکاربرد را که مبتنی بر طیف متنوعی از الگوریتم ها و معناری ها هستند، یعنی درخت تصمیم و روش مبتنی بر نمونه گیری مجدد، جنگل های تصادفی، RF، مقایسه من کنیم.^{۲۳} الگوریتم یادگیری مبتنی بر هسته، SVM^{۲۴} روش کلاسیک با انگیزه آماری، LDA^{۲۵} و الگوریتم مبتنی بر نمونه (نژدیک ترین نمونه های آموزشی) k-نژدیک ترین همسایه، kNN^{۲۶}.

۲.۲ بهینه سازی پارامترهای تنظیم

اکثر الگوریتم های طبقه بندی پارامترهای تنظیم کننده مخصوص به خود را دارند که در حالت ایده آل نیاز به بهینه سازی روی هر مجموعه داده ای که روش ها روی آن اعمال می شوند، دارند. ما پارامترهای مهم قابل تنظیم توسط کاربر را برای هر روش روی هر مجموعه داده ی شبیه سازی شده با استفاده از جستجوی شبکه ای روی فضاهای پارامتر ارائه شده، بهینه من کنیم. فضاهای جستجوی ما برای پارامترهای تنظیم کننده همیشه شامل مقادیر پیش فرض نرم افزار برای پارامترهای مربوطه بودند تا اطمینان حاصل شود که تخمین های عملکرد در پارامترهای بهینه شده حداقل به خوبی مقادیر در گزینه های پیش فرض هستند.

پیشنهادی شود که روش RF نسبت به تغییرات پارامترهای تنظیم خود کاملاً مقاوم باشد. لئو بریمن (در راهنمای برنامه اصلی FORTRAN برای RF) پیشنهاد کرد که متري (تعداد متغیرهای که به عنوان کاندید در هر گره استفاده می شوند) تنها پارامتری است که نیاز دارد

مقداری قضاوت برای تعیین. ما بهینه سازی می کنیم متريکبا استفاده از جستجوی شبکه ای روی یک نمونه تصادفی با اندازه ۵، شامل از مقدار پیش فرض (متريک $\frac{1}{2} \log(\frac{n}{P})$)^{۲۷} از دنباله فا، ..., $\frac{1}{2} \log(n)$ که در آن نمادگذاری [ایکس][نشان دهنده تابعکف، که بزرگترین عدد صحیح غیر بزرگتر از را من دهدایکس. برای ص، ..., $\frac{1}{2} \log(n)$ کمتر از پنج عنصرخواهد داشت که در این صورت جستجوی شبکه ای را روی خود دنباله انجام من دهیم. این موضوع در مقالات پیشنهادشده است (مثلًا دیاز-اوریارت و آلوارز د آندرس)^{۲۸} که برای عملکرد عالی، نیاز کمی به تنظیم دقیق سایر پارامترهای RF وجود دارد. با این حال، ما بهینه سازی را در نظر گرفتیم. اندازه گره (حداقل اندازه گره های ترمیمال) و نتری (تعداد درختان برای رشد) علاوه بر متريک پارامتر، پارامتر گره بندی طول درختان را به صورت no node با تعداد موارد کمتر از ... کنترل می کند. گره بندی تقسیم خواهد شد. اندازه های گره کوچکتر معمولاً دقت بهتری ارائه می دهد و اندازه گره بزرگتر برای مجموعه داده های بزرگتر، معمولاً با هزینه کاهش اندک دقت، مزیت محاسباتی ایجاد می کند. با این حال، ما جستجوی شبکه ای را روی فضای پارامتر اجرا می کنیم. فراز پیش فرض، ..., $\frac{1}{2} \log(n)$ برای هر مجموعه داده شبیه سازی شده برای یافتن مقدار بهینه اندازه گره پارامتر ان تری تعداد درختانی که قرار است در جنگل کاشته شوند را کنترل می کند. به طور شهودی، جنگل بزرگتر (درختان بیشتر) باید برای پایداری بهتر باشد. در اینجا، ما تنظیم می کنیم ان تری برای دقت بهینه در فضای پارامتر ف، ..., $\frac{1}{2} \log(n)$ شامل مقدار پیش فرض randomForest (500) است.

ما SVM را با هسته های خطی، چندجمله ای و تابع پایه شعاعی (RBF) ارزیابی کردیم و RBF را برای همه محاسبات انتخاب کردیم زیرا مشخص شد که در شبیه سازی ما نسبت به هسته چندجمله ای کمتر بایاس دارد و همچنین به این دلیل که از هسته خطی عمومی تر است. این با استفاده از آرگومان هسته (kernel) انتخاب من شود.^{۲۹} "شعاعی" تابع R. عملکرد SVM با هسته RBF ممکن است به ... بستگی داشته باشد. هزینه (ثابت C جمله جریمه در ضرایب لگرانز) و

گاما(پارامتر عرض معمکوس تابع هسته RBF). هزینه پارامتر حاشیه بردارهای پشتیبان را کنترل می‌کند - مقدار کوچکتر، جریمه خطاهای حاشیه را کاهش می‌دهد (نقاط جریمه نزدیک به مرز را نادیده می‌گیرد) و از این رو حاشیه طبقه بندی را فیزیش می‌دهد. مقدار گاما انحصاری مرز تصمیم گیری را کنترل می‌کند - مقادیر بالاتر، مرز تصمیم گیری را انعطاف پذیرتر(غیرخطی) می‌کند. ما بهینه می‌کنیم هزینه و گاما پارامترها با استفاده از جستجوی شبکه ای روی فضاهایی متنشکل از نمونه‌های یکنواخت با اندازه ۵ (شامل مقادیر پیش‌فرض) از محدوده های (۰/۱، ۰/۱) و (۱/۰، ۰/۱)، به ترتیب. ممکن است اشاره شود که فضاهای پارامتری طوری انتخاب می‌شوند که بین ... قرار گیرند. پس از پیش‌فرض و ۱۰-پیش‌فرض که در آن مقادیر پیش‌فرض برای هزینه و گاما پارامترها ۱ و ۰/۱ هستند، به ترتیب. kNN نیازدارد ص

فقط تابعی فی نینگک (تعداد

نژدیکترین همسایه‌ها) که ما در طول محدوده بهینه می‌کنیم، ۲، ... نرم‌گرم‌ما از اعتبارسنجی متقاطع fold-10 استفاده می‌کنیم که از نظر محاسباتی نسبت به اعتبارسنجی متقاطع leave-one-out برای بهینه سازی همه پارامترها، کم زحمت تر است.

۲.۳ برآوردهای عملکرد

عملکرد روش‌ها را از نظر خطای طبقه بندی، حساسیت و ویژگی مقایسه می‌کنیم. ما از اعتبارسنجی متقابل out-one برای تخمین این معیارهای عملکرد استفاده می‌کنیم. این انتخاب بر اساس ارزیابی و مقایسه چندین تخمین گر، یعنی $CV_{10\text{-fold}}$, $bootstrap_{0.632}$ و $leave-one-out$. به علاوه بوت استرپ اغلب برای مقادیر تخمین گرها، تخمین گرها از نظر بایاس و تغییرپذیری با استفاده از شبیه سازی برای نمونه‌های آموزش با اندازه‌های مختلف مقایسه شدند. ما اندازه‌های نمونه‌های آموزش (ن) از ۰ تا ۲۵۰ متغیر است، اما هر چهار تخمین گر را نمی‌توان برای کل بازه مقایسه کرد زیرا تخمین گرهای بوت استرپ و ۶۳۲ به علاوه بوت استرپ اغلب برای مقادیر کوچک تزمینجیر به مشکلات محاسباتی می‌شوند. بنده دلیل ماهیت بسیار نامتوازن داده‌های بازنمونه گیری شده برای اینکه برآوردهای مبتنی بر بوت استرپ در مقایسه باقی بمانند، اندازه نمونه را حداقل به ۵۰ محدود کردیم و مقایسه را بر اساس میانگین عملکرد در ۰-۱ اندازه نمونه آموزشی مختلف انجام دادیم. (ن ۵۰٪، ۶۰، ...، ۱۴۰) نشان می‌دهد که برآوردهای مبتنی بر بوت استرپ کمتر از ۰/۱ هستند، اما حتی برای مقادیر متوسط (ن

(۵) اندازه نمونه‌های آموزشی (داده‌ها نشان داده نشده اند). همانطور که قصد داریم مقایسه کنیم.

روش‌هایی برای طیف وسیعی (بان با اندازه نمونه ای به کوچکی ۰/۱)، ما بین برآوردهای واریانس یک طرفه (leave-one-out) و واریانس ده طرفه (fold CV-10) یکی را انتخاب می‌کنیم. برآوردهای واریانس یک طرفه برای نمونه‌های کوچک تر، نسبت به برآوردهای واریانس ده طرفه متناظر، اریبی کمتری دارند اما متغیرتر هستند. مقایسه ما روی داده‌های شبیه سازی شده نشان می‌دهد که برآوردهای واریانس یک طرفه تغییرپذیری بسیار مشابهی با برآوردهای واریانس ده طرفه دارد (شکل تکمیلی S1 را ببینید). در نهایت، برآوردهای واریانس یک طرفه را برای همه مقایسه انتخاب می‌کنیم، زیرا برای داده‌هایی با هر اندازه‌ای قابل اجرا است و نسبت به برآوردهای واریانس ده طرفه، اریبی کمتری دارد.

۲.۴ مجموعه داده‌های نمونه

عملکرد روش‌ها را روی تعدادی از مجموعه داده‌های واقعی تولید شده از طیف وسیعی از پلتفرم‌های با توان عملیاتی بالا مانند داده‌های بیان زن از ریزآرایه‌های DNA، داده‌های تصویربرداری عصبی از سیستم MRI با وضوح بالا و داده‌های ERP اندازه گیری فعالیت مغز مشتق شده از سیستم EEG ارزیابی کردیم.

۲.۴.۱ داده‌های بیان زن دوقطبی

این مجموعه داده بر اساس مطالعه بیان زن ریزآرایه است. بافت مغز پس از مرگ بزرگسالان (قشر پیش‌پیشانی پشتی-جانبی) از افراد مبتلا به اختلال دوقطبی و افراد سالم Affymetrix

از پلتفرم HG-U133A GeneChips 22000 رونوشت mRNA از 61 فرد (30 فرد مبتلا به اختلال دوقطبی و 31 فرد سالم) استفاده شد. داده های نرمال شده RMA پیش پردازش شده این آزمایش از(/geo/GSE5388، شماره دسترسی GEO)<http://www.ncbi.nlm.nih.gov>

۴.۲.۳ داده های تصویربرداری مغز MRI

داده های تصویربرداری مغز از پاگاه داده طرح تصویربرداری عصبی بیماری آلزایمر (ADNI) (PI Michael M Weiner) در سال ۲۰۰۳ توسط موسسه ملی پیری (NIA)، موسسه ملی تصویربرداری زیست پزشکی و مهندسی زیستی (NIBIB)، سازمان غذا و دارو (FDA)، شرکت های داروسازی خصوصی و سازمان های غیرانتفاعی، به عنوان یک مشارکت دولتی-خصوصی پنج ساله ۶۰ میلیون دلاری راه اندازی شد. هدف اصلی ADNI آزمایش این موضوع بوده است که آیا MRI سریالی، PET و سایر نشانگرهای بیولوژیکی در آزمایشات بالینی MCI و مراحل اولیه بیماری آلزایمر (AD) مفید هستند یا خیر. تعیین نشانگرهای حساس و اختصاصی پیشرفته بسیار زودهنگام آلزایمر به منظور کمک به محققان و پزشکان برای توسعه درمان های جدید و نظرارت بر آنربخشی آنها و همچنین کاهش زمان و هزینه آزمایشات بالینی در نظر گرفته شده است. افراد ADNI تا ۵۵ ساله از پیش از ۵۰ مکان در سراسر ایالات متحده و کانادا در این تحقیق شرکت کردند و اطلاعات دقیق تر در www.adni-info.org موجود است. داده های MRI با شدت ۱.۵ تسلای وب سایت ([ADNI](http://www.loni.ucla.edu/ADNI/index.shtml)) دانلود شدند. شرح جمع آوری داده های مطالعه ADNI را می توانید در www.loni.ucla.edu/ADNI/search/Cores پیابید. به طور خلاصه، داده های اسکنرهای ۱.۵ تسلای با داده های جمع آوری شده از سیستم های مختلف MR با پروتکل های بهینه سازی شده برای هر نوع اسکنر استفاده شدند. پوشش کامل مغز و جمجمه برای مجموعه داده های MRI مورد نیاز بود و کنترل کیفیت دقیق بر روی تمام تصاویر MR مطابق با معیارهای کنترل کیفیت منتشر شده قبل انجام شد.^{۳۲} مطابق با [Freesurfer pipeline](http://freesurfer.org) (نسخه ۵.۱.۰) برای MRI استفاده کردیم تا ضخامت قشر مغز منطقه ای و اندازه گیری های حجمی را تولید کنیم. تمام اندازه گیری های حجمی از هر فرد با حجم داخل جمجمه ای فرد نرمال سازی شدند، در حالی که اندازه گیری های ضخامت قشر مغز نرمال سازی نشدن و به شکل خام خود استفاده شدند.^{۳۳}

۴.۳ داده های الکتروانسفالوگرافی

سیگنال های EEG نوسانات ولتاژ ثبت شده از الکترودهای روی پوست سر را اندازه گیری می کنند و شاخصی از فعالیت مغز را ارائه می دهند. این داده ها از ۴۱ بزرگسال با تشخیص فعلی ADHD (اختلال بیش فعالی کمبود توجه) و ۴۷ فرد بدون مشکلات سلامت روان به دست آمده است.^{۳۴} ثبت شد و همکاران شرح داده شده است. McLoughlin : که به تفصیل در CPT-OX (CPT-OX) در طول یک وضعیت استراحت ۳ دقیقه ای (چشمان باز) و یک تکلیف عملکرد مداوم نشانه گذاری شده EEG^{۳۵} و ثبت EEG، و همچنین ارجاع مجدد، نمونه برداری کاهشی و روش های حذف آرتیفیکت چشمی معادل مواردی بود که در Ty_e و همکارانش ذکر شده است.^{۳۶} آزمایش هایی با تعداد مصنوعات بیش از ۲۰۰ بینجمنیک تا پیک در هر کانال از داده های فیلتر شده دیجیتال پایین گذر (0.1-30 هرتز، فیلتر ناچ 50 هرتز، ۱2 دسی بل بر اکتاو) حذف شدند. داده های EEG پیوسته به فواصل 2 ثانیه ای تقسیم شدند و سپس طیف توان با استفاده از تبدیل سریع فوریه با پنجره هینینگ محاسبه شد. معیارهای کم توان طیفی EEG در باندهای فرکانسی EEG در فواصل فرکانسی مربوطه ۰.۵-۳.۵ هرتز، ۷.۵-۱۲.۵ و ۳۰-۳۵ هرتز) میانگین گیری شدند. چگالی توان مطلق (پنجم=۲ هرتز)، چگالی توان نسبی (نسبت هر دامنه فرکانسی منفرد که به چگالی توان جمع شده کمک می کند) و نسبت های - = در مکان های الکترود پوست سر به طور متوسط فرونتال، مرکزی و جداری محاسبه شدند. CPT-OX ها از ERP همانطور که در McLoughlin و همکاران توضیح داده شده است، استخراج شدند.

۱۳۰ اشپیه سازی

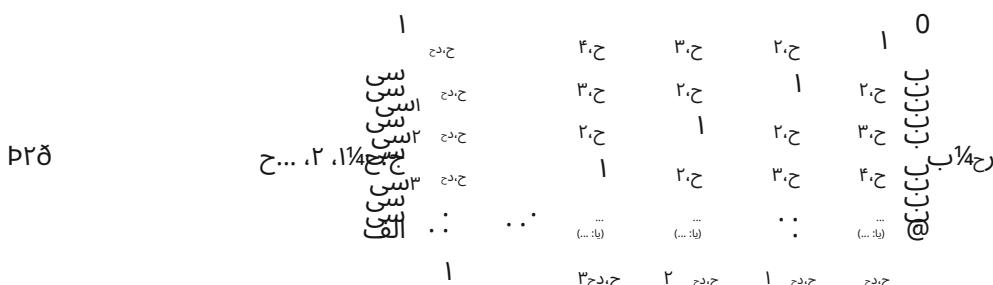
۱۰.۵.۱ شبیه سازی اصلی

مسائل طبقه بندی مبتنی بر داده های با ابعاد بالا استفاده از داده های بیان ژن ریزآرایه نشان داده می شوند. بنابراین، ما مدل شبیه سازی خود را برای تولید داده های بیان ژن واقع گرایانه طراحی می کنیم که در آن می توانیم ویژگی های مختلف داده ها (تغییرپذیری، اندازه اثر، همیستگی و غیره) را به طور سیستماتیک تغییر دهیم تا اثرات چین ویژگی های داده ای را بر عملکرد الگوریتم های طبقه بندی بررسی کنیم. برای اینکه داده های شبیه سازی شده ما تاحد امکان واقعی باشند، مطالعه شبیه سازی خود را بر اساس یک مجموعه داده ریزآرایه واقعی انجام می دهیم. ما مجموعه ای از عبارات پایه () را با میانگین گیری لگاریتم نرمآل شده تعریف می کنیم. داده های بیان ژن روی ۲۸ نمونه خون انسانی گرفته شده از ۲۸ فرد سالمند مجموعه بیان پایه شامل بیان ژن برای ۵۴۳۵۹ نشانگر است که با استفاده از پلتفرم بیان ژن ریزآرایه کل ژنوم انسان CodeLink سنجیده شده اند. تغییرپذیری سیستماتیک و نویز تصادفی در مجموعه بیان پایه را می توان حداقل فرض کرد زیرا داده ها نرمآل سازی و میانگین گیری شده روى افراد زیادی هستند. بنابراین منطقی است که فرض کنیم سطوح بیان پایه () متناسب با سینکل فراوانی mRNA واقعی نشانگرهای مربوطه هستند. سپس از یک مدل اثرات تصادفی برای معرفی مقدار از پیش تعیین شده نویز تصادفی در داده ها استفاده می کنیم. ما دو سطح از تغییرپذیری تصادفی را در نظر می گیریم: تغییرات بین آزمودنی ها و درون آزمودنی ها که بیشتر به عنوان شناخته می شوند. بیولوژیک و فنی تغییرات در مقالات مربوط به ریزآرایه. برای یک اندازه نمونه آموزش مشخص (ن)، ما در ابتدا داده ها را برای هر نشانگر به طور مستقل طبق ... شبیه سازی می کنیم.

Plö

بـ ١/٤ بـ bـ

استقلال به ندرت یک فرض واقع بینانه برای هر داده چندبعدی از جمله بیان $\hat{\theta}$ است که در آن گروه هایی از $\hat{\theta}$ ها به صورت هماهنگ با هم عمل می کنند و روابط شبکه ای را برای انجام عملکردهای بیولوژیکی خاص تشکیل من دهنند. متغیرهای درون چنین گروه ها یا شبکه هایی عموماً همبستگی بالایی با یکدیگر دارند، اما احتمالاً همبستگی های ناچیزی با متغیرهای گروه دیگر نشان می دهنند. ما یک ماتریس همبستگی بلوکی-قطري را برای مدل سازی چنین شبکه ای با استفاده از تولیدیت^{۳۹} ساختار همبستگی برای جبلوک هفتم:



كجا

که از کاهش می یابد کاکتیو برد برای ۲ خط برابر ۱/۴٪. ساختار همبستگی هاب-توپلیتز، یک همبستگی شناخته شده بین یک هاب شبکه (معمولًاً متغیر اول) و هریک از متغیرهای دیگر درون بلوکی که همبستگی بین هاب و ... وجود دارد را فرض می کند. متغیر ام به صورت زیر کاهش می یابد افزایش می یابد. این فرض واقع بینانه تری نسبت به ماتریس همبستگی بلوک-قطري با ساختار تبادلی درون بلوک های است که توسط هوا و همكارانش در نظر گرفته شده است.^{۱۷}

تعداد بلوک‌ها را تعیین می‌کنیم (ح) به ۱ برای ص ۵۵. برای ص ۵۶. برای ص ۵۷=۳-گرم کجا [ایکس] نشان دهنده تابع کف (بزرگترین عدد صحیح که بزرگتر از ازمجموعه انتخاب شده‌است). اگر مضری از استاج، همه بلوک‌ها دارای ابعاد یکسان در نظر گرفته می‌شوند، $D = H \cdot p$ اگر نباشد (ایکس). ص مضری از نیستاج، یکی از بلوک‌ها (معمولًاً اولین بلوک) دارای بُعد در نظر گرفته می‌شود $D = H \cdot \frac{1}{2} \cdot p$ وزارت دفاع β ، $H = D \cdot \beta$ برای بقیه (ح) ۲، ۳، ...، $H = D \cdot \beta$ کجا وزارت دفاع β ، $H = D \cdot \beta$ یادآور تقسیم است پ/ح.

ماز تبدیل ریشه چولسکی برای اعمال ساختار هاب-توپلیتز بلوک-قطری استفاده می کنیم. $\frac{1}{4}$ دیاگ فرا، ...، رح گرم به داده های غیرهمبسته. فرض کنید $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ است-پ-بردار متغیرهای تصادفی که میانگین نشانگرها را در توزارهای مربوطه نشان می دهد. سپس ماتریس کوواریانس $A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$ -توسط داده شده است

نجم-ڏ ۲۰ ۱/۴ گ= رہمن، ای= ڦ

کجا من میک است پ- ماتریس همانی ابعادی. سپس ریشه چولسکی را محاسبه می کنیم سی از پنجم
داده های تبدیل شده با ساختار کوواریانس مورد نظر به صورت زیر بدست می آیندی ۴-ایکس سن.
به منظور معرفی یک منبع سیستماتیک از تغییرات (یعنی تفاوت گروهی)، نمونه آموزش را به دو گروه تقسیم
من کنیم، جی ۱ و جی ۲، از نظر اندازه ن $\frac{1}{2}$ ن = ۲ - و ن $\frac{1}{2}$ ن سپس گروه ها با اندازه اثر متفاوت من شووند ترکیبدر آن تفاوت
(لگاریتم) ۰.۵ تغییر شکل. (یعنی مقداری اضافه می کنیم)
نشان دهنده گروه غیر استاندارد برای مقداری
بیان شبیه سازی شده در جی ۲
برای ایجاد داده ها در جی ۲. بسته به مقدار، من توان آن را افزایش یا کاهش داد. زیه طور تصادفی انتخاب شده از فا، ۱.
۰.
گرمگما از یک توزیع نرمال کوتاه شده با پشتیبانی تولید می کنیم الف، ۱۵، $\frac{1}{2}$ ، $\frac{1}{4}$
یعنی، از توزیع نرمال کوتاه شده ی زیر پیروی می کند:

فجز الالف $\frac{1}{4}$ بـ ، _____

کجا: ۰-۵: به ترتیب نشان دهنده تابع چگال و توزیع یک متغیر نرمال استاندارد است و مقیاس متغیر نرمال صفر میانگین اصلی (کوتاه نشده) است. آستانه تضمین من کند که نشانگرهای که در تحلیل لحظه من کنیم، اندازه اثر غیر صفر داشته باشند. ما عبارت - را بیان می کنیم. $\text{دققه} \times 2\% = \text{عنوان حداقل تعییر برابر نشانگرها}$ که باید در طبقه بندی در نظر گرفته شود. مقدار موردن انتظار معادله (۵) به صورت زیر دارد و شود.

پڻڻ : $\frac{\text{ڦ} \text{ پ}}{-\text{ڦ} \text{ پ}}$ ۱/۴ زولalf ای

جدول ۱. مقادیر اندازه نمونه آموزشی (ن)، تنوع زیستی (-ای)، تغییرات تجربی (-ای)، حداقل تغییر تاشدگی (-دقیقه) و تکرار (۲) در شبیه سازی در نظر گرفته شده است.

ویژگی های داده ها (فاکتورها)				
ر	-دقیقه-	-ای-	-ب-	ن
۱	(۱.۷۴)(۱..)	.۱	.۱	۱
۳	(۲.۳۰)(۱.۵	.۵	.۵	۲
۵	(۲.۸۸)۲..	۱..	۱..	۳
۷	(۳.۴۵)۲.۵	۱.۵	۱.۵	۴
۹	(۴.۰۳)۳..	۲..	۲..	۵
۱۱	(۴.۶۱)۳.۵	۲.۵	۲.۵	۷۵
۱۳	(۵.۱۸)۴..	۳..	۳..	۱۰
۱۵	(۵.۷۵)۴.۵	۳.۵	۳.۵	۱۵
۱۷	(۶.۳۳)۵..	۴..	۴..	۲..

میانگین تغییر برابر (-) نشانگرهای زیستی موجود در طبقه بندی را می‌توان محاسبه کرد.

از ۱۲۴-۱۲۰ به عنوان تابعی از $\bar{x} = \frac{\sum x_i}{n}$ ما مقدار اثرات اندازه نمونه آموزشی (ن)، اندازه مجموعه ویژگی (ص)، تنوع زیستی (-ای)، تغییرات تجربی (-ای)، اندازه اثر (-)، تکرار (ر) و همبستگی (-) بین متغیرها. ما مجموعه ویژگی های با اندازه های ۵، ۲۵، ۵۰، ۷۵ و ۱۰۰ را در نظر می‌گیریم. این باید یک محدوده معقول برای درک الگوهای تأثیر اندازه مجموعه ویژگی ها بر عملکرد طبقه بندی باشد. اگرچه داده های با توان عملیاتی بالا می‌توانند متغیرهای بسیار بیشتری داشته باشند، هدف معمول در طبقه بندی با چنین داده هایی، دستیابی به عملکرد پیش‌بینی کلاس خوب با کمترین تعداد ممکن ویژگی ها، به نام تعداد بهینه، خواهد بود.^۴ عامل بازدارنده برای در نظر گرفتن مجموعه ویژگی های بزرگتر، هزینه محاسباتی است که برای هر متغیر اضافی در مجموعه ویژگی ها به طور قابل توجه افزایش می‌یابد. مقادیر سایر عوامل در نظر گرفته شده در شبیه سازی ما در جدول ۱ خلاصه شده است. ما به طور تصادفی انتخاب من کنیم

حداکثر و دقیقه از توزیع های یکنواخت یو $\bar{x} = \frac{\sum x_i}{n}$ ، $\bar{x} = \frac{1}{n} \sum x_i$ برای هر مجموعه داده ون شانگرهای زیستی به ترتیب نزولی اندازه اثر اضافه می‌شوند. تمام شبیه سازی ها ۵۰۰ بار تکرار شدند و میانگین نرخ خطاروی این ۵۰۰ مجموعه داده برای هر ترکیب از سطوح عوامل محاسبه شد.

۲.۵.۲ داده های غیر نرمال

ما بک تحیل حساسیت انجام می‌دهیم تا پیامدهای احتمالی انحراف از داده های توزیع نرمال را بررسی کنیم. ما برای این تحیل حساسیت، داده هایی از خانوارده پواسون را شبیه سازی کردیم. داده های پواسون از چندین جهت با همتای گاووس خود متفاوت هستند: (i) پواسون یک توزیع گسسته است که داده های عدد صحیح را به جای داده های مقیاس بندی شده فاصله ای (شبیه گاووس) تولید می‌کند، (ii) رابطه میانگین-واریانس برای داده های پواسون محدودتر از داده های گاووس است زیرا آنها با پارامتر نرخ شدت (λ) یکسان هستند و (iii) داده های پواسون عموماً دارای چولگی مثبت هستند، به خصوص برای مقادیر کوچکتر. تطبیق دقیق مشاهده می‌شود که داده های پواسون با همتایان گاووس آنها امکان پذیر نیست، اما ما مشاهده (i) را حفظ کردیم. هاب-توپلیتز(اساختار کوواریانس قطری بلوکی) بین متغیرها را مانند داده های گاووس در نظر گرفت و شبیه سازی را در ترکیب های مختلفی از اندازه های نمونه آموزشی و مجموعه ویژگی ها انجام داد. (ن۱۰۰، ۱۰۰، ۷۵، ۵۰، ۴۰، ۳۰، ۲۰، ۱۰). ما تنظیم می‌کنیم ۴% برای شبیه سازی یک تکرار واحد از هر متغیر و معرفی تفاوت گروهی سیستماتیک (B) به همان روش که برای

داده های گاووسی. انتخاب به عنوان مصالحه ای بین عدم تقارن و تغییرپذیری داده ها انجام شد: ۴۱٪ به اندازه کافی بزرگ است تا تغییرپذیری معقول داده ها را تضمین کند و در عین حال به اندازه کافی کوچک است تا چولگی توزیع را تضمین کند.

۳.۱ نتایج و بحث

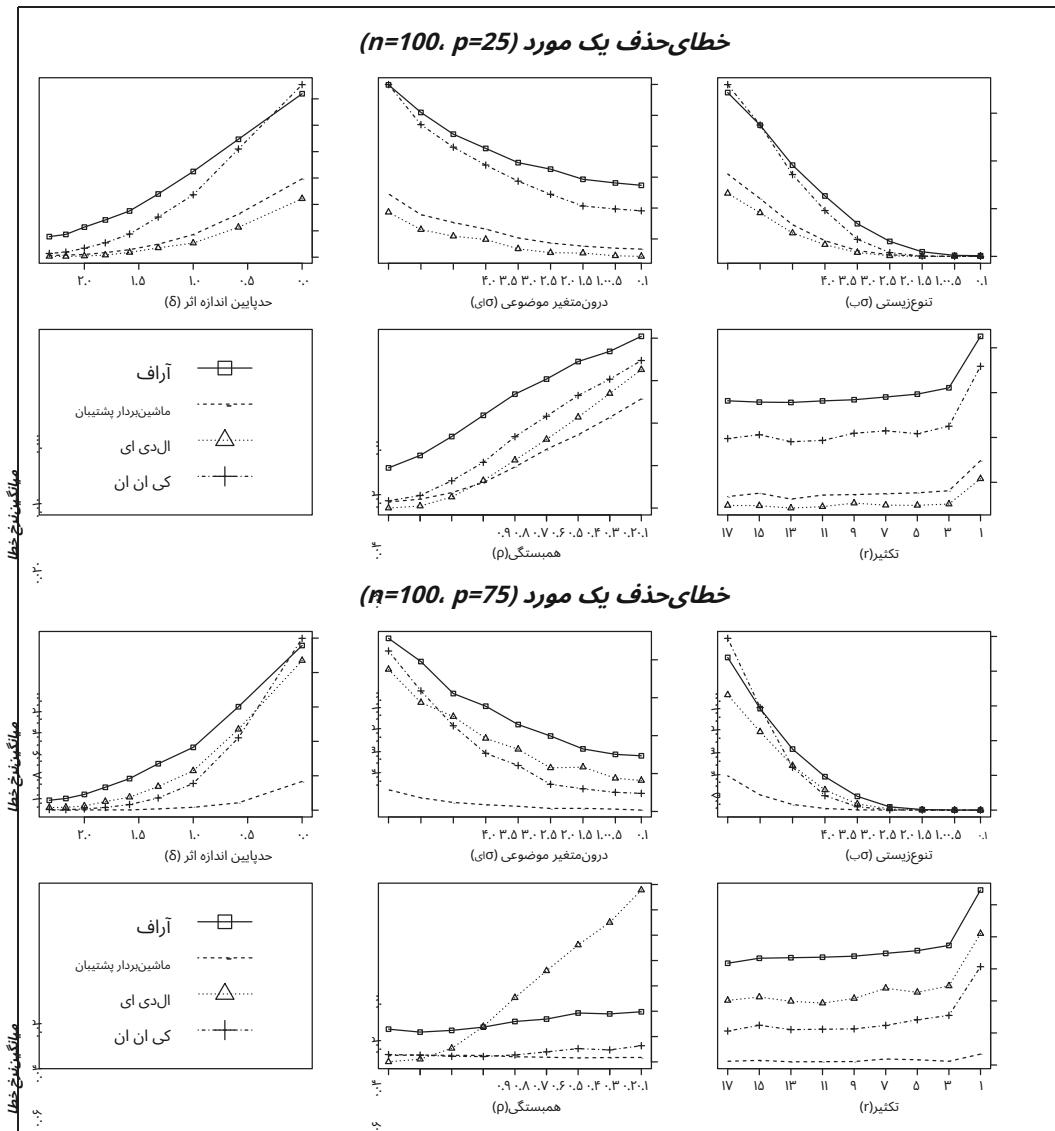
۳.۱ نتایج شبیه سازی اصلی

میانگین تخمین خطای طبقه بندي با استفاده از اعتبارسنجی متقطع با حذف یک باره در 500 تکرار از مجموعه داده های شبیه سازی شده برای هر چهار روش kNN، RF، SVM، LDA در مقابل مقادیر ویژگی های مختلف داده ها در شکل های ۱ و ۲ رسم شده است. پبل های بالا و پایین در شکل ۱ بر اساس () هستند. نتایج در شکل ۲ نمایش داده شده است. نتایج در شکل ۱ و نمودارهای اضافی در فایل تکمیلی (شکل S2) نشان می دهد که شرایط عملکرد بهینه برای LDA، که در این صورت از تمام روش های دیگر مورد مطالعه در اینجا بهتر عمل می کند. اندازه مجموعه ویژگی کوچکتر (نسبت به اندازه نمونه آموزشی) و همبستگی بالاتر است. به نظر من رسن ناحیه قدرت این روش $p=0.55$: (تعداد ویژگی ها کوچکتر از تقریباً نصف اندازه نمونه) و همبستگی بالاتر از حد متوسط (۰.۶۴): بین ویژگی ها با بزرگتر شدن مجموعه ویژگی ها ($p=0.5$ ارا برای leave-one-out دارد. پبل پایین شکل ۱ میانگین خطای RF و kNN پیش می گیرد و همچنین عملکرد بهتری نسبت به از LDA (۰.۵٪)، یعنی $p=0.05$ نسبت به از ۰.۴٪: نشان می دهد که SVM با اختلاف قابل توجه از سایر روش ها بهتر عمل می کند. اختلاف عملکرد در تغییرپذیری داده های بالاتر، اندازه اثر کوچکتر و همبستگی کمتر، بیشتر است. عملکرد kNN همچنین با افزایش اندازه مجموعه ویژگی های بیش از LDA و RF بهتر عمل می کند، مگر اینکه تغییرپذیری داده ها خیلی زیاد / یا اندازه اثر خیلی کوچک باشد. مشخص شد که RF فقط در برخی موارد که داده ها متغیرتر و اندازه اثر کوچکتر دارند، از kNN بهتر عمل می کند، که در این موارد، تخمین های خطای پایدارتری نسبت به kNN و LDA نیز ارائه می دهد (به شکل تکمیلی S3 مراجعه کنید).

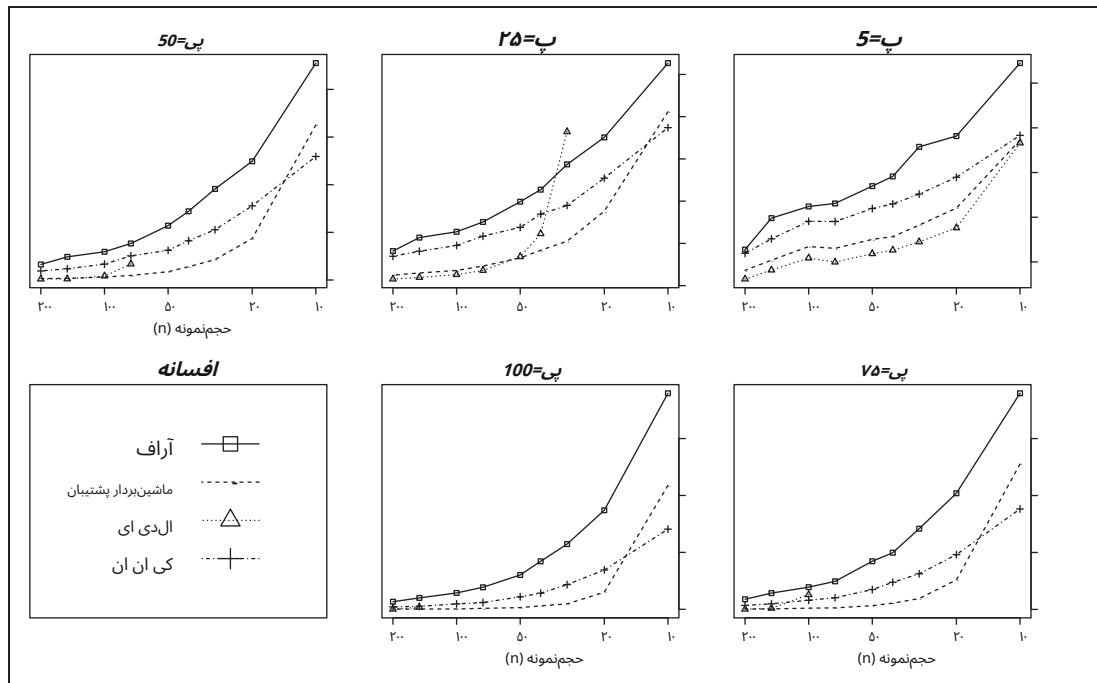
نمودارهای مربوطه برای حساسیت و ویژگی در شکل های تکمیلی S4 و S5 ارائه شده اند. عملکرد تمام روش های موردمطالعه از نظر حساسیت و ویژگی متقاضان بود، که انتظار می رود برای داده های توزیع شده متعادل و متقارن نیز صادق باشد. الگوهای ترتیب عملکرد از نظر حساسیت و ویژگی مشابه خطای کلی طبقه بندي بود.

شکل 2 نرخ خطای را در برابر نه مقدار مختلف نشان می دهدن همانطور که در جدول 1 آمده است. پنج نمودار مربوط به پنج مقدار مختلف از p (اندازه مجموعه ویژگی ها): به ترتیب ۵، ۲۵، ۵۰، ۷۵ و ۱۰۰. اگرچه مشخص شد که SVM برای نمونه های آموزشی بزرگتر و مجموعه ویژگی ها، عملکرد بهتری نسبت به سایر روش ها دارد (شکل ۱ را ببینید)، اما این روش برای نمونه های آموزشی کوچکتر (۰.۵٪) نمونه. نمودار نشان می دهد که اندازه نمونه صرف نظر از تعداد متغیرها باید حداقل ۲۰ باشد (ص) برای اینکه SVM عملکرد بهتری داشته باشد. نرخ خطای برای LDA برای اندازه های نمونه کوچکتر از تعداد متغیرها نشان داده نشده است (ص) از آنجایی که این روش برای ... منحط است (ص). اگرچه LDA از نظر تئوری برای هر چیزی معتبر است (ص)، وقتی اندازه مجموعه ویژگی ها بسیار نزدیک به اندازه نمونه آموزشی باشد، می تواند عملکرد ضعیف داشته باشد. نمودار نشان می دهد که نباید حداقل به بزرگ ۲ باشد ص برای اینکه LDA عملکرد قابل مقایسه ای داشته باشد. این یافته با یافته های قبلی ما سازگار است.

برای درک بهتر الگوهای عملکرد طبقه بندي و سهولت مقایسه های بصری، نمودارهای سه بعدی میانگین خطای اعتبارسنجی متقطع با حذف یک مورد خاص به عنوان تابعی مشترک از اندازه مجموعه ویژگی ها و تنوع زیستی در شکل 3 نمایش داده شده اند. همه نمودارها مربوط به نه ۱۰۰٪ و اندازه مجموعه ویژگی (ص) بین ۵ تا ۱۰۰٪ متغیر است. شکل، نرخ خطای را به عنوان تابعی مشترک از (-) نشان می دهد. (ص) و نشان می دهد که با افزایش اندازه مجموعه ویژگی ها، نرخ خطای کاهش می یابد (ص) برای همه روش ها به جز LDA رشد می کند. همانطور که قبل از روش دو بعدی دیده شد



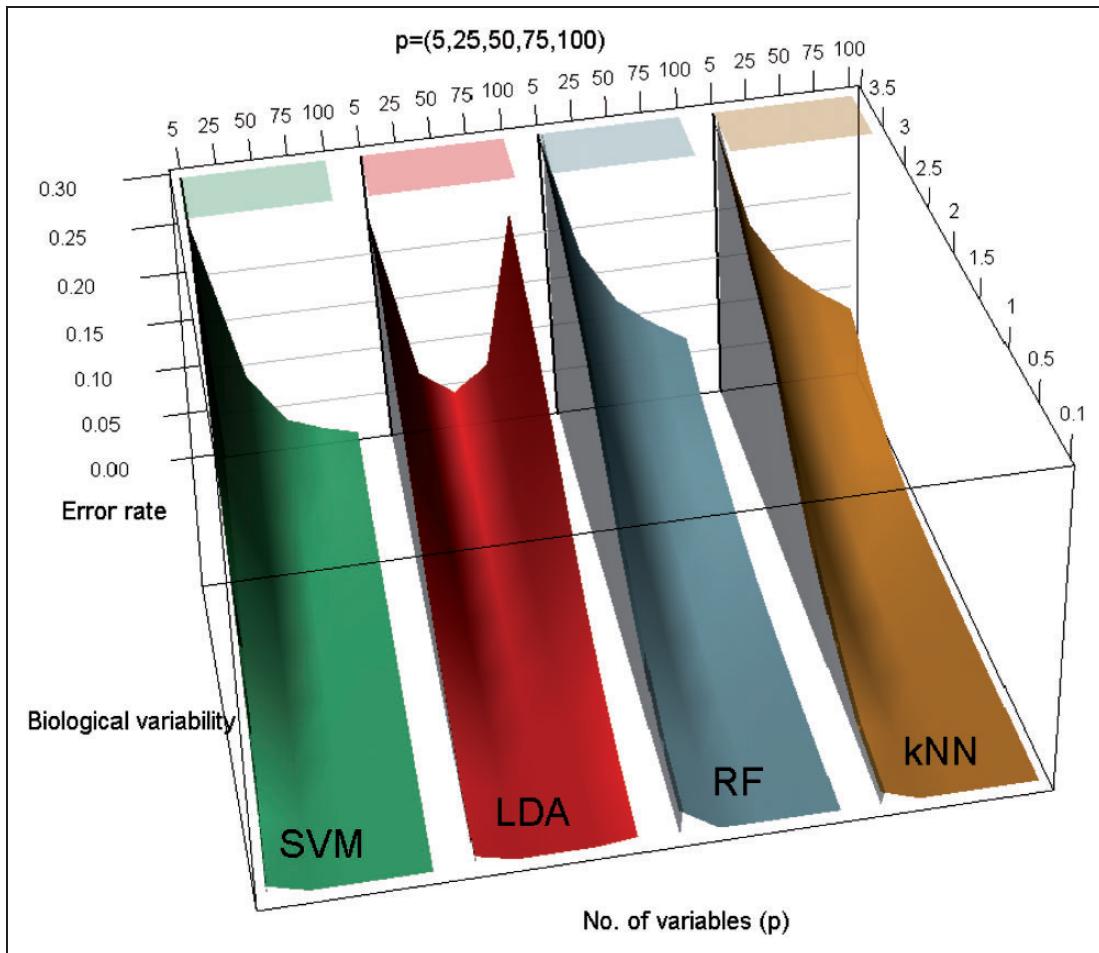
شکل ۱. میانگین خطای اعتبارسنجی متقابل با حذف یک باره در سطوح مختلف ویژگی‌های داده‌های مختلف: تغییرپذیری بیوکوژنیک (ب)، تنوع زیستی (ای)، حد پایین اندازه اثر (آ)، تکرار (ر) و همبستگی بین متغیرها (د). بین‌های بالا و بایین به ترتیب مربوط به اندازه‌های مجموعه ویژگی ۲۵ و ۷۵ متغیر و یک اندازه نمونه آموزشی راچ (ن $\frac{1}{4}$ ۰۰). هر نمودار، نزخ خطا را برای چهار روش در سطوح مختلف یک پارامتر خاص، همانطور که در شکل نشان داده شده است، مقایسه می‌کند. ایکس-محور برای مقادیر داده از پارامترهای دیگر، مقادیر داده شده از مجموعه (انتخاب می‌شوند. ن $\frac{1}{4}$ ۰۰، -ب $\frac{1}{4}$ ۰۵، -آ $\frac{1}{4}$ ۰۱، -د $\frac{1}{4}$ ۰۳)، ساختار همبستگی پارامترهای دیگر، مقادیر داده شده از مجموعه (انتخاب می‌شوند. ن $\frac{1}{4}$ ۰۰، -ب $\frac{1}{4}$ ۰۵، -آ $\frac{1}{4}$ ۰۱، -د $\frac{1}{4}$ ۰۳)، معنادارکنند. برای مجموعه ویژگی های کوچکتر (ص $\frac{1}{4}$ ۰۶۴) و همبستگی بالاتر (۰:۶۴)، نظر من رسید LDA در تمام سطوح ویژگی‌های داده ای مورد بررسی، به طور یکنواخت عملکرد بهتری داشته است. SVM برای مجموعه ویژگی های بزرگتر، از تمام روش‌های دیگر از جمله LDA بهتر عمل کرد (ض $\frac{1}{4}$ ۰۷۵) و kNN در این سطح از اندازه مجموعه ویژگی، دومین الگوریتم برتر شناخته شد. مشخص شد که RF در بیشتر سناریوهای عملکرد ضعیفی دارد، به جز داده‌های بسیار نویزی (دارای تغییرپذیری بالا و اندازه اثر کوچکتر) که در این موارد عملکرد بهتری نسبت به kNN دارد.



شکل ۲. میانگین خطای اعتبارسنجی مقایبل با حذف یک باره در سطوح مختلف نمونه آموزشی و اندازه مجموعه ویژگی‌ها. نرخ خطا در برابر نه مقدار مختلف رسم شده استن‌همانطور که در جدول ۱ آمده است. پنج نمودار مربوط به پنج مقدار مختلف از P (اندازه مجموعه ویژگی): به ترتیب ۵، ۵۰، ۲۵ و ۱۰۰. سایر پارامترها روی مقداری ثابت تنظیم شده اند: ($-P/4$ ، $-2.5/4$ ، $-1.5/4$ ، $-1/4$ ، $-0.5/4$). در قیمت $P=4$ ، $P=3$ ، $P=2$ ، $P=1$ را ببینید (تخصیص داده شد، اگرچه عملکرد SVM برای نمونه‌های آموزشی بزرگتر و مجموعه ویژگی‌ها بهتر از سایر روش‌ها (شکل ۱ را ببینید) تشخیص داده شد، اما این روش برای نمونه‌های آموزشی کوچکتر (ن=۲۰۰) نمونه، نمودار نشان می‌دهد که اندازه نمونه باید حداقل ۲۰ باشد تا SVM عملکردبتر خود را داشته باشد. نرخ خطا برای LDA برای اندازه‌های نمونه کوچکتر از تعداد متغیرها نشان داده است (ص). از آنجایی که این روش برای ... منحط است تصمیم اگرچه LDA از نظر تئوری برای هر چیزی معتبر استن α ، وقتی اندازه مجموعه ویژگی هابسیار نزدیک به اندازه نمونه آموزشی باشد، می‌تواند عملکرد بسیار ضعیفی داشته باشد. نمودار نشان می‌دهد که نیاز به حداقل به بزرگی ۲ باشد صریح اینکه LDA عملکرد قابل مقایسه‌ای داشته باشد.

با توجه به نمودارها، عملکرد LDA با افزایش تعداد متغیرها شروع به کاهش می‌کند. ص. تقریباً از نصف حجم نمونه تجاوز می‌کند (ن=۲). با این حال، LDA برای مقادیر کوچکتر () از SVM بهتر عمل می‌کند. ص. ۵۰٪ ویژگی بر اساس نسبت حجم نمونه تنظیم شده و برعکس.

خطاهای استاندارد (SE) تخمین‌های خطای اعتبارسنجی مقاطعه با حذف یک باره در سطوح مختلف ویژگی‌های داده‌های مختلف در شکل تکمیلی S3 رسم شده اند. به نظر می‌رسد SE‌های تخمین‌های خطای از الگوهای بسیار مشابه با تخمین‌های خطای میانگین پیروی می‌کنند. به عنوان مثال، SVM پایدارترین (کمترین SE) تخمین‌های خطای حذف یک باره را در موقعیت‌های ارائه می‌دهد که مجموعه ویژگی‌ها بزرگتر از نصف اندازه نمونه همان‌جایی آموزشی باشند ($P=n=4$)، در حالی که مشخص شد LDA دقیق ترین تخمین‌ها از نرخ خطا را برای $P=n=5$ همبستگی‌های 5% و بالاتر بین ویژگی‌ها. با این حال، به نظر می‌رسد LDA برای نسبت اندازه مجموعه ویژگی‌ها بزرگتر به اندازه نمونه () کمترین پایداری را دارد. $P=n=4$. قدرت RF در موقعیت‌هایی که داده‌ها متغیرتر و اندازه‌اش را کوچکتری دارند، قابل مشاهده است، که در این موارد، تخمین‌های خطای پایدارتری نسبت به LDA و KNN ارائه می‌دهد.



شکل ۳. نمودار سه بعدی میانگین خطای اعتبارسنجی متقطع با حذف یک متغیر به عنوان تابعی مشترک از اندازه مجموعه ویژگی‌ها (تعداد متغیرها) و تنواع زیستی. همه نمودارها مریبوط بهن ۰.۱۴٪ و پنج اندازه مجموعه ویژگی مختلف (ص ۵، ۲۵، ۵۰، ۷۵، ۱۰۰) همانطور که در سربرگ نمودارنامه داده شده است. پارامترهای باقیمانده همانطور که در نمودارهای دو بعدی نشان داده شده است، روی مقادیر ثابت تنظیم شده اند. نمودارها نشان می‌دهند که نرخ خطای افزایش اندازه مجموعه ویژگی‌ها (کاهش می‌یابد. ص) برای همه روش‌ها به جز LDA افزایش می‌یابد. همانطور که قبله در نمودارهای دو بعدی مشاهده شد، نرخ خطای برای LDA با افزایش تعداد متغیرها شروع به افزایش می‌کند. ص فراتر از $n=2$. اگرچه SVM برای مجموعه ویژگی‌های بزرگتر از LDA بهتر عمل می‌کند، اما LDA برای مجموعه ویژگی‌های کوچکتر بهتر از SVM عمل می‌کند (ص ۲۰۵) ...

۱.۱.۳.۱ نتایج روی داده‌های غیر نرمال این بخش نتایج شبیه سازی بر اساس توزیع‌های غیر نرمال (پواسون) را همانطور که در بخش شبیه سازی توپیچ داده شده است، ارائه می‌دهد. شکل تکمیلی S6 میانگین تخمین خطای طبقه بندی با روش اعتبارسنجی متقطع یک مرحله‌ای را در 500 تکرار داده‌های شبیه سازی شده برای هر چهار روش (kNN, RF, SVM, LDA) در ترکیبات مختلف مجموعه ویژگی‌ها و اندازه نمونه آموزشی نشان می‌دهد. الگوهای و ترتیب عملکرد برای داده‌های پواسون بسیار شبیه به آنچه برای داده‌های گاووسی مشاهده کردیم (شکل ۲) است. به عنوان مثال، مانند داده‌های گاووسی، SVM برای نمونه‌های کوچکتر عملکرد خوبی ندارد (ن ۲۰۵) اگرچه برای مجموعه ویژگی‌های بزرگتر، از سایر روش‌ها بهتر عمل می‌کند

نمونه های بزرگتر و اگرچه LDA از لحاظ تئوری می تواند اندازه مجموعه ویژگی ها را به بزرگی اندازه نمونه (پ.ن). به نظر من رسید عملکرد آن با افزایش اندازه مجموعه ویژگی ها (ص) تقریباً از نصف اندازه نمونه فراتر می رود (مثلًاً، به نموادر مراجعة کنید) (ص ۲۵٪). از بین چهار روش مقایسه شده، تنها LDA مبتنی بر فرض نرمال بودن است، اما بقیه به هیچ فرض توزیعی متکی نیستند. این نتایج نشان می دهد که LDA در برابر برخی انحرافات از نرمال بودن مقاوم است و سایر روش ها روی داده های نرمال و غیرنرمال عملکرد مشابه دارند.

۳.۲ نتایج روی داده های واقعی

داده های دنیای واقعی عموماً از نظر الگوها و منابع تغییرپذیری بسیار پیچیده تر از داده های شبیه سازی شده هستند. برای اینکه بتوانیم نتیجه گیری ها را تعمیم دهیم، برسی این موضوع مهم است که آیا یافته های حاصل از داده های شبیه سازی شده توسط یافته های حاصل از داده های واقعی پشتیبانی می شوند یا خیر. بنابراین، ما عملکرد روش ها را روی چندین مجموعه داده واقعی تولید شده از فناوری های مختلف با توان عملیاتی بالا مانند داده های بیان ژن از ریزآرایه های DNA، داده های تصویربرداری عصبی از سیستم MRI با وضوح بالا و داده های ERP که فعالیت مغز را از سیستم EEG اندازه گیری می کنند، مقایسه می کنیم. مجموعه داده های نمونه ما مدتی از مطالعات در تحقیقات سلامت روان هستند، اما یافته ها باید به هر بیماری یا وضعیت دیگری قابل تعمیم باشند.

۳.۲.۱ داده های بیان ژن دوقطبی

این مجموعه داده بر اساس مطالعه بیان ژن ریزآرایه است.^{۳۱} بافت مغز پس از مرگ بزرگسالان (قشر پیش پیشانی پشتی-جانی) از افراد مبتلا به اختلال دوقطبی و افراد سالم. جزئیات بیشتر را می توانید در بخش روش ها ببایدید. ما ۲۵ نشانگر برتر را بر اساس ... انتخاب من کنیم. پ-مقادیر آزمون بیز تجربی^{۳۲} برای بیان افتراقی بین گروه های دوقطبی و کنترل برای تحلیل طبقه بندی. این انتخاب به منظور جلوگیری از متفاوتی های که فقط نویز هستند و هیچ سیگنال تمایز کلاسی ندارند، انجام شده است. همین رویکرد برای سایر مجموعه داده های نمونه ای که در این مقاالت استفاده کرده ایم، استفاده می شود. خلاصه ویژگی های داده های بیان ژن دوقطبی بر اساس نشانگرهای انتخاب شده در جدول ۲ آورده شده است. خطاهای اعتبارسنجی متقطع تخمینی leave-one-out برای طبقه بندی دوقطبی در مقابل گروه کنترل بر اساس ۲۵ نشانگر برتر در شکل ۴ نمایش داده شده است.

اندازه های اثر برای این مجموعه داده چندان بزرگ نیستند و به نظر من رسید عملکرد طبقه بندی با افزایش تعداد نشانگرها از حدود ۵ بهبود یابد. با این حال، از نظر عملکرد مقایسه ای، SVM در بیشتر موارد بهتر (و پس از آن LDA) به نظر من رسید که سازگار است.

جدول ۲. خلاصه ای از ویژگی های داده های تخمینی برای زیرمجموعه ای از ۲۵ نشانگر برتر از داده های بیان ژن دوقطبی [31].

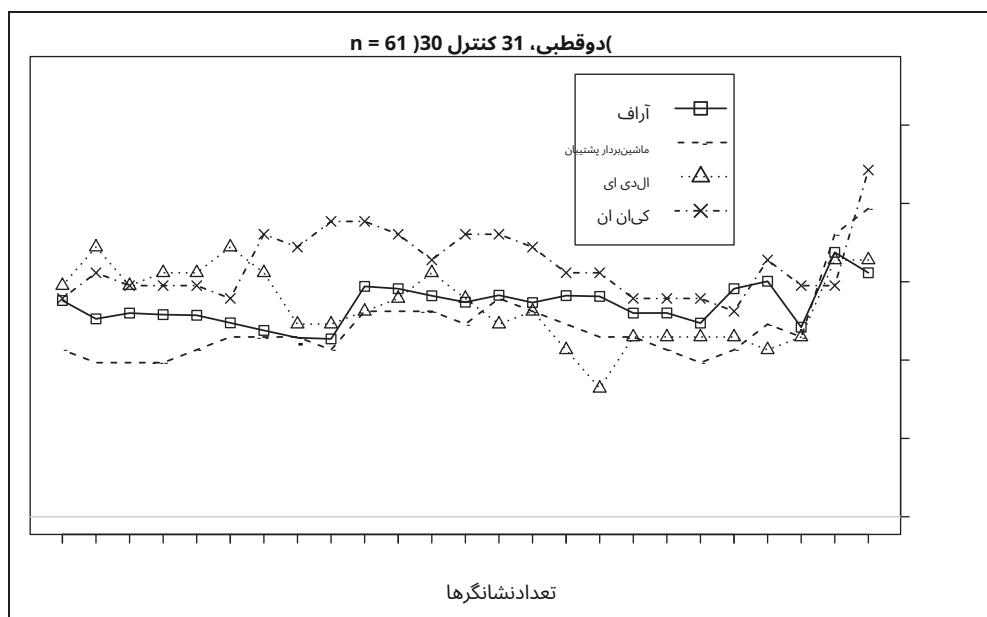
گروه	اندازه گیری	حداقل	میانگین	پرشکی	مکس
دوقطبی	(-) نوع*	.۱۴	.۳۶	.۳۲	.۹۱
کنترل	(+) همبستگی (واریاسیون)	.۶۸	.۲۷	.۴۰	.۹۰
ترکیبی	(-) همبستگی (واریاسیون)	.۱۴	.۲۷	.۲۱	.۱۹
دوقطبی در مقابل کنترل	(+) همبستگی (نسیتاکاربمی (جی جی))	.۱۶	.۳۴	.۳۰	.۸۵

* نوع زیستی (-) و نوع فنی (+) / تجربی (ffifi-ffififi) اففیفی برای این داده ها نمی توان cfffffafffffffinfffffin را جداگانه تخمین زد زیرا هیچ وجود ندارد.

همانند سازی (۱/۴) جدول، تغییرپذیری ترکیبی را نشان می دهد.^{۳۳-۳۵} بـ^{۳۶} ای.

با یافته های حاصل از داده های شبیه سازی شده، لازم به ذکر است که عملکرد LDA زمانی که تعداد نشانگرها از ۱۸ فراتر رفت، که بسیار کمتر از آستانه است، شروع به بدتر شدن کرد. (تعداد=230) از مطالعات شبیه سازی پیش بینی شده است. این ممکن است به دلیل اندازه اثر کوچکتر نشانگرها باشد که به راحتی تحت تأثیر نویز قرار می گیرند، که همچنین با این واقعیت که RF همانطور که با داده های شبیه سازی شده از این نوع یافتیم، عملکرد بهتری نسبت به kNN داشت، پشتیبانی من شود.

۱۳.۲.۲ داده های تصویربرداری مغز بیماران آزمایمرو گروه کنترل پیشرفت های اخیر در فناوری های تصویربرداری عصبی مانند سیستم MRI با وضوح بالا، اندازه گیری مؤثر ضخامت قشرمنطقه ای مغز و حجم منطقه ای را با استفاده از طرح تقسیم بندی خودکار تصویر عصبی مبتنی بر اطلس، امکان پذیرساخته است. چنین معیارهایی معمولاً برای ارتباط دادن بیماری آزمایمرو به تغییرات فیزیکی در مناطق مختلف مغز، مانند هیپوکامپ و قشر انترورینال، استفاده می شوند. در این مطالعه، داده های تصویربرداری مغز از مطالعه ADNI مستقر در ایالات متحده (www.loni.ucla.edu/ADNI) را برای طبقه بندی بیماران بر اساس τ_7 ضخامت و حجم قشر منطقه ای آنها در نظر می گیریم. اطلاعات بیشتر در مورد داده ها و پروتکل های تولید چنین معیارهایی را می توان در بخش روش یافت. ما 25 معیار برتر را بر اساس نمونه ای از 418 بیمار (186 بیمار آزمایمرو، 222 بیمار گروه کنترل) انتخاب می کنیم. پ مقادیر آزمون بیز تجربی^{۱۰} برای تفاوت میانگین ها بین گروه های AD و کنترل (CTR) برای تجزیه و تحلیل طبقه بندی. خلاصه ویژگی های داده های تصویربرداری بر اساس نشانگرها^{۱۱} انتخاب شده در جدول ۳ آورده شده است. عملکرد طبقه بندی چهار روش (kNN و RF، SVM، LDA) برای طبقه بندی AD در مقابل CTR در شکل ۵ نمایش داده شده است. اندازه های اثر () برای این مجموعه داده بالاتر است و عملکرد روش های سیار نزدیک به یکدیگر به نظر می رسد. این مورد انتظار است و توسط یافته های شبیه سازی (به عنوان مثال، شکل ۱) پشتیبانی می شود که نشان می دهد تفاوت های عملکرد با افزایش سن کمتر می شود.



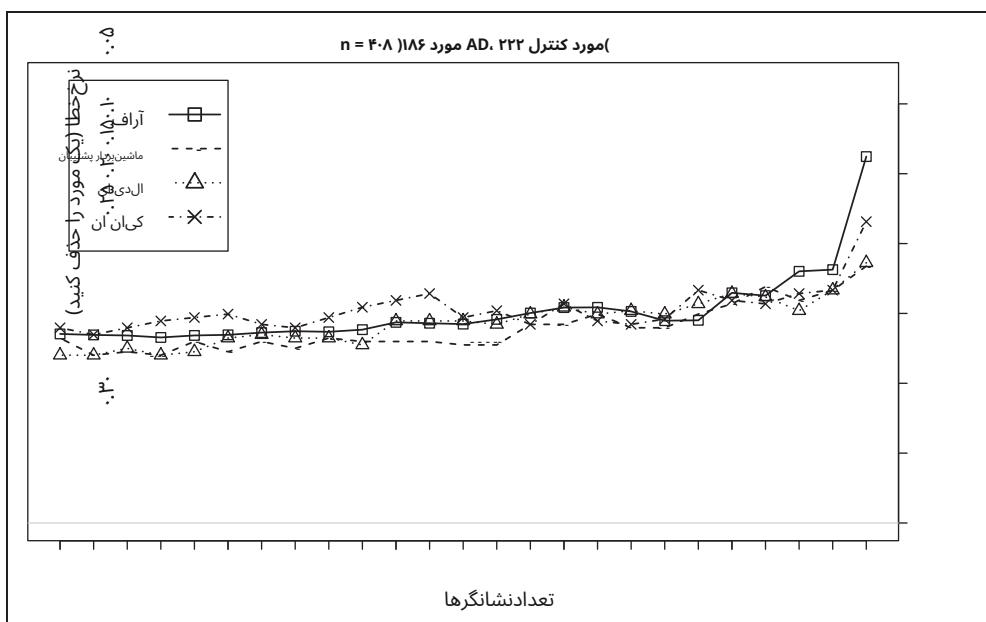
شکل ۴. خطاهای حذف یک مورد از CV برای طبقه بندی ۶۱ بیمار به گروه های دوقطبی و کنترل بر اساس ژن های با بیان متفاوت برتر از داده های بیان ژن دوقطبی.^{۱۱}

جدول ۳. خلاصه ای از ویژگی های داده های تخمینی برای زیرمجموعه ای از ۲۵ نشانگر برتر تصویربرداری مغز از مطالعه ADNI.

خلاصه تخمین ها					
گروه	اندازه گیری	حداقل	میانگین	پژشک	مکس
آگهی	تنوع*(+)	.۷۸	.۱۰	.۱۰	.۱۷
کنترل	همبستگی(-) واریاسیون(+)	.۰۵۴	.۰۶	.۰۲۱	.۰۷۸
ترکیبی	همبستگی(+) واریاسیون(-)	.۰۵۲	.۰۱۴	.۰۱۳	.۰۷۳
در مقابله کنترل AD	همبستگی(+)	.۰۵۱	.۰۱۵	.۰۱۸	.۰۷۳
	جیتی	.۰۷۶	.۰۹۷	.۰۹۲	.۱۲۶

* به باور قی جدول ۲ مفاجعه کنید.

۲۵



شکل ۵. خطاهای حذف یک مرحله ای CV برای طبقه بندی AD و کنترل بر اساس معیارهای تصویربرداری برتر. تفاوت های بسیار کمی بین عملکرد روش ها وجود دارد، که به دلیل اندازه های اثر بزرگتر مورد انتظار است و توسط یافته های حاصل از شبیه سازی پشتیبانی می شود که نشان می دهد تفاوت های عملکرد با افزایش اندازه اثر کمتر می شود.

اندازه اثر افزایش می یابد. در مجموع، به نظر می رسد SVM و LDA در بیشتر موارد عملکرد بهتری داشته اند، همانطور که الگوهای کلی مشاهده شده در شبیه سازی نیز آن را تأیید می کنند.

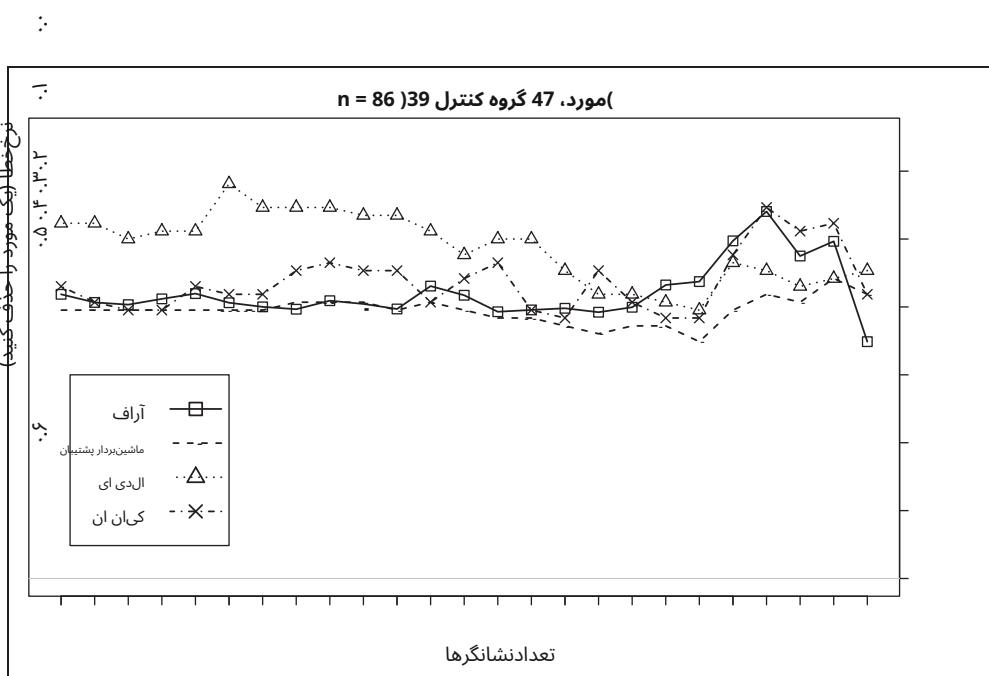
داده های الکتروانسفالوگرافی EEG نوسانات ولتاژ ثبت شده از الکترودهای روی پوست سر را اندازه گیری می کنند و شاخصی از فعالیت مغز را ارائه می دهند. داده های EEG برای این مثال از ۴۱ بزرگسال با تشخیص فعلی به دست آمده است.

جدول ۴. خلاصه ای از ویژگی‌های داده‌های تخمینی برای زیرمجموعه‌ای از ۲۵ ویژگی برتر از داده‌های EEG.

خلاصه تخمین‌ها					
مکس	پزشکی	میانگین	حداکثر	اندازه‌گیری	گروه
۱.۱۲	.۹۳	.۹۳	.۷۴	تنوع*(-)	اختلال کم توجه بیش فعال
.۹۶	.۰۴	.۰۷	.۹۰	همبستگی()	
۱.۱۷	۱.۰۶	۱.۰۵	.۸۸	واریاسیون(-)	
.۹۷	.۰۴	.۱۱	.۹۱	همبستگی()	
۱.۰۰	.۹۹	.۹۹	.۹۸	واریاسیون(-)	ترکیبی
.۹۵	.۰۳	.۰۹	.۸۷	همبستگی()	
.۰۴۵	.۲۷	.۰۲۸	.۱۸	جنس	
در مقابله کنترل ADHD					

اختلال کم توجه بیش فعال

* به پاور F جدول ۲ مراجعه کنید. Δ



شکل ۵. خطاهای CV حذف یک طرفه برای طبقه بنده ۸۶ بیمار به گروه‌های ADHD و کنترل بر اساس معیارهای برتر EEG. این مثالی است که در آن اندازه اثر بسیار کوچک است و بنا برای نرخ خطا با افزایش تعداد متغیرها، الگوی کاهشی ثابتی را نشان نمی‌دهد. با این حال، این موضوع از مشاهده شبیه سازی در مورد نقاط قوت RF در برخورد با داده‌هایی با سیگنال ضعیف تر و نویز بالاتر پیشیابی می‌کند، زیرا در این مثال، kNN از RF و LDA عمل می‌کند.

مبتلای ADHD (اختلال بیش فعال کمبود توجه) و ۴۷ نفر بدون مشکلات سلامت روان، توضیحات بیشتر داده‌ها در بخش روش آمده است. از بین ۶۳ معیار، تنها تعداد کمی از آنها بین گروه‌های مورد و کنترل تفاوت معنی داری داشتند. با این حال، ما ۲۵ معیار برتر را بر اساس ... انتخاب می‌کنیم. پ-مقادیر آزمون بیز تجربی^{۱۴} برای تحلیل طبقه بنده

خلاصه ویژگی های داده های EEG بر اساس معیارهای انتخاب شده روي آزمودنی (دو مورد ADHD به دليل مقادیراز دست رفته در برجی از متغیرهای انتخاب شده حذف شدند) در جدول ۴ آمده است. عملکرد طبقه بندي چهار روش(kNN، RF، SVM، LDA) برای اختلال کم توجهی بیش فعالی در مقابل کنترل طبقه بندي در شکل ۶ نمایش داده شده است.

ازین سه مجموعه داده مثال، داده های EEG کوچکترین سیگنال (میانگین) را دارندجی جی ۰.۲۸٪. برای تمايز قائل شدن بین گروه های مورد و کنترل. به ندرت اندازه گیری های EEG وجود دارد که تفاوت معناداری بین گروه ها داشته باشد. این احتمالاً توضیح می دهد که چرا منحنی های خطای افزایش تعداد ویژگی ها، کاهش ثابتی را نشان نمی دهند. با این حال، مشاهده من شود که RF از LDA و kNN بهتر عمل می کند و از مشاهدات حاصل از شبیه سازی در مورد نقاط قوت RF در برخورد با داده هایی با سیگنال ضعیف تر و نویز بالاتر پشتیبانی می کند.

۴.نتیجه گیری

مایک مطالعه شبیه سازی گسترده انجام دادیم تا عملکرد طبقه بندي تعدادی از الگوریتم های یادگیری ماشین یا آماری پرکاربردرا از نظر خطاهای تعمیم، حساسیت و ویژگی برای مسائل طبقه بندي نظارت شده، به طور عین مقایسه کنیم. مرکز اصلی مطالعه ما بررسی این بود که «کدام روش در چه شرایط عملکرد بهتری دارد» و این کار را با مقایسه عملکرد هادر ترکیبات مختلف سطوح عوامل متعدد (ویژگی های داده) انجام دادیم. نتایج مطالعه شبیه سازی ما و مثال های بعدی روی چندین مجموعه داده واقعی از پلتفرم های مختلف فناوری با توان عملیاتی بالا منجر به نتیجه گیری های زیر شد:

. برای تعداد کمتری از ویژگی های همبسته، تعداد ویژگی هایی که تقریباً از نصف اندازه نمونه تجاوز نمی کنند، LDA ازنظر میانگین خطاهای تعمیم و همچنین پایداری (دقت) تخمین خطای روش انتخابی شناخته شد. به نظر من رسد نایخیه قدرت LDA عبارت است از: $p=0.55$: (تعداد ویژگی ها کوچکتر از تقریباً نصف اندازه نمونه) و همبستگی (بالاتر ۰.۶۴): بین ویژگی ها.

. همانطورکه مجموعه ویژگی ها بزرگتر من شود ($p=0.5$): پیش می گیرد و همچنین LDA از RBF با هسته SVM با اختلاف فاحشی بهتر از RF و kNN عمل می کند. اختلاف عملکرد در تغییرپذیری داده های بالاتر، اندازه اثر کوچکتر و همبستگی کمتر، بیشتر است. با این حال، اندازه نمونه صرف نظر از تعداد ویژگی ها باید حداقل 20 باشد (ص) برای اینکه SVM به عملکرد برتر خود دست یابد.

. عملکرد kNN همچنین با افزایش اندازه مجموعه ویژگی ها بهبود می یابد و از LDA و RF پیش می گیرد، مگر اینکه تغییرپذیری داده ها خیلی زیاد و/یا اندازه های اثر خیلی کوچک باشند. . مشخص شد که RF فقط در برجی موارد که داده ها متغیرتر بودند و اندازه اثر کوچکتری داشتند، از kNN بهتر عمل می کند، که در این موارد، تخمین های خطای پایدارتری نسبت به kNN و LDA نیز ارائه می دهد.

. همه روش ها تمایل به عملکرد بهتر در همبستگی بالاتر را نشان دادند، اما به نظر من رسد RF وقتی متغیرها همبستگی بسیار بالایی دارند، عملکرد نسبتاً بدتری دارد. . عملکرد تمام روش های مورد مطالعه از نظر حساسیت و ویژگی متقاضی ویژگی متقاضی بود، که انتظار من رود برای داده های توزیع شده متعادل و متقاضی نیز صادق باشد.

هیچ یک از روش های مورد مطالعه (به جز LDA) نیازی به پیروی داده ها از توزیع احتمال خاصی ندارند و نتایج شبیه سازی باید در برابر انحراف از فرض نرمال بودن مقاوم باشند. ما این را با شبیه سازی از توزیع غیر نرمال (پواسون) نشان دادیم که نشان می دهد LDA در برابر برجی انحراف از نرمال بودن مقاوم است و سایر روش ها نیز به طور مشابه روی داده های نرمال و غیر نرمال عمل می کنند.

تقدیرنامه‌ها

از پروفسور اندر و پیکلز و داوران ناشناس برای نظرات مفیدشان در مورد پیش نویس اولیه که به بهبود مقاله کمک کرد، تشکر من کنیم. از دکتر گرین مک لافلین و پروفسور فیلیپ اشرسون برای ارائه داده‌های EEG و از بخش پزشکی مسیر (دانشگاه ادینبورگ) برای داده‌های بیان ژن پایه سپاسگزاریم.

اعلام تضاد منافع

نویسنده‌گان هیچ گونه تضاد منافع بالقوه‌ای را در رابطه با تحقیق، تأثیف و/یا انتشار این مقاله اعلام نکردند.

بودجه

نویسنده‌گان دریافت حمایت مالی زیر را برای تحقیق، تأثیف و/یا انتشار این مقاله اعلام کردند: RD و AS از مرکز تحقیقات زیست پزشکی سلامت روان موسسه ملی تحقیقات سلامت (NIHR) در جنوب لندن و بنیاد NHS مادزلی و کالج کینگ لندن، حمایت مالی دریافت می‌کنند. دیدگاه‌های بیان شده متعلق به نویسنده‌گان است و لزوماً دیدگاه‌های NHS، NIHR با وزارت بهداشت نیست.

این کار با استفاده از منابع ارایه شده توسط (<http://www.ecdf.ed.ac.uk/>) ECDF و مرکز تحقیقات زیست پزشکی سلامت روان (<http://compbio.brc.iop.kcl.ac.uk/cluster/>) NIHR (<http://www.edikt.org.uk/>) eDIKT (<http://www.edikt.org.uk/>) پشتیبانی می‌شود.

جمع آوری و اشتراک گذاری داده‌های تصویربرداری مغز که در یکی از نمونه‌های این مطالعه استفاده شده است، توسط طرح ابتکاری تصویربرداری عصبی بیماری آزمایش (ADNI) (موسسات ملی بهداشت، کمک هزینه U01 AG024904) تأمین مالی شده است. ADNI توسط مؤسسه ملی پیری، مؤسسه ملی تصویربرداری زیست پزشکی و مهندسی زیستی و از طریق کمک‌های سخاوتمندانه مواد زیر تأمین مالی می‌شود: Abbott؛ انجمن آزمایش؛ بنیاد کشف داروی آزمایش؛ Amorfix Life Sciences Ltd؛ Eisai Inc؛ Bristol-Myers Squibb Company؛ Biogen Idec Inc؛ BioClinica, Inc؛ Bayer HealthCare؛ AstraZeneca Research & Development, LLC؛ IXICO Ltd؛ Innogenetics, NV؛ GE Healthcare؛ Genentech, Inc Johnson & Johnson Pharmaceutical Research & Development LLC؛ Janssen Alzheimer Immunotherapy؛ Medpace, Inc؛ شرکت مرک و شرکا؛ شرکت تشخیص بیماری مزو اسکیل؛ شرکت داروسازی نووارتبیس؛ فایزر؛ سروبر؛ سینارک؛ شرکت داروسازی تاکدا. موسسات تحقیقات سلامت کانادا بودجه‌ای را برای حمایت از مراکز بالینی ADNI در کانادا فراهم‌من کنند. کمک‌های مالی بخش خصوصی در ۷ نوامبر ۲۰۱۲ توسط بنیاد موسسات ملی بهداشت (www.fnih.org/) تسهیل می‌شود. سازمان دریافت کننده کمک هزینه، موسسه تحقیقات و آموزش کالیفرنیای شمالی است و این مطالعه توسط مطالعه مشارکتی بیماری آزمایش کالیفرنیا، سن دیگو هماهنگ شده است. داده‌های ADNI توسط آزمایشگاه تصویربرداری عصبی در دانشگاه کالیفرنیا، لس آنجلس منتشر می‌شود. این تحقیق همچنین توسط کمک‌های مالی NIH P30 AG010129 و NIH K01 AG030514 پشتیبانی شده است.

منابع

۳. من، دایسون جی، جانسون کی و همکاران. ارزیابی روش‌های طبقه‌بندی داده‌های بیان ژن. مجله بیوفارم استات. ۲۰۰۴؛ ۱۶۵: ۱۶۱-۱۸۴.

۴. استاتیکوپ آ، آلیفریس سی، تسامارینوس آی و همکاران. ارزیابی جامعه روش‌های طبقه‌بندی چنددسته ای برای تشخیص سرطان با استفاده از ریزآرایه بیان ژن. بیوانفورماتیک. ۲۰۰۵؛ ۳۱: ۶۳۳-۶۴۳.

۵. لی جی، لی جن، پارک ام و همکاران. مقایسه ای گستره از ابزارهای طبقه‌بندی اخیر که بر روی داده‌های ریزآرایه اعمال می‌شوند. تحلیل داده‌های آماری محاسباتی. ۲۰۰۵؛ ۴۸: ۸۶۹-۸۸۵.

۱. دودوبت اس، فریدلیاند جی و اسپیدت. مقایسه روش‌های تمايز برای طبقه‌بندی تومورها با استفاده از داده‌های بیان ژن. انجمن آمار آمریکا. ۲۰۱۲؛ ۷۷: ۹۷-۸۷.

۲. تان ای و گلبرت دی. مقایسه تجربی تکنیک‌های بادگیری ماشین تحت نظر از پیونافورماتیک. در: مجموعه مقالات اولین کنفرانس بیوانفورماتیک آسیا و اقیانوسیه در مورد بیوانفورماتیک. ۲۰۱۳ - جلد ۱۹، دارلینگ‌هرست، استرالیا. انجمن کامپیوت استرالیا، ۲۰۱۳، صفحات ۲۲۹-۲۲۲.

۱۶. دودا آر، او، هارت پ. ای. و استورک دی. جن. طبقه بندی الگو. نیویورک: واپلی، ۲۰۰۱.
۱۷. داساراتن پی. وی. هنجارهای نزدیکترین همسایه (NN): تکنیک های طبقه بندی الگوی NN. لوس آلامینوس، کالیفرنیا، ایالات متحده: انجمن کامپیوتر ۹۶ IEEE، ۱۹۹۶.
۱۸. شاخنازوریج ج، دارل تی و ایندیک پی. روش های نزدیکترین همسایه برای دادگیری و سیاست: نظریه و عمل کمربیج، ایالات متحده آمریکا: MIT Press، ۲۰۰۵.
۱۹. دیاز-اوریار آر و آدرس اس. انتخاب ژن و طبقه بندی داده های ریزآرایه با استفاده از جنگل تصادفی. بیوانفورماتیک BMC ۳:۷-۲۰. ۲۰۰۵.
۲۰. افرون ب. تخمین نزخ خطای یک قانون پیش بینی: بهبود اعتبارسنجی مقایل. انجمن آمار آمریکا ۱۹۸۳:۷۸:۱۳۳-۱۴۳.
۲۱. رایان ام، لاکستون اچ، هافاکر اس و همکاران. تحلیل بیان ژن اختلال دوپاکی، کاهش تنظیم چرخ یوپیکوپتین و تغییرات در ژن های سیستاپس ای نشان می دهد. روانپژوهش مول ۲۰۰۶:۵۶-۶۵:۱-۶.
۲۲. S, et al. AddNeuroMed بیماری آذینی و مطالعه MRI آندازه گیری. Simmons A, Westerm E, Muehlboeck ساینس ۲۰۰۹:۴۷:۱۱۸:۲-۹.
۲۳. سیموز ای، وستمن ای، موهلبیک اس، و همکاران. چارچوب از ۲۴ ماه اول. روانپژوهش سالمندان بین المللی ۲۰۱۰:۷۵-۷۲:۲۶-۳۱.
۲۴. وستمن ای، آگلار س، موهلبیک ج، و همکاران. معیارهای تصویربرداری شدید مغناطیسی منطقه ای برای تحلیل چند متغیره در بیماری آذینی و اختلال شناختی خفیف. توبوگرافی مغز ۲۰۱۱:۱-۱۵.
۲۵. اسکیرو س و اشرسون پ. بیانی عاطفی، همبودی و اختلال در بزرگسازان مبتلا به اختلال پیش فعالی-کمود توجه. اختلال عاطفی ۲۰۱۳:۷۲:۱۱-۸۶.
۲۶. مک لافلن جن، آبرشت پی، پاشفیسکی ت و همکاران. شواهد الکتروفیزیولوژی برای حالت های اماده سازی غیرطبیعی و پردازش مهاری در ADHD بزرگسالان. عملکرد مغز بهای ۲۰۱۰:۶۶:۲-۱.
۲۷. تای سن، ریچسیدیک اف، گون سی و همکاران. تأثیرات زنگنه مشترک بر علائم ADHD و فعالیت EEG با فرائنس بسیار پایین: مطالعه ای بر روی دوقلوها. مجله روانشناسی کودک، روانپژوهش ۲۰۱۳:۵۰(۴):۷۵-۷۰.
۲۸. اسپیت سی ای، دیکنیسون پی، فورست تی و همکاران. ارزیابی کمی RNA خون کامل انسان به عنوان یک نشانگر زیستی بالقوه برای بیماری های غفعون. تحلیلکرای ۲۰۱۰:۱۲۲:۱۰۰-۱۳۰.
۲۹. هاردن جن، گارسیا اس آر و گولان دی. روش برای تولید preprint arXiv:1106.5834.
۳۰. خودنودک ام، باخمن تی، میوسن ام و همکاران. تحلیل چند عاملی خطای پیش بینی کلاس: تخمین تعداد بهینه نشانگرهای زیستی برای قوانین طبقه بندی مختلف. مجله بیوانفورماتیک محسسانی ۹۴۵:۸:۳۱-۴۵.
۳۱. اسپیت جی کی. مدل های خط و روش های تجزیی بیز برای ارزیابی بیان اتفاقی در آزمایش های ریزآرایه. کاربرد آماری ژن مول بیولوژی ماده ۳:۱۰۰-۳.
۳۲. هوانگ ایکس، بن دبلیو، گریندل اس و همکاران. مطالعه تطبیقی تشخیص علت نارسایی قلبی انسان با استفاده از پروفایل های بیان ژن. بیوانفورماتیک ۲۰۰۶:۴-۵.
۳۳. دیاز-اوریار آر و دی آدرس آ. انتخاب ژن و طبقه بندی داده های ریزآرایها استفاده از جنگل تصادفی. بیوانفورماتیک BMC ۳:۷-۲۰. ۲۰۰۵.
۳۴. دوسات ن، مانگه آ، سولاسول جی و همکاران. مقایسه روش های سلطان اینفوژم ۲۰۰۵:۱۲-۱۵.
۳۵. بیرونیا م، بانگ ج، بانگ م و همکاران. مطالعه تطبیقی روش های مختلف بادگیری ماشین بر روی داده های بیان ژن ریزآرایه. زنومیک ۹:۲۰-۲۷. ۲۰۰۵.
۳۶. استاتنیکوف آ، وانگ ال و آلیفیریس سی. مقایسه جامع جنگل های تصادفی و ماشین های بردار پشتیبان برای طبقه بندی سلطان مبتنی بر ریزآرایه. بیوانفورماتیک BMC ۳:۱۹-۹:۳-۱۹.
۳۷. بولستیکس ای، استرتوبل سی، آگوستین تی و همکاران. ارزیابی طبقه بندی کنده های مبتنی بر ریزآرایه: یک مرور کلی. سلطان اینفوژم ۹:۷-۷۷:۶:۲۰-۸.
۳۸. راک دی، ایدکرت، ترویانسکاپا او و همکاران. مقاولاتی در مورد نرم افزار سازی، انتخاب متغیر، طبقه بندی یا خوش بندی داده های ریزآرایه. بیوانفورماتیک ۲۰۰۷:۱-۲:۵۱-۲۰۰-۹.
۳۹. یوسفی م، هوا ج، سیما س و همکاران. گزارش سوگیری هنگام استفاده از مجموعه داده های واقعی برای تحلیل عملکرد طبقه بندی. بیوانفورماتیک ۲۰۱۱:۱-۱:۶۸-۷۶.
۴۰. هانزار بی و دوگرت ای. در مورد مقایسه طبقه بندی کنده ها برای داده های ریزآرایه بیوانفورماتیک Curr ۲۰۱۱:۱-۱:۹۳-۹۵.
۴۱. یوسفی م، هوا ج و داگرت ای. سوگیری چند قاعده ای در مقایسه قوانین طبقه بندی. بیوانفورماتیک ۲۰۱۱:۷:۱۷-۲۰:۱-۶۸-۳.
۴۲. فیشر آر، استفاده از اندازه گیری های چندگانه در مسائل طبقه بندی آن. هیومن ۱۹۹۳:۷:۱۷-۱۸.
۴۳. هوا ج، شیونگ زد، لووی ج و همکاران. تعداد بهینه ویزگی ها به عنوان تابعی از اندازه نمونه برای قوانین طبقه بندی مختلف. بیوانفورماتیک ۲۰۰۵:۱-۱:۱۵-۹:۱۲-۲۰.
۴۴. اوپنیک وی، ان. ماهیت نظریه یادگیری آماری. نیویورک: اشپرینگر، ۱۹۹۵.
۴۵. کوکوتین سی و اوپنیک وی. شبکه های بردار پشتیبان. یادگیری ماشین ۱۹۹۵:۲-۲۷:۱-۲.
۴۶. اوپنیک وی، ان. نظریه یادگیری آماری. نیویورک: جان واپلی و پسران، ۱۹۹۸.
۴۷. اسکروندال ای، طراحی و تحلیل آزمایش های مومنت کارلو: حمله به خرد متعارف. پژوهش های رفتار چند مقیمه ۳۵:۱-۱۷.
۴۸. ولبرت دی. فقدان تمایزات پیشینی بین الگوریتم های یادگیری. محاسبات عصبی ۱۹۹۶:۱۰:۱۳۴-۱۳۹.
۴۹. بیرین ل. جنگل های تصادفی. یادگیری ماشین ۰:۲۰۵-۳۲-۵.
۵۰. فیشر آر ای. استفاده از اندازه گیری های چندگانه در مسائل طبقه بندی آن. یوگن ۱۹۹۶:۷:۱۷-۱۸.
۵۱. مک لکلن جی جن. تحلیل تشخیص و تشخیص الگوی آماری. نیویورک: ایالات متحده: انتشارات جان واپلی و پسران، ۱۹۹۲.