مطالعات علوم کاربردی در مهندسی

دوره ۳، شماره ۱، بهار ۱۳۹۶

صفحات ۲۱۸- ۲۰۸

Online ISSN: ٢٤٧٦-٤٥٠٧

Print ISSN: ٢٥٣٨-٢٠٥٥

www.irijournals.com

تکامل تفاضلی با ترکیب شبکه عصبی برای استخراج قوانین از پایگاه داده‌های پزشکی

مهدی جبه داری [۱]، حمید پایگذار [۲]

[۱] گروه کامپیوتر، دانشکده فنی و مهندسی، واحد خمین، دانشگاه آزاد اسلامی، خمین، ایران

[۲] گروه کامپیوتر، دانشکده فنی و مهندسی، واحد خمین، دانشگاه آزاد اسلامی، خمین، ایران (نویسنده مسئول)

چکیده

امروزه در علم پزشکی سیستم‌های پشتیبانی برای تصمیمات بالینی تعریف شده‌اند که با استفاده از دو یا چند آیتم از ویژگی-های بیمار، پیشنهادهای مفیدی ارائه می‌دهند. این پیشنهادهابه متخصص در تشخیص بیماری یا روند درمان کمک می‌کنند. طبقه‌بندی داده‌ها در این حوزه نیازمند شفافیت و ارائه‌ی علت برای تخصیص داده‌ها می‌باشد. ازاین‌رو لازم است علاوه بر طبقه-بندی درست داده‌های پزشکی، علت تخصیص هر داده شرح داده شود. در این تحقیق، یک روش جدید بر پایه‌ی تکامل تفاضلی برای طبقه‌بندی خودکار آیتم‌ها در پایگاه داده‌های پزشکی موردنظر است. بر اساس آن، یک ابزار به نام DEREx معرفی شده است که به‌طور خودکار، دانش واضح را از پایگاه داده، به شکلی از قوانین IF-THEN (اگر-سپس) شامل شرط‌های متصل به AND (و) بر روی متغیرهای پایگاه داده، استخراج می‌کند. هر فرد از تکامل تفاضلی یک قانون برای طبقه‌بندی محسوب می-شود. مجموعه افراد تکامل تفاضلی مجموعه‌ای از قوانین می‌باشند که از تمام آن‌ها برای طبقه‌بندی داده‌های موجود در پایگاه‌داده استفاده می‌شود. این قوانین می‌توانند به‌صورت منطقی متصل به OR (یا) نیزدر نظر گرفته شوند؛ بنابراین تمام قوانین طبقه‌بندی برای تمام دسته‌ها یک‌باره در یک مرحله پیدا می‌شوند. ابزار DEREx اولین ابزار طبقه‌بندیاست که بر پایه‌ی تکامل تفاضلی می‌باشد و به‌طور خودکار و بدون دخالت سازوکار دیگری دسته‌هایی از قوانین IF-THEN را استخراج می‌کند.

واژه‌های کلیدی: طبقه بندی، تکامل تفاضلی، پایگاه داده، استخراج قوانین.

دوره ۳، شماره ۱، بهار ۱۳۹۶، صفحات ۲۱۸- ۲۰۸

۱- مقدمه

در حال حاضر داده‌کاوی مهم‌ترین فناوری جهت بهره‌برداری مؤثر از داده‌های حجیم است و اهمیت آن رو به افزایش است. چند سال پیش تخمین زده شده بود که مقدار داده‌ها در جهان هر ۲۰ ماه به حدود دو برابرمی‌رسد. در یک تحقیق مربوط به سال ۲۰۰۴ که بر روی گروه‌های تجاری بسیار بزرگ در جمع‌آوری داده‌ها صورت گرفت مشخص گردید که ۱۹ درصد از این گروه‌ها دارای پایگاه داده‌هایی با سطح بیشتر از ۵۰ گیگابایت می‌باشند و ۵۹ درصد از آن‌ها انتظار دارند که در آینده‌ای نزدیک در چنین سطحی قرار گیرند (هلت[1]، ۱۹۹۳). لذا در چنین شرایطی که پیشرفت‌های فناوری امکان ذخیره‌سازی این حجم عظیم از داده‌ها را برای ما فراهم آورده دیگر بعید است که بتوان با روش‌های معمول و سنتی پردازش اطلاعات به تمام واقعیات نهفته در داده‌ها دست یافت. لذا داده‌کاوی برای یاری در این عرصه با روش‌ها و تکنیک‌های جدید و رو به بهبود به میدان آمده است.

بهادرا[2] و همکاران (۲۰۱۲) یک الگوریتم توسعه‌یافته از بهینه‌سازی پارامترهای الگوریتم ماشین بردار پشتیبان[3] (SVM) با استفاده از تکامل تفاضلی ارائه داده‌اند که دقت طبقه‌بندی را نسبت به استفاده انفرادی از SVM افزایش می‌دهد. اخیراً SVM به‌عنوان تکنیکی مؤثر برای حل مسائل طبقه‌بندی‌های واقعی، توجه زیادی را به خود جلب نموده است. به‌هرحال عملکرد SVM حساس به انتخاب تابع اساسی مناسب و مقادیر پارامترها به کمک تابع اساسی است.

جایانث[4] و همکاران (۲۰۱۵)، داده‌های گرفته شده توسط ماهواره‌های فضایی را با استفاده از یک الگوریتم کلونی زنبور مصنوعی انجام داده‌اند. علت استفاده‌ی آن‌هاازاین‌روش جدید این است که روش‌های آماری قدیمی موفقیت محدودی در طبقه‌بندی داده‌ها ارائه می‌دهند و بسیاری از جنبه‌های اطلاعات مانند همبستگی را در نظر نمی‌گیرند. الگوریتم کلونی زنبور مصنوعی[5] (ABC) یک الگوریتم مبتنی بر هوش جمعی است. این الگوریتم برای داده‌های حس شده از راه دور توسط ماهواره‌ها استفاده شده است و نتایج نشان می‌دهند که ۵٪ افزایش دقت نسبت به روش سنتی آماری یعنی طبقه‌بند درستی ماکزیمم[6] (MLC) و شبکه‌های عصبی مصنوعی و همچنین ۳٪ افزایش دقت نسبت به الگوریتم SVM دارد.

وینودهینی[7] و همکاران (۲۰۱۴) یک ارزیابی عملکرد مقایسه‌ای برای روش‌های مبتنی بر شبکه‌های عصبی مصنوعی برای طبقه‌بندی عواطف انسانی در فضاهای آنلاین ارائه داده‌اند. کمک اصلی طبقه‌بندی عواطف تعیین هویت عواطفی است که در یک پیام متنی نهفته است. روش‌های یادگیری ماشین برای طبقه‌بندی عواطف انسانی به‌طور گسترده‌ای موردمطالعه قرار گرفته‌اند ولی روش‌هایی که بر اساس شبکه‌های عصبی مصنوعی باشند به ندرت در طبقه‌بندی عواطف مورداستفاده قرار گرفته‌اند. سانینو[8] و همکاران (۲۰۱۴)، یک روش استخراج خودکار قوانین با استفاده از تکامل تفاضلی برای نظارت بر روی آپنه انسدادی خواب[9] (OSA) توسط یک سیستم موبایل ارائه داده‌اند. چن[10] و همکاران (۲۰۱۳) در مقاله‌شان روشی برای بهبود عملکرد ماشین بردار پشتیبان دوقلوی لاپلاسی[11] (Lap-TSVM) با استفاده از تکامل تفاضلی ارائه داده‌اند. ماشین بردار پشتیبان دوقلوی لاپلاسی اخیر تولید خوبی با حل یک جفت مسئله‌ی برنامه‌ریزی خطی (QPP) حاصل می‌کند. به‌هرحال فرایند آموزش Lap-TSVM وابسته به زمان است. علاوه بر آن درمقایسه با SVM، Lap-TSVM پارامترهای بیشتری برای تعدیل نیاز دارد که بر روی برنامه‌های عملی آن تأثیر می‌گذارد.

---

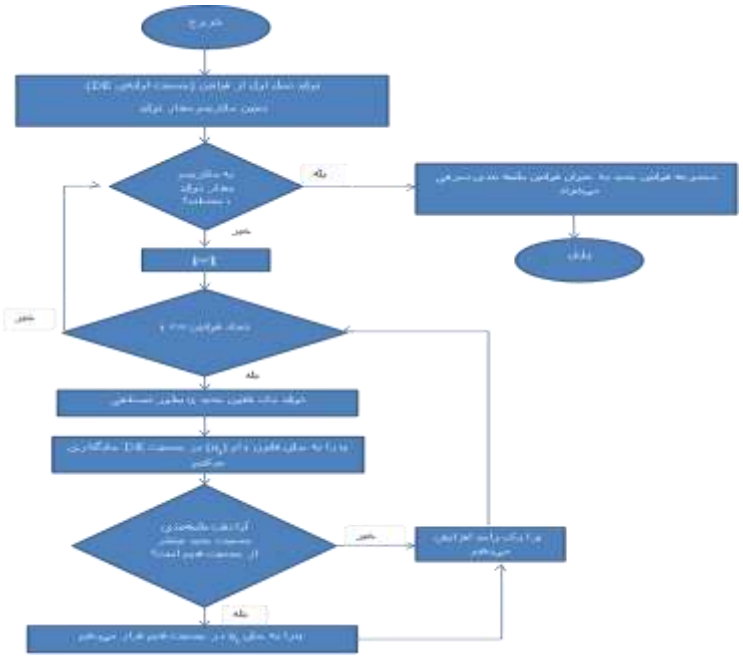
[1] Holte
[2]- Bhadra
[3]- Support Vector Machine
[4]- Jayanth
[5]- Artificial Bee Clony
[6]- Maximum Legitimacy Classifier
[7]- Vinodhini
[8]- Sannino
[9]- occlusion Sleep Apne
[10]- Chen
[11]- Laplasian Twin Support Vector Machine

دوره ۳، شماره ۱، بهار ۱۳۹۶، صفحات ۲۱۸- ۲۰۸

وو کای[1] از الگوریتم DE برای مشخص کردن وزن‌های خصیصه‌ها در پایگاه‌داده استفاده کرده‌اند. سپس‌ازآن وزن‌ها در بیز بی‌دقت وزن‌دار[2] (WNB) گسترش داده‌شده‌یقبلی‌شان استفاده کرده‌اند. به‌عنوان نتیجه، آن‌ها یک پیکربندی جدید به نام DE-WNB تعریف نمودند. کای[3] و همکاران یک برنامه مبتنی بر تکامل برای اجرای طبقه‌بندیداده‌هایابر طیفی ارائه داده است. برنامه آن‌ها بر اساس روش‌های کلاسیک طبقه‌بندی مانند Naïve Bayes،NBtree و J48 است درحالی‌که DE برای انتخاب نوارهای ابر طیفی استفاده شده است. داس[4] و همکاران را برای طبقه‌بندی مجموعه داده‌های بزرگ و بدون برچسب اجرا کرده‌اند. در مقابل بسیاری از تکنیک‌هایطبقه‌بندی موجود، الگوریتم پیشنهادشده نه‌تنها نیازی به دانش اولیه برای طبقه‌بندی ندارد بلکه تعداد پارتیشن‌های بهینه از داده‌ها را در اجرا مشخص می‌کند. داس[5] و همکاران را برای طبقه‌بندی مجموعه داده‌های بزرگ و بدون برچسب اجرا کرده‌اند. در مقابل بسیاری از تکنیک‌هایطبقه‌بندی موجود، الگوریتم پیشنهادشده نه‌تنها نیازی به دانش اولیه برای طبقه‌بندی ندارد بلکه تعداد پارتیشن‌های بهینه از داده‌ها را در اجرا مشخص می‌کند.

۲- روش انجام

ایده‌ی اصلی روش حل مسئله در پژوهش این است که هر فرد از DE یک قانون را نشان می‌دهدکه در اصل می‌تواند اشاره به یکی از کلاس‌های مختلف داشته باشد؛ بنابراین یک مجموعه از قوانین، افراد جمعیت DE می‌باشند که در طول تولیدهای آینده اصلاح می‌شوند. ویژگی این روش این است که برای هر کلاس بیش از یک قانون می‌تواند یافت شود، به شرطی که الگوریتم تشخیص دهد این عمل کارایی طبقه‌بندی را بهبود می‌بخشد. علاوه بر این اینجا تمام قوانین طبقه‌بندییک‌باره در یک مرحله به دست می‌آیند؛ که این روش با روش‌های دیگر متفاوت است. شکل ۱ روند کلی الگوریتم را نشان می‌دهد.



شکل ۱. فلوچارت الگوریتم پیشنهادی

[1]- WuandCai
[2]- Weighted Naïve Bayes
[3]- Cai
[4]- Das
[5]- Das

دوره ۳، شماره ۱، بهار ۱۳۹۶، صفحات ۲۱۸- ۲۰۸

۲-۱ فرایند کار DEREx

با ترکیب تمام اطلاعات بالا یک پایگاه‌داده داده شده با N_V متغیر، DEREx به‌عنوان شکل کلی نشان داده شده در شبه کد الگوریتم DE کار می‌کند. قبل از پرداختن به جزییات DEREx لازم است توجه شود که این ابزار می‌تواند فوراً خصیصه‌های پایگاه‌داده را که با مقادیر حقیقی نشان داده شده‌اند، بررسی کند. در این فرایند یک کدگذاری ساده برای خصیصه‌های بولی، بی قید و صحیح باید بررسی شود. یک خصیصه‌ی صحیح در یک محدوده‌ی صحیح برای مثال [10,35]، با یک مقدار حقیقی در محدوده‌ی مقادیر حقیقی [10.0,35.0] کدگذاری می‌شود. یک خصیصه‌ی بی‌قاعده می‌تواندبه‌عنوان یک مورد صحیح کنترل شود. یک خصیصه‌ی بی‌قاعده می‌تواند مثل یک خصیصه‌ی صحیح که یک رابطه‌ی سفارشی را تأمین کرده، کنترل شود و می‌تواند توسط مقادیر احتمالی‌اش تعریف شود. ممکن است نمونه‌ای از یک خصیصه مثل سن = {جوان، میان‌سال، پیر} به‌صورت سن={۱ و ۲ و ۳} کنترل شود. یک خصیصه‌ی بولی در [0,1]نیز به‌صورت عددی حقیقی در [0.0,1.0] کدگذاری می‌شود. سپس مقادیر پارامترهای DEREx تعیین می‌شوند. این در میان دیگران اشاره دارد به بیشترین تعداد قوانین مطلوب یعنی N_R. در اینجا با دادن پایگاه‌داده‌ای با N_V متغیر، DEREx به‌طور تصادفی جمعیتی از N_{pop} راه‌حل ممکن مقداردهی اولیه می‌کند. هر مقدار از یک توزیع تصادفی یکنواخت که دارای کران بالا و پایین تعیین شده‌ایست انتخاب می‌شود. در پایان اجرا قوانین پیداشده با استفاده از مکانیسم اصلی شرح داده شده به‌طور اتوماتیک کدگذاری می‌شوند. آن مکانیسم اصلی کدگذاری خصیصه‌های پایگاه‌داده مقدار حقیقی را شرح می‌دهد. برای مقادیر صحیح، مقادیر بی‌قاعده و خصیصه‌های بولی پایگاه‌داده در قوانین باید یک فرایند کدگذاری ثانویه و سرراست با شروع از مقدار حقیقی تأمین شده توسط DEREx اجرا شود. برای تمام این‌گونه متغیرها این عمل بدین معناست که مقدار حقیقی را به نزدیک‌ترین مقدار صحیح گرد کنیم. برای مثال برای یک خصیصه‌ی صحیح، یک مقدار ۳۱.۵۷ که توسط DEREx تولید شده به شکل ۳۲ کدگذاری می‌شود. درحالی‌که مقدار حقیقی ۳۱.۲۴ به شکل ۳۱ کدگذاری می‌شود. به‌طور مشابه برای خصیصه‌یبی‌قاعده‌ی سن، یک مقدار ۲.۱۴ به عدد ۲ گرد می‌شود که به مقدار بی‌قاعده‌ی*میان‌سال* اشاره دارد. درنهایت برای یک خصیصه‌ی بولی مقدار حقیقی ۰.۱۴ به‌صورت صفر کدگذاری می‌شوددرحالی‌که مقدار ۰.۷۴ به‌عنوان ۱ کدگذاری می‌شود.

خلاصه‌ی DEREx در این مراحل قابل ذکر است:

۱. ایجاد مجموعه‌ای از قوانین به‌صورت تصادفی نام مجموعه را $pop=\{x_1,x_2,\ldots x_{Nr}\}$میگذاریم.

۲. برای هر قانون (x_i) از مجموعه‌ی pop روند زیر اجرا می‌شود:

- یک قانون جدید (x'_i) با یکی از استراتژی‌های تکامل تفاضلی (به‌عنوان مثال، تصادفی) ایجاد می‌شود.
- قانون x' را به‌جای x وارد مجموعه‌ی pop می‌کنیم و دقت طبقه‌بندی با استفاده از کل مجموعه را اندازه می‌گیریم. اگر دقت طبقه‌بندی افزایش یافت، قانون x'_i را به‌جای x_i قرار داده و قانون x_i را از مجموعه حذف می‌کنیم. درغیر اینصورت قانون x_i در مجموعه باقی می‌ماند.

۳. مرحله‌ی دوم برای چندین بار (مثلاً ۱۰۰ دور) تکرار می‌شود. به عبارتی مجموعه‌ی pop چندین بار اصلاح می‌شود تا بهترین قوانین برای طبقه‌بندی به وجود بیایند.

٤. قوانین استخراجی از مرحله‌ی سوم به‌عنوان قوانین نهایی معرفی شده و ازین پس از این قوانین برای طبقه‌بندی داده‌های جدید استفاده خواهد شد.

۲-۲ آموزش شبکه عصبی مصنوعی با تکامل تفاضلی

همان‌طور که پیش‌تر گفته شد از شبکه‌های عصبی مصنوعی نیز می‌توان برای طبقه‌بندی داده‌ها استفاده نمود. شبکه‌های عصبی مصنوعی همانند دیگر روش‌های هوش مصنوعی نیازمند آموزش می‌باشند. داده‌هایی که برای ارزیابی صلاحیت در روش پیشنهادی با تکامل تفاضلی داده می‌شود، به شبکه عصبی مصنوعی نیز داده می‌شود و شبکه با تنظیم وزن‌ها و بایاس‌ها آموزش می‌بیند. در این قسمت برخلاف روش معمول، قصد داریم شبکه‌های عصبی مصنوعی را توسط تکامل تفاضلی

دوره ۳، شماره ۱، بهار ۱۳۹۶، صفحات ۲۱۸- ۲۰۸

پیشنهادی آموزش داده و طبقه‌بندی را انجام دهیم. برای ارائه روش پیشنهادی یک شبکه عصبی چندلایه پرسپترون[1] پیشخور[2] در نظر گرفته شده است.

خروجی شبکه‌های عصبی پیشخور یک تابع از وزن‌های پیوندگاهی W و مقادیر ورودی x می‌باشد؛ یعنی y=f(x,W). در پردازش‌های استاندارد آموزش، هم بردار ورودی x و هم بردار خروجی y شناخته شده هستند و بردار وزن W برای دست‌یابی به یک مسیردهی درست برای رسیدن ورودی x به خروجی y تنظیم می‌شود. به صورت عمومی وزن‌ها می‌توانند با حداقل-سازی تابع خطای شبکه E به دست آیند. تابع خطای شبکه به صورت زیر قابل تعریف است (ایلونن و همکاران[3]، ۲۰۰۳):

$$E\big(y,f(x,W)\big):(y^{D1},x^{D2},W^{D3},f)\to R. \qquad (۱)$$

هدف بهینه‌سازی حداقل‌سازی تابع هدف E توسط بهینه‌سازی مقادیر وزن‌های شبکه است.

$$W=(w_1,\dots,w_D) \qquad (۲)$$

مشابه دیگر الگوریتم‌های تکاملی، تکامل تفاضلی مبتنی بر یک جمعیت (P_G) از راه‌حل‌های کاندیدا کار می‌کند. این راه‌حل‌های کاندیدا افراد جمعیت می‌باشند. تکامل تفاضلی یک جمعیت با اندازه‌ی ثابت NP، بردارهایی با مقادیر حقیقی، $W_{i,G}$ که i اندیس جمعیت و G شماره‌ی تولیدی را نشان می‌دهد که جمعیت به آن متعلق است.

$$P_G=\big(w_{1,G},\dots,w_{NP,G}\big),\ G=0,\dots,G_{max} \qquad (۳)$$

علاوه بر این در آموزش شبکه‌ی عصبی مصنوعی هر بردار شامل D وزن شبکه است (کروموزوم‌های افراد):

$$W_{i,G}=\big(w_{1,i,G},\dots,w_{D,i,G}\big),\ i=1,\dots,NP,\qquad G=0,\dots,G_{max} \qquad (۴)$$

طرح تولید مجدد جمعیت تکامل تفاضلی با دیگر روش‌های تکاملی متفاوت است. پس از مقداردهی اولیه‌ی جمعیت نخست، بردارها در جمعیت جاری یعنی P_G به طور تصادفی نمونه‌برداری و ترکیب می‌شوند تا بردارهای کاندیدا را برای نسل بعدی یعنی P_{G+1} تولید کنند. جمعیت کاندیداها یا بردارهای آزمایشیِ $P'_{G+1}=U_{i,G+1}=u_{j,I,G+1}$ به صورت زیر تولید می‌شوند:

$$v_{j,i,G+1}=w_{j,r3,G}+F.\big(w_{j,r1,G}-w_{j,r2,G}\big)$$
$$u_{j,i,G+1}=\begin{cases}v_{j,i,G+1}, & if\ rand_j[0,1]\le CR\\ w_{j,i,G}, & otherwise\end{cases} \qquad (۵)$$

به طوریکه:

$$i=1,\dots,NP,\quad j=1,\dots,D$$
$$r_1,r_2,r_3\in[1,\dots,NP],\text{randomly},r_1\neq r_2\neq r_3\neq i \qquad (۶)$$
$$CR\in[0,1],\qquad F\in(0,1].$$

و CR عبارت است از مقدار حقیقی فاکتور گذردهی که احتمال پارامترهای بردار آزمایشی را کنترل می‌کند. عموماً مقادیر F و CR سرعت همگرایی و قدرت فرایند جستجو را تحت تأثیر قرار می‌دهند. مقادیر بهینه‌ی آن‌ها بستگی به ویژگی‌های تابع هدف و اندازه‌ی جمعیت، NP دارد.

طرح انتخاب تکامل تفاضلی نیز متفاوت از دیگر روش‌هاست. جمعیت برای تولید بعدی، P_{G+1} از جمعیت جاری P_G به صورت زیر انتخاب می‌شود:

$$W_{i,G+1}=\begin{cases}U_{i,G+1}, & if\ E(y,f(x,W_{i,G+1})\le E(y,f(x,W_{i,G}))\\ W_{i,G}, & otherwise\end{cases} \qquad (۷)$$

[1]- Multi Layer Perceptron Neural Network
[2]- Feed Forward
[3] Ilonen

بنابراین هر فرد از جمعیت موقت، با همتای خود در جمعیت جاری مقایسه می‌شود. با فرض حداقل‌سازی تابع هدف، برداری که مقدار کمتری برای تابع هدف ایجاد کند در مقایسه پیروز شده و در جمعیت قرار می‌گیرد؛ بنابراین افراد در جمعیت بعدی یا به خوبی جمعیت جاری هستند، یا از آن‌ها بهتر می‌باشند. نکته جالب در طرح جایگزینی DE این است که یک بردار آزمایشی فقط با یک فرد مقایسه می‌شود نه تمام افراد. این نکته از این بابت ارزشمند است که طرح جایگزینی اطمینان می-دهد که جمعیت از بهترین پاسخ‌ها واگرا نمی‌شود (ایلونن و همکاران، ۲۰۰۳).

۴- نتایج

۴-۱ پایگاه‌داده‌های مورداستفاده

در این تحقیق از سه پایگاه داده استاندارد پزشکی برای بررسی روش پیشنهادی استفاده شده است. این هشت پایگاه داده عبارت‌اند از:

۱. دیتابیس هابرمن شامل مواردی روی بقای بیماری است که برای سرطان سینه تحت جراحی است. این داده‌ها توسط دانشگاهِ بیمارستان بیلینگ شیکاگو جمع‌آوری شده است. هر آیتمی توسط سه مقدار روشن صحیح نشان داده شده که به ترتیب شامل: سن بیمار هنگام عمل، سال عمل بیمار (به‌صورت: سال منهای ۱۹۰۰) و تعداد گره‌هایگوشه‌ای مثبت پیداشده، می‌شوند. هر آیتم به کلاس صفر اختصاص می‌یابد به شرطی که بیمار ۵ سال یا بیشتر زنده بماند و اگر بیمار در کمتر از ۵ سال فوت شود به کلاس یک اختصاص می‌یابد (هابرمن[1]، ۱۹۷۶).

۲. دیتابیس بریست[2] موارد سرطان سینه را کهتوسط دکتر ویلیام اچ وولبرگ از بیمارستان دانشگاه ویسکانسین[3] جمع‌آوری شده، نشان می‌دهد. کلاس صفر شامل موارد خوش‌خیم و کلاس یک شامل موارد بدخیم است (ولبرگ[4]، ۱۹۹۰).

۳. دیتابیس کلورلند[5] ویژگی‌های بیماران با مرض قلبی را از مرکز پزشکی V.A، Long Beach و موسسه‌ی بالینی کلورلند لیست می‌کند. آیتم‌های آن هم می‌تواند به صورت ۲ کلاسه و هم ۵ کلاسه طبقه‌بندی شوند. در این پژوهش مورد پنج کلاسهمورداستفاده قرار گرفته است. هدف، یافتن وجود بیماری قلبی در بیمار است. کلاس‌هابه‌صورت افزایش خطر از صفر (بدون بیماری) تا چهار (بیشترین اعلام‌خطر) طبقه‌بندیشده‌اند (دترانو و همکاران[6]، ۱۹۸۹).

۴-۲ نتایج اجرای DEREx

الگوریتم برای هر سه پایگاه داده اجرا شده و نتایج با الگوریتم‌های طبقه‌بندی KNN و NB مقایسه گردیده که به شرح زیر میباشند:

۴-۲-۱ پایگاه‌داده Breast

برای اجرای DERExروی این پایگاه‌داده مقادیر CT، RT و CR هر سه را برابر ۰.۵ در نظر گرفته و تعداد قوانین تولیدی را ۱۰ در نظر گرفته‌ایم. همچنین تعداد دورهای تولید را ۱۰۰ دور قرار داده‌ایم.

قوانیناستخراجشده پس از اجرا به‌صورت زیر می‌باشند.

if 1 && 1 && 1 && 1 && A5>=0.29804 && in(A6,1.5961,2.6723) && 1 && in(A8,2.1922,4.3447) && 1 then CLASS= 1;

[1] Haberman
[2] Breast
[3] Wisconsin
[4] Wolberg
[5] Cleverland
[6] Detrano et al.

if A1<15.5653 && A2==0.82261 && A3>-0.43475 && out(A4,-0.12868,13.7069) && A5==-0.088695 && A6<=0.82261 && A7>-0.088695 && 1 && A9==-0.088695 then CLASS= 1;

if 1 && 1 && in(A3,24.5059,77.5088) && 1 && A5<=0.41012 && A6<=1.8202 && out(A7,0.41012,1.4702) && 1 && A9>0.41012 then CLASS= 1;

if out(A1,32.275,67.9236) && 1 && 1 && A4>7.892 && 1 && A6>1.491 && out(A7,0.2455,0.95847) && in(A8,1.982,4.8339) && 1 then CLASS= 1;

if 1 && A2>1.3633 && A3>13.0828 && in(A4,6.3597,24.7892) && 1 && A6>1.3633 && 1 && A8>1.7266 && 1 then CLASS= 1;
else CLASS=0;

این پایگاه داده ۹ متغیرودوکلاس (۰ و ۱) دارد. هر متغیر یکه در شرط لحاظ نشده به جای آن ۱ قرار گرفته است که در کل شرط بی تأثیر است.

۴-۲-۲ پایگاه‌داده هابرمن

برای اجرای DEREx روی این پایگاه‌داده مقادیر CT، RT و CR هر سه را برابر ۰.۵ در نظر گرفته و تعداد قوانین تولیدی را ۴ در نظر گرفته‌ایم. همچنین تعداد دورهای تولید را ۱۰۰ دور قرار داده‌ایم.

قوانین استخراج شده:

if 1&& in(A2,61.8457,66.2149) && A3>18.1797 then CLASS= 1;
else CLASS=0;

این پایگاه‌داده شامل سه متغیر و دو کلاس (۰ و ۱) است.

۴-۲-۳ پایگاه‌داده کلورلند

برای اجرای DEREx روی این پایگاه‌داده مقادیر CT، RT و CR هر سه را برابر ۰.۵ در نظر گرفته و تعداد قوانین تولیدی را ۱۵ در نظر گرفته‌ایم. همچنین تعداد دورهای تولید را ۱۰۰ دور قرار داده‌ایم.

قوانین استخراج شده:

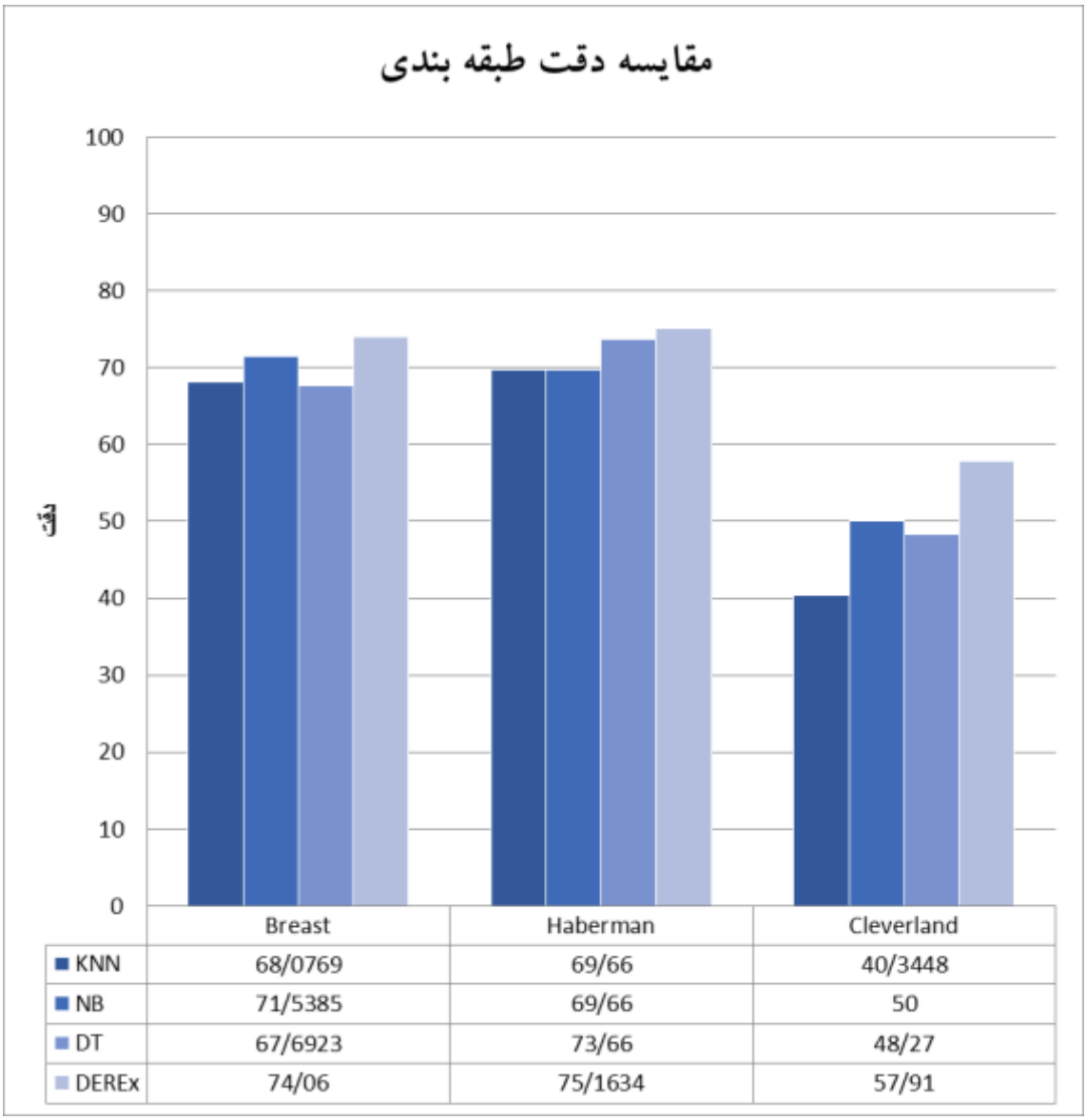
if out(A1,35.1223,56.3106) && A2==0.12755 && 1 && 1 && 1 && 1 && 1 && 1 && A10>=0.79079 && out(A11,1.2551,2.1379) && 1 && A13<=3.5102 then CLASS= 1;
if A1>=46.3485 && out(A2,0.36143,0.40984) && 1 && out(A4,132.3114,137.4434) && A5>=284.3055 && A6<=0.36143 && 1 && A8<118.3471 && 1 && in(A10,2.2409,2.541) && 1 && out(A12,1.0843,1.2295) && 1 then CLASS= 1;
if 1 && 1 && A3>1.4445 && 1 && A5==190.8943 && 1 && 1 && A8>90.409 && A9>=0.14816 && 1 && A11<=1.2963 && out(A12,0.44448,2.3476) && A13>=3.5926 then CLASS= 1;
if 1 && 1 && 1 && 1 && A5==191.9729 && 1 && A7==0.30125 && 1 && 1 && 1 && A11<1.3012 && 1 && 1 then CLASS= 2;
if 1 && 1 && A3>1.7791 && 1 && 1 && A6>0.25972 && 1 && in(A8,105.0227,157.0504) && 1 && 1 && 1 && 1 && A13>=4.0389 then CLASS= 3;
if 1 && A2>=0.37964 && 1 && out(A4,134.2421,142.0223) && 1 && 1 && 1 && out(A8,120.7331,130.3483) && A9<0.37964 && out(A10,2.3538,2.8089) && 1 && A12>1.1389 && A13>4.5186 then CLASS= 3;

if out(A1,47.4131,53.7349) && 1 && 1 && A4>=134.6623 && A5<=294.0196 && out(A6,0.38361,0.51531) && 1 && A8<121.2524 && 1 && 1 && 1 && A12<=1.1508 && 1 then CLASS= 3;
if A1>=31.2227 && A2>=0.046306 && out(A3,1.1389,3.5945) && 1 && 1 && A6<=0.046306 && A7>0.092611 && 1 && out(A9,0.046306,0.86482) && 1 && 1 && 1 && A13>3.1852 then CLASS= 4;
else CLASS=0;

این پایگاه‌داده نیز دارای ۱۳ متغیر و ۵ کلاس (۰ و ۱ و ۲ و ۳ و ۴) می‌باشد.

توجه شود که: نتایج بالا در هر سه پایگاه داده برای ۱۰۰ دور از تکامل تفاضلی اجرا شده است. با بیشتر کردن تعداد دورها طبیعتاً دقت افزایش خواهد یافت.

نمودار دقت طبقه‌بندی DEREx در مقایسه با سه الگوریتم دیگر برای هر سه پایگاه‌داده در شکل ۴-۱ قابل مشاهده است:



شکل ۲. مقایسه دقت طبقه‌بندی DEREx با الگوریتم‌های دیگر

دوره ۳، شماره ۱، بهار ۱۳۹۶، صفحات ۲۱۸- ۲۰۸

۵- جمع‌بندی

وظایف طبقه‌بندی در حوزه‌ی پزشکی با اشاره به فرایند درمان داری اهمیت هستند. در این حوزه و در بسیاری از حوزه‌های دیگر، یک ویژگی خوش‌آیند طبقه‌بندی کننده این است که برای کاربران دانسته‌های واضحی از پایگاه داده به‌طور اتوماتیک استخراج می‌کند. البته دانش به‌دست‌آمده هرگز نباید جانشین کار کارشناسانه شود، بلکه باید به‌عنوان پشتیبانی برای تصمیم‌گیری در نظر گرفته شود.

در این پژوهش، یک روش جدید مبتنی بر تکامل تفاضلی برای طبقه‌بندی خودکار آیتم‌های پایگاه‌داده پیشنهاد شده است. بر اساس آن ابزاری که DEREx نام دارد معرفی شده است که به‌طور خودکار دانش واضحی را از پایگاه‌دادهبه‌صورت قوانین IF-THEN متصل به شرط‌های AND روی متغیرها استخراج می‌کند. نتیجه برای مجموعه‌ای از قوانین کدگذاری می‌شود و این قوانین می‌توانند به‌صورت متصل به OR منطقی نیز دیده شوند. بعلاوه تمام قوانین طبقه‌بندی‌شده برای تمام کلاس‌ها یک‌باره در یک مرحله به دست می‌آیند.

بحث اصلی این پژوهش این است که DEREx اولین ابزار طبقه‌بندی است که بر اساس DE است و به‌طور خودکار مجموعه‌ای از قوانین IF-THEN را بدون مداخله‌ی مکانیسم دیگری استخراج می‌کند.

برای آزمایش‌ها سه پایگاه‌داده از حوزه‌ی پزشکی انتخاب شده است. ابزار در سرتاسر اینسه پایگاه داده با یک دسته قوانین طبقه‌بندی‌کننده استفاده شده است. نتایج،تأثیر روش موردنظر را ثابت کرده و تصدیق می‌کنند که DEREx در مقایسه با دیگر الگوریتم‌ها بهترین ابزار طبقه‌بندی است.

مزیت اصلی DEREx تأمین اطلاعات واضحِ استخراج شده از پایگاه‌داده برای کاربر است؛ زیرا برخلاف RBF، Naïve Bayes، Bagging و NBtree می‌تواند به‌طور مستقیم قوانین IF-THEN را برای تشخیص ارائه کند. علاوه بر این DEREx می‌تواند استخراج ویژگی‌ها را نیز انجام دهد. چون قوانین به‌دست‌آمده شامل تعدادی از متغیرها می‌باشند که می‌توانند برای تشخیص درست بیماری بسیار مهم باشند. در این روش با استفاده از اطلاعات مفید به پزشکان کمک شده است. البته نظر آن‌ها درباره‌ی درستی و مفید بودن این قوانین بیشترین اهمیت را در عمل دارد. وقتی‌که DEREx با دیگر ابزار طبقه‌بندی مبتنی بر قانون مقایسه می‌شود، DEREx نیازمند کمترین میانگین از قوانین برای روبرو شدن با مسئله است که این امتیاز قابل‌توجهی نسبت به دیگر تکنیک‌هاست. علاوه بر این میانگین تعداد شرط‌ها در قوانین زیاد بزرگ نیست. این ویژگی DEREx را کاربرپسند می‌کند.

لازم به یادآوری است که این نتیجه باوجود استفاده از AND به‌عنوان تنها متصل‌کننده‌ی محلی در شرط‌ها به‌دست‌آمده است. این نقص با این واقعیت که سیستم قادر است مجموعه‌ای از قوانین و درصورت لزوم برای هر کلاس بیش از یک قانون داشته باشدتا حد زیادی برطرف شده است؛ زیرا قوانین می‌توانند با استفاده از OR در نظر گرفته شوند.

یک توضیح ممکن برای کیفیت نتایج به‌دست‌آمده این است که در اصل روش ارائه شده در اینجا تکاملی است. لذا تمام مزایای الگوریتم‌های تکاملی را داراست:

- می‌تواند بدون دانش اولیهیا هیچ‌گونه فرضی وارد مسئله شود.
- می‌تواند مقادیر بهینه‌ی محلی را خارج کند.
- نیازی نیست تابع صلاحیت قابل‌تشخیص باشد

تا اینجا این ویژگی‌ها برای تمام روش‌های مبتنی بر تکامل در روبرو شدن با مسائل متداول ند. مخصوصاً ابزار معرفی شده که بر اساس تکامل تفاضلی است که در نوشته‌ها به‌طور گسترده به‌عنوان روشی سریع‌تر از دیگر روش‌ها در جستجوی محلی شناخته می‌شود.

منابع

1. G. Vinodhini, R.M. Chandrasekaran,A comparative performance evaluation of neural network based approach for sentiment classification of online reviews, Journal of King Saud University – Computer and Information Sciences (2016) 28, 2–12
2. Giovanna Sannino, Ivanoe De Falco, Giuseppe De Pietro,Monitoring Obstructive Sleep Apnea by means of a real-time mobile system based on the automatic extraction of sets of rules through Differential Evolution, Journal of Biomedical Informatics 49 (2014) 84–100
3. J. Ilonen, J. K. Kamarainen, J. Lampinen, Differential Evolution Training Algorithm for Feed-Forward Neural Networks, Neural Processing letters 17 93-105, 2003.
4. J. Ilonen, J. K. Kamarainen, J. Lampinen, Differential Evolution Training Algorithm for Feed-Forward Neural Networks, Neural Processing letters 17 93-105, 2003
5. J. Jayanth Shivaprakash Koliwad, Ashok KumarT. Classification of remote sensed data using Artificial Bee Colony algorithm, The Egyptian Journal of Remote Sensing and Space Sciences (2015) 18, 119–126
6. J. Wu, Z. Cai, Attribute weighting via differential evolution algorithm for attribute weighted naive bayes (WNB), Journal of Computational Information Systems 7 (5) (2011) 1672–1679.
7. R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, K. Sandhu, K. Guppy, S. Lee, V. Froelicher, International application of a new probability algorithm for the diagnosis of coronary artery disease, American Journal of Cardiology 64 (1989) 304–310.
8. R.C. Holte, Very simple classification rules perform well on most commonly used datasets, Machine Learning 11 (1993) 63–91.
9. S. Das, A. Abraham, A. Konar, Automatic clustering using an improved differential evolution algorithm, IEEE Transactions on Systems, Man and Cybernetics Part A: Systems and Humans 38 (1) (2008) 218–237.
10. S.J. Haberman, Generalized residuals for log-linear models, in: Proceedings of the 9th International Biometrics Conference, 1976, pp. 104–122.
11. Tapas Bhadra, Sanghamitra Bandyopadhyay, Ujjwal Maulik, Differential Evolution Based Optimization of SVM Parameters for Meta Classifier Design, Procedia Technology 4 (2012) 50 – 57
12. W.H. Wolberg, O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proceedings of the National Academy of Sciences of the United States of America 87 (1990) 9193–9196.
13. Wei-Jie Chena, Yuan-Hai Shaoa, Ya-Fen Ye, Improving Lap-TSVM with successive overrelaxation and differential evolution, Procedia Computer Science 17 (2013) 33 – 40
14. Z. Cai, H. Chengyu, Z. Kang, L. Yong, L. Xiaobo, Y. Chao, Differential evolution based band selection in hyperspectral data classification, in: Proceedings of ISICA 2010 in Lecture Notes in Computer Science, Springer, 2010, pp. 86–94.

مطالعات علوم کاربردی در مهندسی

دوره ۳، شماره ۱، بهار ۱۳۹۶، صفحات ۲۱۸- ۲۰۸

A Combined Differential Evolution and Neural Network Approach to Extract Rules from Medical Databases

Mehdi Jobedari, Hamid Paygozar

Department of Computer Engineering, Islamic Azad University, Khomein Branch, Khomein, Iran

Abstract

Today, clinical decision support systems have been defined in medical science to offer useful suggestions using two or more items of the patient's characteristics. These suggestions help specialists in disease diagnosis or in the treatment process. The classification of data in this area requires transparency and reasoning for the allocation of data. Therefore, in addition to the correct classification of medical data, it is necessary to explain the reason for the data allocation. This research uses a new method based on differential evolution for automatic classification of items in the medical database. This method has introduced a tool called DEREx, which automatically extracts explicit knowledge from the database in the form of an IF-THEN rule, including the conditions connected by AND (and) on the database variables. Each item of differential evolution is considered a rule for classification. A set of differential evolution items is a set of rules, all of which are used to classify data in the database. These rules can also be logically linked by OR (or), so all classification rules are found for all categories at once in one step. The DEREx tool is the first classification tool based on differential evolution which automatically extracts categories of IF-THEN rules without the involvement of another mechanism.

Keywords: Classification, Differential Evolution, Database, Rule Extraction