ارائه یک راهکار ترکیبی داده کاوی مبتنی بر تئوری دمپستر شافر برای تشخیص بیماری دیابت

سعید دلکانی [۱]، مهدی صادق زاده [۲]

[۱] *گروه مهندسی کامپیوتر ،واحد بوشهر،دانشگاه آزاد اسلامی ، بوشهر ،ایران*
k.max1392@gmail.com

[۲] *گروه مهندسی کامپیوتر،واحدماهشهر،دانشگاه آزاد اسلامی، ماهشهر ، ایران*
sadegh_1999@yahoo.com

چکیده

امروزه پزشکان بیش از هر چیز با تکیه بر تجربیات و دانسته های خود، آزمایشات پیچیده و وقت گیر به تشخیص بیماری دیابت پی می برند. با این وجود خطاهای انسانی اجتناب ناپذیر است.در این پژوهش روشی ترکیبی برای تشخیص بیماری دیابت ارائه شده است چراکه یکی از مشکلات اساسی مربوط به این بیماری عدم تشخیص به موقع و صحیح این بیماری است. هدف از انجام این پژوهش، ارائۀ ساز وکاری برای بهبود دقت در تشخیص بیماری دیابت می باشد که این سازوکار بر اساس تجزیه و تحلیل داده‌های دیتاست PID با استفاده از سیستم های داده کاوی انجام می‌شود. بر اساس مطالعات انجام شده، ثابت شده است که سیستم‌های یادگیری مرکب نسبت به سیستم‌های ساده از دقت و عملکرد بهتری برخودار هستند. بنابراین در این پژوهش از یک سیستم ترکیبی داده کاوی مبتنی بر دمپستر شافر برای تشخیص بیماری دیابت ارائه شده است که در آن انتخاب ویژگی مبتنی بر همبستگی پیرسون با استفاده از الگوریتم ژنتیک و از روش‌های طبقه بندی متداول مانند شبکه عصبی، درخت تصمیم و ماشین بردار پشتیبان به عنوان سیستم‌های یادگیری پایه و برای ترکیب طبقه بندها از تئوری دمپستر – شافر استفاده شده است. بر اساس نتایج آزمایش‌های انجام شده، روش پیشنهادی نسبت به سیستم‌های پایه از عملکرد بهتری برخوردار بوده بیماران دیابتی را با دقت بهتری از یکدیگر تشخیص می دهد. دقت در مجموعه داده از ۸۷,۲۴ ٪ به ۸۹,۵۸ ٪ رسیده است.

کلمات کلیدی: دیابت، تشخیص بیماری، الگوریتم ژنتیک، شبکه عصبی، تئوری دمپستر شافر.

۱. مقدمه

بیماری دیابت یکی از شـایعترین بیماری های حاضر دنیا شناخته شده است که علیرغم گستردگی شیوع این بیماری هنوز روشی به منظور ریشه کن کردن و از بین بردن آن در دنیا شناخته نشده است هرچند که روش های مختلفی جهت تشخیص و کنترل آن در حال حاضـر مورد استفاده قرار می گیرد. از جمله عوارضی که به دنبال مبتلا شدن افراد به این بیماری گریبانگیر آنها خواهد شد میتوان به گرفتگی عروق قلبی و در نوع پیشرفته آن به نابینایی، قطع اعضای بدن، اختلالات فکری و غیره اشاره نمود.

بیماری دیابت را از نظر تقسـیم بندی میتوان به دو نوع وابسته به انسولین که در این نوع لوزالمعده شخص مبتلا به دیابت قادر به ترشـح انسـولین نمی باشـد و یا نوع غیر وابسته به انسولین که در آن لوزالمعده شخص مبتلا به دیابت قادر به تولید و ترشح انسولین می باشد اما میزان جذب آن در بدن بسیار اندک است. مشکل عمده ای که در رابطه با بیماری دیابت وجود دارد عدم تشـخیص به موقع و یا به طور کلی ضـعف در تشـخیص این بیماری اسـت که این ضـعف نیز به دلیل عدم انتخاب الگوی مناسـب توسـط پزشک و یا عدم استفاده مناسب از الگوهای استاندارد است. بنابراین پیاده سازی روشی که بتواند هر فرد را در تشـخیص صـحیح ابتلا یا عدم ابتلا به این بیماری یاری رسـاند میتواند گام مهمی در جهت پیشـگیری و کنترل این بیماری به خصوص در مراحل ابتدایی آن باشد .

امروزه با پیشـرفت های بیولوژیکی و توسعه تکنولوژی و استفاده از فناوری های روز و تجهیزات مدرن پزشکی، متخصصین قادرند تا به جمع آوری اطلاعات دقیق تری در مورد بیماران گردند که تحلیل آنان به دلیل حجم بالایی اطلاعات و متعدد بودن نمونه ها مشکل است و نیاز به فناوری جدیدتری می باشد که تکنولوژی های داده کاوی به کمک الگوریتم های قدرتمند خود، به این مسـئله مهم دست یافته اند. استخراج دانش از میان حجم انبوه داده های مرتبط با سوابق بیماری و پرونده های پزشکی افراد با استفاده از فرآیند داده کاوی، میتواند منجر به شناسایی قوانین حاکم بر ایجاد رشد و تسریع بیماری گردیده و اطلاعات ارزشـمندی را به منظور؛ شـناسـایی علل رخداد بیماری ها، پیش بینی و درمان بیماری ها با توجه به عوامل محیطی حاکم در اختیار متخصـصـین حوزه سـلامت قرار دهد. هدف از روش های پیشـگویی داده کاوی در پزشکی، ساخت یک مدل پیشگویانه است که به پزشکان کمک می کند تا روش های پیشگیری، تشخیص و برنامه های درمانی خود را بهبود بخشند؛ به همین دلیل داده کاوی در حوزه پزشـکی از اهمیت بالایی برخوردار اسـت و به عنوان تکنیکی برای شناسایی و تشخیص بیماری ها و دسته بندی بیماران در مدیریت بیماری و پیدا کردن الگوهایی برای تشـخیص سریعتر بیماران و جلوگیری از بروز عوارض در آنها می تواند کمک بسـیار بزرگی باشـد. افزایش دقت تشـخیص، کاهش هزینه ها و کاهش منابع انسـانی به عنوان مزایای معرفی داده کاوی در تجزیه و تحلیل پزشکی می باشد[۲و۱].

همچنین توانایی گروه بندی داده های پیچیده به یک تعداد متناهی از کلاس ها در داده کاوی مهم است و بدان معنی است که تصمیمگیری مفیدتر را میتوان بر اساس اطلاعات موجود ساخت. برای مثال، در زمینه تشخیص پزشکی استفاده از روشهایی که با دقت باال میتواند باعث تمایز بین دادههای غیرعادی و سالم شود، ضروری است [۳].

روش های داده کاوی مختلفی برای تشـخیص و مدیریت بیماری در علم پزشکی ارائه شده است. همچنین نشان داده شده اسـت که اسـتفاده از سـیسـتم های تشخیصی با کمک به عنوان "نظر دوم" منجر به بهبود تصمیم گیری در تشخیص بیماری شده است[۳] .

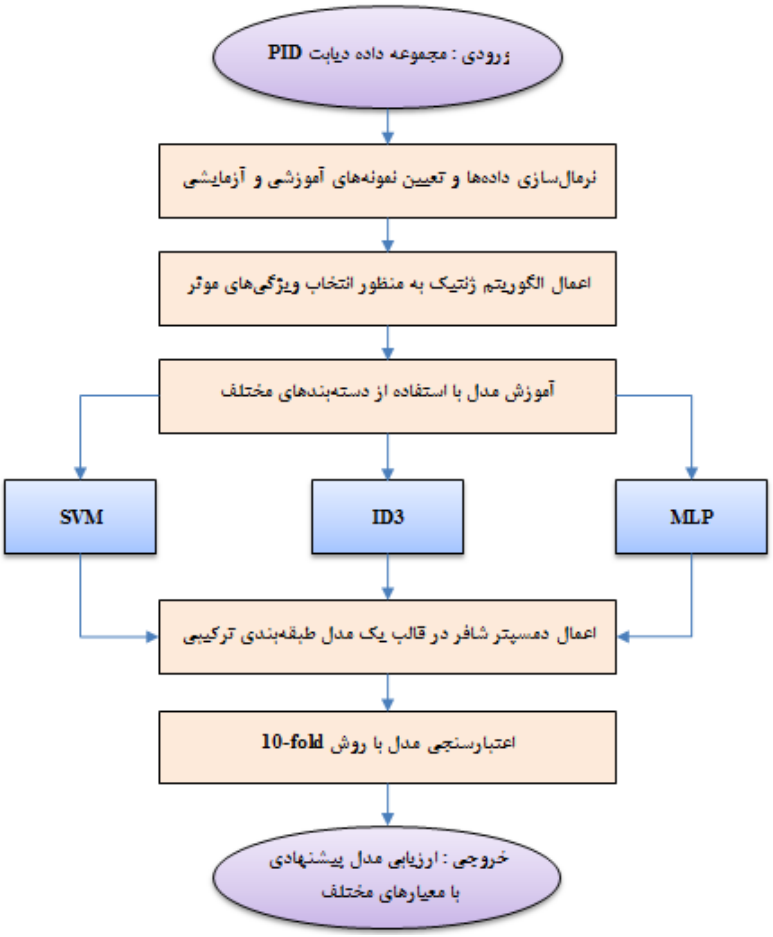
تمرکز کنونی تحقیقات شـدید در کلاسـه بندی الگوی ترکیبی از چندین سـیستم طبقه بندی است که می تواند به صورت مدل یکسان و یا متفاوت ساخته شود. این سیستم اطلاعات، تصمیم گیری های طبقه بندی را ترکیب می کند که برای غلبه بر محدودیت های روش های سنتی بر اساس طبقه بندی است[۳].

تاکنون روش های هوشمند گوناگونی به منظور حل این مشکل بنیادی در سراسر جهان ارائه گردیده که از آن جمله میتوان به استفاده از روش های تکاملی و نیز استفاده از روش الگوریتم های فازی تشخیص الگو در استخراج ویژگی و همچنین استفاده از روش شـبکه های عصـبی RBF اشـاره نمود. در این پایان نامه سـعی بر این اسـت که یک سیستم ترکیبی مبتنی بر تئوری دمسـتر-شـافر برای تشخیص بیماری دیابت ارائه شود. هدف از طراحی این سیستم ارایه یک سیستم هوشمند برای بالا بردن دقت پیش بینی بیماری دیابت می باشد.

۲. روش پیشنهادی

در روش پیشنهادی ابتدا داده های موجود به دو دسته داده های آموزشی و آزمایشی تقسیم بندی می شوند تا با استفاده از داده های آموزشی سیستم مورد نظر ایجاد شده و با استفاده از داده های آزمایشی سیستم ایجاد شده مورد ارزیابی قرار گیرد. عملیات نرمال سازی داده ها نیز در مرحله بعدی انجام می پذیرد. سپس در مرحله بعد، با استفاده از الگوریتم ژنتیک، عملیات انتخاب ویژگی انجام می شود. انتخاب ویژگی به معنی انتخاب ویژگی های بهتر و موثرتر در ویژگی های موجود می باشد. هدف از این عملیات، کاهش هزینه اجرایی و حافظه و همچنین افزایش دقت سیستم پیشنهادی می باشد. در انتها سیستم پیشنهادی با استفاده از داده های آموزشی و الگوریتم پیشنهادی ایجاد می گردد. الگوریتم پیشنهادی یک سیستم ترکیبی است که از چندین دسته بند تنها (مانند دسته بندها درخت تصمیم، ماشین بردار پشتیبان، شبکه های عصبی) استفاده کرده و درنهایت با استفاده از تئوری دمپستر شافر، عملیات ترکیب نتایج انجام خواهد شد.

در این تحقیق از الگوریتم ژنتیک به منظور انتخاب زیرمجموعه ای از ویژگی های موثر استفاده می شود. الگوریتم ژنتیک گونه ای از الگوریتم های تکاملی می باشد که یک روش تصادفی است و اصول اولیه آن از علم ژنتیک اقتباس گردیده است و دارای توانایی بهینه سازی بهتر می باشد و می تواند فضای تحقیقاتی بهینه شده را حاصل و ایجاد کند و جهت جست و جو را بصورت خودکار تنظیم کند.



شکل ۱ : بلوک دیاگرام روش پیشنهادی

۲-۱. انتخاب ویژگی ها با الگوریتم ژنتیک

یکی از اهداف اصلی در مسئله انتخاب ویژگی حذف ویژگی‌های نامرتبط و همچنین دارای افزونگی است که باعث کاهش دقت مدل‌های طبقه‌بندی می‌شود. هدف این مرحله از روش پیشنهادی انتخاب زیرمجموعه‌ای از ویژگی‌های موثر در داده‌های جرائم اینترنتی است که به حداکثر قابلیت پیش‌بینی کلاس هدف منجر شود. در این تحقیق از یک الگوریتم ژنتیک برای انتخاب ویژگی‌ها استفاده می‌کنیم. بخش‌های مختلف الگوریتم ژنتیک به شرح زیر است.

الف) ساختار نمایش راه‌حل‌ها : به منظور ایجاد راه‌حل‌هایی که علاوه بر ویژگی‌های موثر، تعداد آنها را نیز تشخیص دهد از سیاست طول متغیر ویژگی‌ها استفاده می‌شود. شکل ۲ ساختار کروموزوم‌ها جهت انتخاب زیر مجموعه‌ای از ویژگی‌های بهینه را نشان می‌دهد.

F_1	F_2	...	F_j	...	F_M
X_1	X_2	...	X_j	...	X_M

شکل ۲ : ساختار کروموزوم‌ها در مسئله انتخاب ویژگی‌ها

بردار F با مقادیر باینری به طول M در نظر گرفته می‌شود جائیکه M تعداد کل ویژگی‌ها در مجموعه داده را نشان می‌دهد ($j = 1,2, \dots, M$). در صورت انتخاب یک ویژگی عدد ۱ و در صورت عدم انتخاب ویژگی عدد ۰ در سلول متناظر با هر ویژگی درج می‌شود.

ب) ایجاد جمعیت اولیه : جمعیت اولیه به صورت تصادفی با توجه به ساختار کروموزوم پیشنهادی ایجاد می‌شود. اندازه جمعیت با توجه به پارامتر ورودی N_p مشخص می‌شود.

ج) تابع ارزیابی مبتنی بر همبستگی پیرسون : تابع ارزیابی یا تابع برازش استفاده شده بوسیله الگوریتم ژنتیک بر اساس این فرضیه است که یک ویژگی مربوط به میزان زیادی با متغیر(های) پاسخ همبسته است و با سایر ویژگی ها در زیر مجموعه ویژگی ها همبستگی کمتری دارد[۴]. همبستگی یک معیار دو جهته از قدرت ارتباط بین دو متغیر است.

در این مقاله از تابع برازش زیر برای ارزیابی استفاده می کنیم :

$$Fitness_s = \frac{N\overline{r}_{cf}}{\sqrt{N + N(N-1)\overline{r}_{ff}}} \qquad (۱)$$

در این رابطه S زیر مجموعه‌ای شامل N ویژگی است، $\overline{r}_{cf}$ میانگین همبستگی ویژگی-طبقه و $\overline{r}_{ff}$ میانگین همبستگی ویژگی-ویژگی است. این رابطه همبستگی پیرسون است که در آن همه متغیرها همگون شده‌اند. صورت کسر می‌تواند میزان تخمین گروهی از افراد را ارائه دهد، و مخرج کسر میزان افزونگی بین آنها را نمایش می‌دهد. این تابع مکاشفه‌ای ویژگی های نامربوط را بررسی می‌کند زیرا آنها تخمین زننده ضعیفی برای طبقه می‌باشند. صفات خاصه زاید نیز متمایز می‌شوند زیرا با یک یا چند صفت خاصه دیگر همبستگی بالایی خواهند داشت. بدلیل اینکه صفات خاصه بصورت مستقل در نظر گرفته می‌شوند، این الگوریتم نمی‌تواند صفات خاصه با تعامل قوی مانند آنچه که در یک مسئله توازن داریم را تشخیص دهد. البته نشان داده شده است که می‌تواند صفات خاصه مفید با سطوح متوسطی از تعامل را شناسایی نماید[۴].

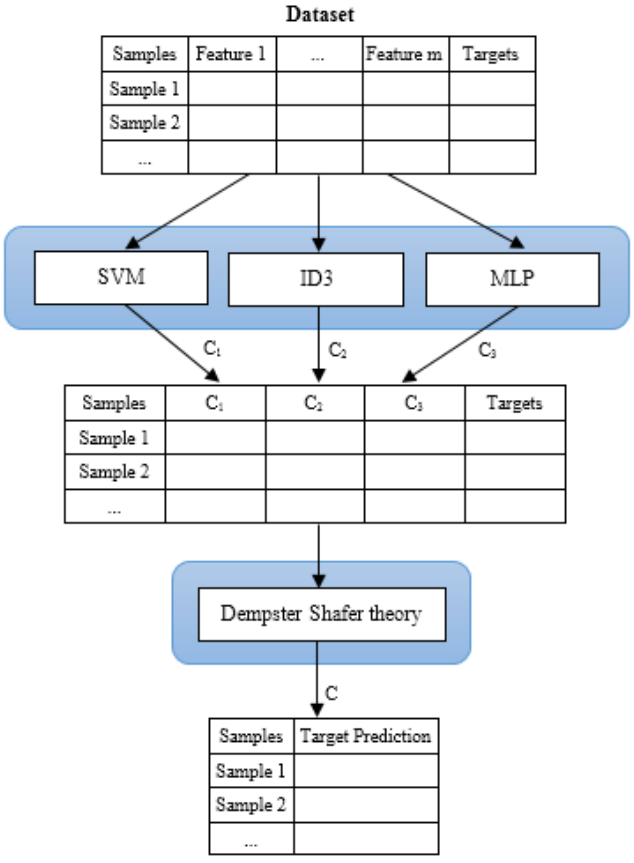
د) عملگرهای ژنتیکی : در این تحقیق از روش چرخ رولت برای انتخاب والدین جهت تولید و مثل استفاده می‌شود. عملگر ترکیب با احتمال CR براساس روش تک-نقطه‌ای و عملگر جهش با احتمال MR براساس روش جابه‌جایی بیت اعمال می‌شوند.

- انتخاب چرخ رولت : در مکانیزم چرخ رولت هر یک از کروموزوم ها بسته به میزان مناسب بودنشون (بر اساس تابع برازش) احتمال انتخاب شدن رو دارند. به عبارت دیگر هر چه یک کروموزوم بهتر باشه احتمال انتخاب شدنش برای تولید نسل بعدی بیشتر است. در این روش نسبت خوب بودن یک کروموزوم رو با یک میزان احتمال محاسبه کرده و هر چه این عدد بزرگتر باشه، احتمال انتخاب کروموزوم برای تولید مثل بیشتر خواهد بود.
- ترکیب تک-نقطه‌ای : در این روش عملیات تلفیق در یک نقطه انجام می‌شود. روش تولید مثل نیز بدین صورت است که ابتدا بصورت تصادفی، نقطه‌ای که قرار است تولید مثل از آنجا آغاز گردد، انتخاب می‌گردد. سپس فرزندان به صورت ضربدری از کروموزوم‌های والد ساخته می‌شوند.
- جهش جابه‌جایی بیت : پس از اتمام عمل ترکیب، عملگر جهش بر روی کروموزوم‌ها اثر داده می‌شود. عملگر جهش جابه‌جایی بیت دو ژن از یک کروموزوم را به طور تصادفی انتخاب نموده و سپس محتوای آنها را جابه‌جا می‌کند.

ه) **شرط توقف و جمعیت نسل بعد :** شرط توقف الگوریتم ژنتیک پیشنهادی تعداد نسل ثابت است که با متغیر max_{Gen} به عنوان ورودی الگوریتم درنظر گرفته می‌شود.

۲-۲. طبقه‌بندی ترکیبی مبتنی بر دمپستر-شافر

به منظور ارائه یک مدل طبقه‌بندی ترکیبی از الگوریتم‌های SVM، ID3 و MLP استفاده می‌کنیم. خروجی این سه طبقه‌بندی بر اساس نظریه دمپستر- شافر ترکیب می‌شوند. شکل ۳ فرایند ترکیب را نشان می‌دهد.



شکل ۳ : فرایند ترکیب به روش دمپستر- شافر

نظریه ریاضـی شـواهد، توسـط دمپسـتر معرفی شد و توسط شافر بسط داده شده است. این تئوری با بحث درباره باورهای موجود از یک وضعیت و یا سیستمی از وضعیت ها، حائز اهمیت می باشد. باورها در مورد پیشامدها یکسان نیستند؛ اما به کمک این نظریه میتوان شواهد موجود از وضعیت ها را در یک روش مشابه بررسی و ترکیب کرد. تئوری دمستر-شافر براساس باوری است که از شواهد نتیجه می شود؛ به طوری که ساختار باور تئوری شاهد به مدل احتمال کلاسیک مربوط می شود.

تئوری دمپستر- شافر یک ابزار قدرتمند برای بیان نایقینی است. دمستر- شافر با لحاظ نمودن عدم قطعیت در اعتبار فرضیه ها، یک شکل عمومی از تئوری بیز را ارائه نموده که در آن برای تعیین احتمال وقوع پیشامدها از بازه های احتمال و بازه های عدم قطعیت بر مبنای شواهد چندگانه (مثال حاصل از خروجی چند طبقه ند) استفاده می شود. با استفاده از تئوری دمستر-شافر برای ترکیب طبقه بندها، ابتدا با توجه به رفتار طبقه بندها شواهدی از آنها استخراج می شود که به عنوان دانش اولیه در پیدا کردن درجه عضویت الگو به هر یک از کلاس ها مورد استفاده قرار می گیرد. نظریه دمپستر- شافر به عنوان یک ابزار قدرتمند ترکیبی نشان داده شده است و تا به امروز بسیاری از تلاش های تحقیقاتی استفاده از آن را برای ترکیب نتایج تعدادی از تکنیک های طبقه بندی به صورت جداگانه نشان داده اند. به عنوان مثال، کرک پاتریک و همکارانش(۱۹۸۳)نتایج حاصل از طبقه بندی شبکه های بیزی و طبقه بندی مبتنی بر منطق فازی را با هم ترکیب کرده اند، احمدزاده و پترو (۲۰۰۳) نظریه دمستر- شافر را در روش شبکه عصبی برای مسأله تشخیص خطا به موتور القایی اعمال کردند، یانگ و کیم (۲۰۰۶)نیز روش ترکیبی طبقه بندها را بر اساس نظریه دمستر- شافر ارائه دادند و همچنین آنها توانایی قدرتمند موفقیت روش دمپستر-شافر را در ترکیب شواهد از طبقه بندی های متعدد بیان کردند.

در استفاده از تئوری دمپستر- شافر برای ترکیب طبقه‌بندها، ابتدا با توجه به رفتار طبقه‌بندها شواهدی از آن‌ها استخراج می‌شود که به‌عنوان دانش اولیه در پیدا کردن درجه عضویت الگو به هر یک از کلاس‌های الگو مورداستفاده قرار می‌گیرد. شواهد مورد نیاز با توجه به شباهت ماتریس پروفایل نمونه و ماتریس کلیشه تصمیم هر کلاس مشخص می‌شود. ماتریس پروفایل هر نمونه، ماتریسی است که از پاسخ هر یک از طبقه‌بندها در مورد آن نمونه تشکیل می‌شود به‌طوری‌که هر سطر این ماتریس نشان دهنده یکی از طبقه‌بندها و هر ستون نماینده یکی از برچسب‌های کلاس‌هاست. این ماتریس برای هر کلاس، با توجه به خروجی طبقه‌بندها برای نمونه‌های آن کلاس شکل می‌گیرد. به عنوان مثال روش شکل‌گیری کلیشه تصمیم برای کلاس w_i، ماتریس DT_i، به‌صورت زیر است.

نمونه‌های آموزشی را که برچسب‌های کلاسی آن‌ها مشخص است، به سیستم مرکب اعمال‌شده و با توجه به بردار خروجی طبقه‌بندها، ماتریس‌های پروفایل محاسبه می‌شود. در مرحله بعد، تمامی ماتریس‌هایی که مربوط به کلاس w_i هستند جدا شده و با میانگیری از درایه‌های نظیر به نظیر ماتریس کلیشه تصمیم کلاس w_i محاسبه می‌شود. به‌طور خلاصه می‌توان الگوریتم روش ترکیب دمپستر-شافر را در این حالت به‌صورت زیر بیان کرد:

الف) فرض کنید DT_i بیانگر iامین سطر ماتریس کلیشه تصمیم برای کلاس w_j باشد. نزدیکی بین این دو ماتریس محاسبه می‌شود، یعنی $f_{ji}(x)$ برای $j = 1, 2, \ldots, K$ و $i= 1, 2, ..,L$ محاسبه می‌شود. $\Phi_{ji}(x)$ در حقیقت بیانگر شباهت نظر طبقه‌بند D_i (در مورد الگوی x) به میانگین نظراتش در مورد الگوهای کلاس W_j است.

یکی از تعاریف ارائه‌شده برای Φ_{ji} به‌صورت رابطه زیر است.

$$\Phi_{ji}(x) = \frac{(1 + \left|\left|DT_j^i - D_i(x)\right|\right|^2)^{-1}}{\sum_{k=1}^{K}(1 + \left|\left|DT_k^i - D_i(x)\right|\right|^2)^{-1}} \qquad (۲)$$

ب) که در آن ||.|| بیانگر نرم ماتریس است. عبارت مخرج در رابطه فوق برای نرمالیزه کردن مقدار شباهت است به‌طوری‌که مجموع عناصر هر ستون ماتریس شباهت $\phi(x)$ برابر یک باشد.

ج) برای تمام کلاس‌ها و تمام طبقه‌بندها میزان باور را با استفاده از رابطه زیر به دست آورید.

$$b_j\big(D_i(x)\big) = \frac{f_{ji}(x).\prod_{k\neq j}(1-f_{ki}(x))}{1-f_{ji}(x)\{1-\prod_{k\neq j}\big(1-f_{ki}(x)\big)\}} \quad (۳)$$

د) درجه عضویت الگوی x به کلاس W_j را طبق رابطه زیر محاسبه کنید.

$$\mu_D^j(x) = k\prod_{i=1}^{L} b_j\left(D_i(x)\right) \quad j = 1,2,\dots,K \quad (۴)$$

اندیس متناظر با بزرگ‌ترین مؤلفه بردار μ_D کلاس الگوی x است. K در رابطه فوق ضریب مربوط به نرمالیزاسیون است.

۲-۲. بستر شبیه سازی

در این تحقیق از نرم افزار MATLAN ورژن 2017a برای پیاده‌سازی و آنالیز روش پیشنهادی استفاده شده است. همه آزمایش‌ها با یک پردازنده اینتل core i7 با فرکانس 2.5 GHz، حافظه 8 GB و سیستم عامل ویندوز 10 نسخه 64 bit انجام شده است.

در روش پیشنهادی از الگوریتم های ژنتیک به منظور انجام عملیات انتخاب ویژگی استفاده شده است. پارامترهای اولیه مقداردهی شده در این پیاده سازی در جدول ۱ نشان داده شده است.

جدول ۱ : مقداردهی اولیه الگوریتم تکاملی ژنتیک

پارامترها	مقادیر
جمعیت اولیه(تعداد کروموزوم ها)	۵۰
تکرارهای برنامه	۲۰۰
نرخ crossover	۰,۸
نرخ mutation	۰,۲
عملیات انتخاب	چرخ رولت

۲-۳. مروری بر مجموعه داده استفاده شده

در این تحقیق برای مقایسه تجربی روش پیشنهادی در برابر سایر الگوریتم‌های تشخیص بیماری دیابت، از پایگاه‌های داده واقعی دیابت PID استفاده شده که می‌باشد. مجموعه داده PID شامل ۸ ویژگی و ۷۶۸ نمونه می‌باشد که توسط وینسنت سیگیلیتو در دانشگاه جان هاپکینز تولید شده است. این مجموعه داده در اصل از موسسه ملی دیابت و بیماری های گوارشی و کلیه است و هدف آن این است که براساس اندازه گیری‌های تشخیصی، مبتلا بودن بیمار به دیابت را پیش بینی شود.

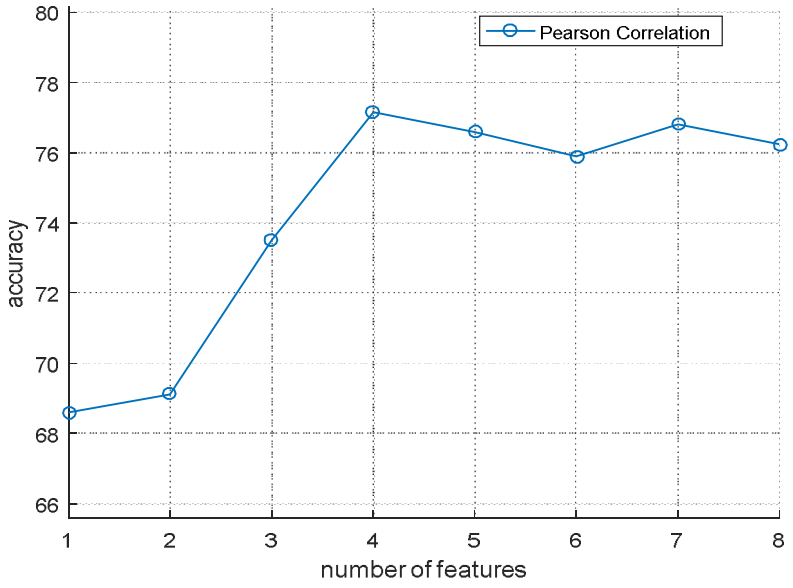
به طور خاص، تمام بیماران انتخاب شده برای ایجاد این مجموعه داده، زنانی با حداقل ۲۱ سال از کشور هند می‌باشند. با استخراج ۸ ویژگی «تعداد دفعات بارداری، غلظت گلوکز پلاسما، فشار خون دیاستولیک، ضخامت ورقه پوستی Triceps، ۲ ساعت سرم انسولین، شاخص توده بدن، تابع عملکرد دیابت و سن» از ۷۶۸ بیمار، مجموعه داده PID تبدیل به یک پایگاه داده بزرگ در این زمینه شده است. از ۷۶۸ نمونه موجود در این مجموعه داده، ۲۶۸ نمونه به بیماری دیابت مبتلا هستند (کلاس 1) و ۵۰۰ نمونه این بیماری را ندارند (کلاس 0). جدول ۲ اطلاعات کاملی از این مجموعه داده را نشان می‌دهد.

جدول ۲ : مشخصات مجموعه داده PID

شماره ویژگی	نام ویژگی	نام کوتاه ویژگی	حداقل مقدار	حداکثر مقدار
۱	تعداد دفعات بارداری	PregCount	۰	۱۷
۲	غلظت گلوکز پلاسما	Glucose	۰	۱۹۹
۳	فشار خون دیاستولیک	DBP	۰	۱۲۲
۴	ضخامت ورقه پوستی Triceps	Thickness	۰	۹۹
۵	۲ ساعت سرم انسولین	Insulin	۰	۸۴۶
۶	شاخص توده بدن	BMI	۰	۶۷,۱
۷	تابع عملکرد دیابت	Pedigree	۰,۰۷۸	۲,۴۲
۸	سن	Age	۲۱	۸۱

۲-۳. نتایج و آزمایش ها

تشـخیص تعداد ویژگی های مطلوب توسـط الگوریتم ژنتیک پیشنهادی بر مبنای تکنیک طول رشته متغیر انجام می شود. این تکنیک علاوه بر انتخاب ویژگی های مطلوب، تعداد بهینه این ویژگی ها را نیز مشخص می کند. در شکل ۴نمودار دقت الگوریتم پیشـنهادی با توجه به تعداد مختلف ویژگی ها نشـان داده شده است. نتایج برای روش پیشنهادی و با تعداد نسل ۲۰۰ گزارش می شـود. محاسبه دقت برای تعداد ویژگی های مختلف در طول روند بهینه سازی و در حالت بهترین معیار accuracy نشان داده شده است.



شکل ۴ : دقت طبقه بندی با تعداد ویژگی های مختلف

نتایج حاصل از این آزمایش نشان می‌دهد که بهترین دقت طبقه‌بندی پایگاه داده PID در ۴ ویژگی و دقت ۷۷,۱۴٪ می‌باشد.

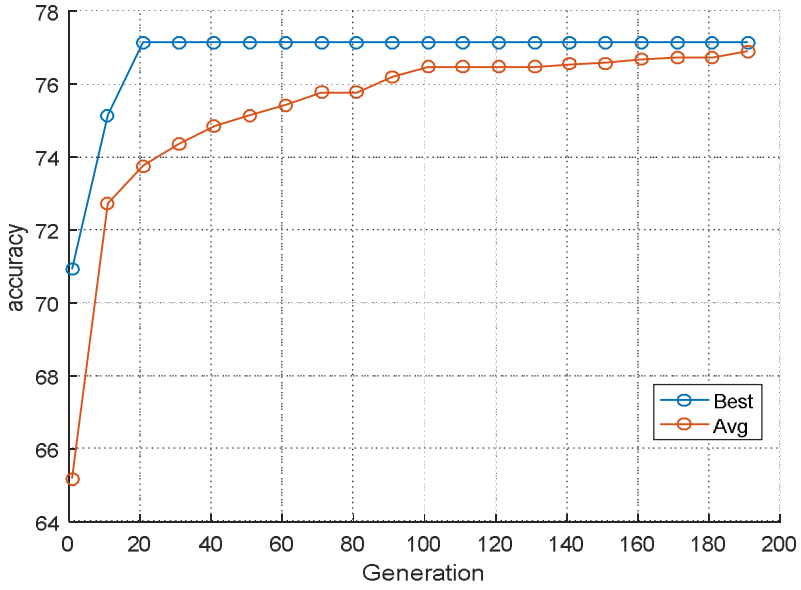
بهترین مدل طبقه‌بندی تولید شده از طریق الگوریتم پیشنهادی با و بدون انتخاب ویژگی با توجه به معیارهای مختلف در جدول ۳ مورد مقایسه قرار می‌گیرد. همچنین در این جدول زیر مجموعه ویژگی‌های انتخاب شده قابل دسترس است. نتایج با توجه به بخش الگوریتم ژنتیک و فرایند انتخاب ویژگی گزارش شده است.

جدول۳ : نتایج الگوریتم پیشنهادی با و بدون انتخاب ویژگی

حالت اجرا		دقت (٪)	ویژگی‌های انتخاب شده
با انتخاب ویژگی	بهترین حالت	۷۷,۱۴	تعداد دفعات بارداری، شاخص توده بدن، تابع عملکرد دیابت و سن
	حالت میانگین	۷۶,۹۰	
بدون انتخاب ویژگی	بهترین حالت	۷۶,۲۳	همه ویژگی‌ها
	حالت میانگین	۷۵,۶۸	

همگرایی این الگوریتم در مرحله انتخاب ویژگی در شکل ۵ نشان داده شده است. همگرایی عملکرد روش پیشنهادی را نسبت به تکرار در طول روند اجرا (تکرار نسل) نشان می‌دهد. نتایج برای روش پیشنهادی و با تعداد نسل ۲۰۰ گزارش می‌شود. نمودار همگرایی در دو حالت میانگین و بهترین برای بخش انتخاب ویژگی با معیار ارزیابی «دقت» نشان داده شده است.

نتایج، همگرایی الگوریتم را در نسل ۱۱ و با دقت ۷۷,۱۴٪ برای بهترین حالت و در نسل ۱۸۰ با دقت ۷۶,۹۰٪ در حالت میانگین نشان می‌دهد.



شکل ۵ : همگرایی الگوریتم پیشنهادی در دو بخش انتخاب ویژگی

همانطور که گفته شد در این تحقیق از 10-Fold برای ارزیابی مدل پیشنهادی استفاده می‌کنیم. این مدل در هر فولد دو مجموعه داده آموزشی و آزمایشی را مطابق با ۱۰٪ آزمایش و ۹۰٪ آموزش ایجاد می‌کند. نتایج در هر فولد بر روی مجموعه داده

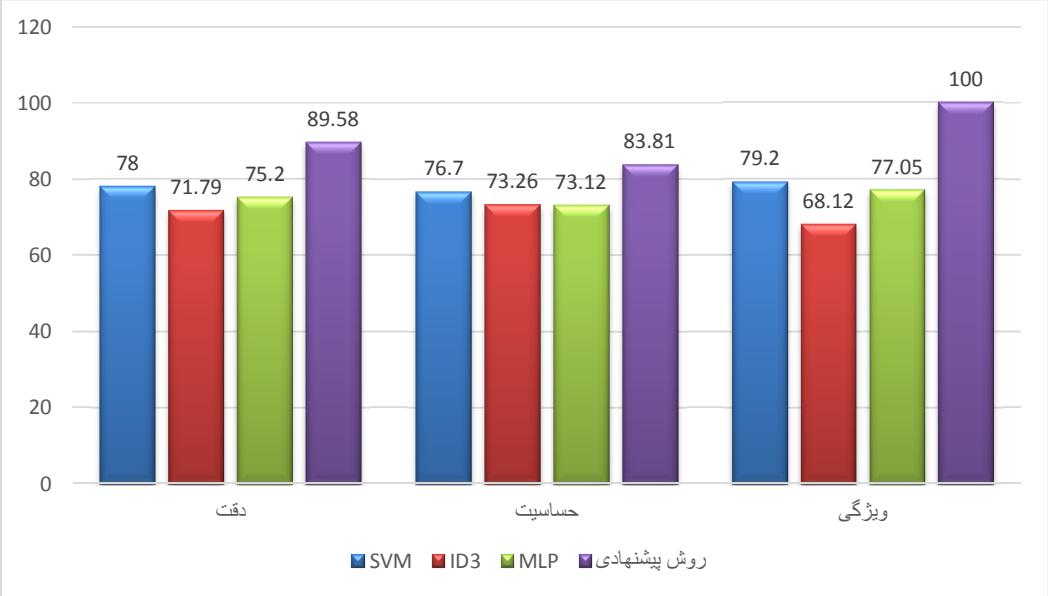
آزمایش برای به دست آوردن نتایج عملکرد اعمال می‌شود. این نتایج در قالب معیارهای دقت، حساسیت و ویژگی در جدول۴ نشان داده شده است.

جدول ۴ : نتایج اعمال 10-fold در روش پیشنهادی

شماره فولد	دقت (٪)	حساسیت (٪)	ویژگی (٪)	شماره فولد	دقت (٪)	حساسیت (٪)	ویژگی (٪)
۱	۹۰,۳۷	۸۳,۴۹	۱۰۰,۰	۶	۹۲,۰۵	۸۸,۴۵	۱۰۰,۰
۲	۸۸,۷۸	۸۱,۸۹	۱۰۰,۰	۷	۹۱,۶۵	۸۷,۴۴	۱۰۰,۰
۳	۸۹,۰۷	۸۲,۳۴	۱۰۰,۰	۸	۸۹,۱۰	۸۳,۴۸	۱۰۰,۰
۴	۸۳,۳۳	۷۹,۳۴	۱۰۰,۰	۹	۸۹,۱۰	۸۳,۰۰	۱۰۰,۰
۵	۸۷,۰۹	۸۴,۱۱	۱۰۰,۰	۱۰	۹۲,۳۳	۹۱,۶۴	۱۰۰,۰

نتایج دقت میانگین را در ۱۰ فود ۸۹,۵۸٪ نشان می‌دهد. نتایج میانگین برای حساسیت و ویژگی به ترتیب ۸۳,۸۱٪ و ۱۰۰٪ می‌باشد.

الگوریتم‌های SVM، ID3 و MLP سه مدل طبقه‌بندی استاندارد و محبوب هستند. در آزمایشی دیگر نتایج این سه الگوریتم در MATLAB اجرا شده و در مقایسه با روش پیشنهادی (ترکیب این سه طبقه‌بند با دمپستر شافر) بررسی شده است. نتایج این بررسی در شکل ۶ برای معیارهای مختلف گزارش شده است.



شکل ۶ : مقایسه عملکرد روش پیشنهادی با طبقه بندهای کلاسیک SVM، ID3 و MLP

از نتایج بدست آمده واضح است که روش پیشنهادی از روش های طبقه بندی کلاسیک در معیارهای دقت، حساسیت و ویژگی برتر است.

به منظور بررسی عملکرد روش پیشنهادی مقایسه‌ای را در برابر دو الگوریتم ترکیب سیستم استنتاج فازی سوگنو و الگوریتم کرم شب تاب برای بهبود تشخیص بیماری دیابت (SF-FA) [Sahebi and Ebrahimi, 2015] و الگوریتم روش یک مدل

طبقه بندی فازی مبتنی بر بهینه سازی ذرات اصلاح شده برای تشخیص بیماری دیابت FPSO [Shirali et al,2017] انجام شده است. نتایج با توجه به میانگین 10-Fold برای معیارهای مختلف به صورت میانگین در جدول ۵ نشان داده شده است.

جدول ۵ : مقایسه عملکرد روش پیشنهادی با روش‌های SF-FA و FPSO

الگوریتم‌ها	دقت (٪)	حساسیت (٪)	ویژگی (٪)
SF-FA	۸۷,۲۴	۸۴,۳۳	۸۸,۵۰
FPSO	۸۵,۱۹	۸۴,۲۰	۸۶,۳۹
روش پیشنهادی	۸۹,۵۸	۸۳,۸۱	۱۰۰,۰

۳. نتیجه‌گیری

در این مقاله هدف ، بهبود دقت تشخیص بیماری دیابت با استفاده از یک مدل ترکیبی می باشد. در این روش به کمک تئوری دمپستر شافر نتایج حاصل از الگوریتم های طبقه بندی با هم ترکیب شده تا نتایج دقیق تری تولید گردد. از نتایج شبیه سازی می توان به این نتیجه رسید که روش پیشنهادی در مقایسه با روش های مقایسه ای توانایی بهتری برخوردار می باشد.در مجموع روش پیشنهادی که بر اساس قانون ترکیب دمپستر شافر بنا شده نتایج کارآمدی را برای طبقه بندی نمونه های دیابتی ارائه نموده است.با توجه به این که هر طبقه بند دارای نقاط ضعف و قوت خود می باشد و هیچ کدام نمی تواند داده ها را بدون هیچ خطایی طبقه بندی کنند بنابراین مدل ترکیبی باعث تلفیق نتایج هر سه طبقه بند ID3 وsvm وmlp می شود و با توجه به این که تئوری دمپستر – شافر یک ابزار قوی برای ترکیب است که تصمیم نهایی بر اساس تلفیق اطلاعات بدست آمده از طبقه بندها تو لید می شود بنابراین روش پیشنهادی با در نظر گرفتن تضاد و همسو بودن شواهد و ترکیب بر اساس تئوری دمپستر – شافر توانسته در مقایسه با روش های مقایسه ای دقت تشخیص بیماری دیابت را بهبود دهد. و همچنین استفاده از الگوریتم ژنتیک با طول رشته متغیر برای انتخاب ویژگی های موثر در مجموعه داده دیابت PID باعث افزایش دقت این بیماری می شود.

در کارهای آتی ، می توان استفاده کرد از دیگر روش های ترکیب طبقه بندها نظیر رای اکثریت و پشته تعمیم یافته و مقایسه آن با روش حاضر. و استفاده از سایر روش ها برای محاسبه فیتنس در الگوریتم ژنتیک برای انتخاب ویژگی های موثر.

مراجع

[1]. Khajehei M. Etemady F. Data Mining and Medical Research Studies. In Computational Intelligence, Modelling and Simulation (CIMSiM), 2010 Second International Conference on, 2010; pp. 119-122.IEEE.

[2]. Jayalakshmi T. Santhakumaran A. A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks. In Data Storage and Data Engineering (DSDE), 2010 International Conference on, 2010; pp. 159-163.IEEE.

[3].Chen, Qi, et al. "Data classification using the Dempster–Shafer method." Journal of Experimental & Theoretical Artificial Intelligence 26.4 (2014): 493-517.

[4]. 1. Chatterjee, S., Hadi, A. S., and Price B., Regression Analysis by Example, Third ed. NewYork: Wiley Series, 2000..

[5]. Shirali, M., Madmoli, Y., Roohafza, J., Karimi, H., Baboli Bahmaei, A., & Ertebati, S. (2017). Improvement Diagnosis Of Diabetes Using A Combination Of Sugeno Fuzzy Inference Systems And Firefly Algorithms. Iranian Journal of Diabetes and Metabolism, 15(3), 172-176.

[6]. Sahebi, H. R., & Ebrahimi, S. (2015). A fuzzy classifier based on modified particle swarm optimization for diabetes disease diagnosis. Advances in Computer Science: an International Journal, 4(3), 11-17.

Provides a solution Data mining Combination based on Dempster–Shafer theory for Diagnosis of Diabetes

Saeed dalkani
Department of Computer Engineering, Faculty of Engineering, University of boushehr, boushehr, Iran, E-mail: k.max1392@gmail.com

Mehdi sadeghzadeh
Department of Computer Engineering, Faculty of Engineering, University of boushehr, boushehr, Iran, E-mail: sadegh_1999@yahoo.com

Abstract. Today, physicians mostly diagnose diabetes by relying on their experiences and knowledge and complicated and time-consuming experiments. Nonetheless, human errors are inevitable. A hybrid method is presented in the current study to diagnose diabetes because one of the main problems regarding this disease is lack of timely and correct diagnosis. The present study aims at presenting a mechanism to improve the accuracy of diabetes diagnosis. This mechanism is conducted based on the PID dataset analysis using data mining systems. According to studies, it is proved that hybrid learning systems are more accurate and outperform simple systems. Therefore, a hybrid data mining system based on Dempster-Shafer was presented in this study to diagnose diabetes, in which property selection is done based on Pearson's correlation and using the genetic algorithm. Common classification methods such as the neural network, decision tree and support vector machine were used as basic learning systems and Dempster-Shafer theory was used to combine the classifications. According to the experiments, the proposed method outperformed the basic systems and diagnosed diabetic patients at a higher accuracy. The dataset accuracy reached 89.58% from 87.24%.

Keywords: diabetes, diagnosis, genetic algorithm, neural network, Dempster-Shafer theory