یک سیستم خبره بر اساس سیستم استنتاج عصبی-فازی تطبیقی جهت تشخیص بیماری دیابت با انتخاب ویژگی مبتنی بر انحنا

زهره کریمی1\*، روح اله رمضانی

1. دانشکده، فنی و مهندسی، دانشگاه دامغان، دامغان، ایران
2. 1 دانشکده ریاضی و علوم کامپیوتر، دانشگاه دامغان، دامغان، ایران

z.karimi@du.ac.ir

چکیده: سیستم­های خبره پزشکی نقش مهمی در پیش­بینی زودهنگام بیماری­ها ایفا می­کنند. به ویژه اهمیت عملکرد صحیح این سیستم­ها با انتشار بیماری کرونا و اختصاص زمان و انرژی زیادی از کادر درمان به مراتب بیشتر از گذشته شده است. بیماری دیابت مورد تمرکز این مقاله است که در سیستم­های خبره معمولاً با یادگیری یک دسته­بند پیاده­سازی می­شود. عملکرد دسته­بندها در یادگیری ماشین به میزان زیادی وابسته به ویژگی­های ورودی است. ویژگی­های افزونه و با ابعاد بالا می­تواند باعث افت کارایی دسته­بند و افزایش هزینه شود. در این مقاله بکارگیری یک روش انتخاب ویژگی مبتنی بر انحنای منگر جهت حذف ویژگی­های نامفید و افزونه پیشنهاد شده است. با توجه به عملکرد موفق سیستم استنتاج عصبی-فازی تطبیقی در یش­بینی دیابت، بعد از انتخاب ویژگی­ها این دسته­بند اعمال شده و نتایج روی مجموعه داده PIMA با زمانی که از انتخاب ویژگی منگر استفاده نشده باشد، مقایسه شده است. ارزیابی انجام شده کارایی روش پیشنهادی را بر اساس ملاک­های دقت، F1 و تشخیص­پذیری تایید می­کند.

**کلید واژه‌ها:** مجموعه داده PIMA، سیستم خبره، سیستم استنتاج عصبی فازی تطبیقی- تشخیص دیابت- انتخاب ویژگی مبتنی بر انحنای منگر

1- مقدمه

هوش مصنوعی ابزار ارزشمندی جهت شناسایی بیماری و حفاظت از بیماران در سیستم­های مونیتورینگ است. هوش مصنوعی میزان اطمینان از تصمیم­گیری انجام شده در مواجهه با بیمار را با بهره­گیری از الگوریتم­ها و سیستم­های مفید افزایش می­دهد. کارکنان عرصه­ی سلامت خصوصا پرستاران و پزشکان به دلیل افزایش گسترده و غیرمنظره تعداد بیماران در طول همه­گیری کرونا بیش از حد کار می­کنند. در چنین شرایطی، تکنیک­های هوش مصنوعی جهت تشخیص بیمارانی که زندگی آن­ها در معرض خطر است بکار می­روند. تشخیص اولیه­ی بیماری­ها خصوصاً بیمارانی که خطر بستری و مرگ را در بیماران ابتلا به ویروس کرونا افزایش می­دهند بسیار اهمیت دارد. تمرکز این مقاله روی تشخیص دیابت بر اساس تکنیک­های داده­کاوی است، از آن­جا که دیابت عامل اصلی نابینایی، نارسایی کلیه و حملات قلبی است. تکنیک­های داده­کاوی شامل مراحل جایگذاری[[1]](#footnote-1) نمونه­های از دست رفته، انتخاب ویژگی و دسته­بندی است. تاکنون دسته­بندهای گوناگونی بدین منظور بکار رفته­اند، ماشین بردار پشتیبان، بیزین ساده، قوانین فازی، شبکه عصبی و درخت­های تصمیم از جمله­ی این روش­ها هستند. یکی از دسته­بندهای موفق در پیش­بینی دیابت، سیستم عصبی-فازی تطبیقی (ANFIS) است که عملکرد بسیار خوبی در سیستم­های خبره­ی پزشکی و از جمله در تشخیص دیابت داشته است [1] [2]. این سیستم، ترکیبی از سیستم استنتاج فازی و شبکه عصبی است با کاربردهای موفقی در زمینه­های گوناگون است [3].

 هر چند، در ادبیات یادگیری ماشین، ویژگی­های داده شده به عنوان ورودی دسته­بند اهمیت زیادی دارد. ویژگی­های افزونه یا نامرتبط با ورودی تاثیر مستقیمی بر افت کارایی دسته­بند داشته و باعث اتلاف هزینه می­شوند. لذا انتخاب ویژگی­های صحیح و مفید از مسائل مطرح در این زمینه است. روش­های انتخاب ویژگی در متون یادگیری ماشین به سه دسته­ی روش­های مبتنی بر فیلتر، روش­های مبتنی بر پوشاننده و روش­های ترکیبی تقسیم می­شوند. روش­های مبتنی بر فیلتر، اهمیت هر ویژگی را به صورت مجزا از سایر ویژگی­ها با بهره­گیری از ملاک­های گوناگون محاسبه می­کنند و ویژگی­های با اهمیت کمتر را حذف می­نمایند. روش­های مبتنی بر پوشاننده، در هر مرحله از انتخاب ویژگی،­ دسته­بند را نیز اعمال نموده و اثر ویژگی­های انتخاب شده را روی کارایی دسته­­بند محاسبه می­کنند. این روش­ها در مقایسه با روش­های مبتنی بر فیلتر، تاثیر صفات بر یکدیگر را در نظر می­گیرند اما هزینه­ی بیشتری در مقایسه با روش­های مبتنی بر فیلتر دارند. در مقابل، روش­های انتخاب ویژگی، روش­های اقتباس ویژگی نیز هستند که ویژگی­های جدیدی را از ویژگی­های ورودی استخراج می­کنند. معروف­ترین این روش­ها روش تحلیل مولفه­ی اصلی (PCA) است که روی داده­های دیابت به عنوان مرحله­ی پیش­پردازش قبل از اعمال به دسته­بند ANFIS مورد استفاده قرار گرفته است.

تمرکز این مقاله روی بحث انتخاب ویژگی به شیوه­ی مبتنی بر فیلتر به دلیل مشخصه­ی کارایی آن است. در این مقاله از یک روش مبتنی بر ریاضیات به نام انتخاب ویژگی مبتنی بر انحنای منگر استفاده شده که عملکرد موفقی در کاربردهای گوناگون از جمله کاربردهای پزشکی داشته است [4]. بعد از انتخاب ویژگی، ANFIS به داده­ها اعمال شده و کارایی آن با کارایی کل ویژگی­های ورودی و نیز PCA مقایسه شده است.

­در ادامه، در ابتدا پژوهش­های مرتبط در حوزه­ی پیش­بینی دیابت شرح داده شده، سپس روش پیشنهادی ارائه می­شود. در بخش چهار، مجموعه داده و نحوه­ی ارزیابی آمده است. در پایان، نیز نتیجه­ی پژوهش ارائه می­شود.

2- پژوهش­های مرتبط

پژوهش­های زیادی جهت پیش­بینی دیابت با استفاده از روش­های یادگیری ماشین انجام شده است. Kandhasamy و همکاران دسته­بندهای SVM، j48، k-نزدیک­ترین همسایه و جنگل تصادفی را به داده­های دیابت در دو حالت با اعمال گام­های پیش­پردازشی و بدون گام­های پیش­پردازشی مقایسه کرده­اند [5]. جزئیات گام پیش­پردازشی آن­ها ذکر نشده است و صرفا بیان شده که نویز حذف شده است. بهترین نتایج با استفاده از j48 برای حالت بدون پیش­پردازش و k-نزدیک­ترین همسایه و جنگل تصادفی بدست آمده است. Yuvaraj و همکاران از بهره­ی اطلاعات برای انتخاب ویژگی استفاده کرده­اند و سپس دسته­بندهای جنگل تصادفی، درخت تصمیم و بیزین ساده را به داده­های دیابت اعمال نموده­اند [6]. بر اساس تنظیمات بکاررفته در آزمایشات آن­ها، جنگل تصادفی به بهترین دقت دست یافته است. Tafa و همکاران مدلی ترکیبی از SVM و بیزین ساده پیشنهاد کرده­اند که نتایج آن از هر دو مدل SVM و بیزین ساده به صورت مجزا بهتر است [7]. Mercaldo و سایرین از شش دسته­بند J48، پرسپترون چندلایه، درخت هافدینگ، JRip، شبکه­ی بیزی و جنگل تصادفی جهت پیش­بینی دیابت استفاده کرده­اند دو الگوریتم GreedyStepwise و BestFit جهت تعیین صفات تمایزی به عنوان گام پیش­پردازشی اعمال شده­اند [8]. چهار صفت از هشت صفت اولیه­ی مجموعه داده­ی استاندارد دیابت به این روش انتخاب شده­اند. منطق فازی نیز جهت تشخیص رتینوپاتی[[2]](#footnote-2) زودهنگام بکار رفته است [9]. ANFIS نیز در پژوهش­های گوناگون جهت تشخیص دیابت اعمال شده است. رمضانی و سایرین تکنیک­هایی را جهت تعامل با داده­های گمشده اعمال کرده­اند و سپس یک مدل ترکیبی از ANFIS و رگرسیون لجستیک را به داده­های دیابت اعمال نموده­اند. Polat و سایرین یک سیستم خبره مبتنی بر ANFIS و تحلیل مولفه­های اصلی را نیز پیشنهادی کرده­اند [10].

3-سیستم خبره­ی پیشنهادی

در این بخش، جزییات سیستم خبره­ی پیشنهادی ارائه می­شود که شامل دو مرحله­ی انتخاب ویژگی و اعمال دسته­بند ANFIS است. دیاگرام مدل پیشنهادی در شکل (1) آمده است. مرحله­ی اول پیش­ پردازش داده­ها است. با توجه به این که در مجموعه داده استاندارد دیابت، داده­های از دست رفته[[3]](#footnote-3) وجود دارد، باید راجع به آن­ها تصمیم­گیری شود. نحوه­ی تعامل ما با داده­ها گمشده رویکرد بکار رفته توسط رمضانی و همکاران است [2]. بعد از حل مساله­ی داده­های گمشده، داده­ها نرمال می­شوند به نحوی که مقدار آن­ها بین 0 و 1 باشد. در گام بعد، ویژگی­های مناسب انتخاب می­شوند. روش انتخاب ویژگی که قسمت اصلی این مقاله است برگرفته از یک معیار ریاضی به نام انحنای منگر است. انحنای منگر عملکرد موفقی در تولید قوانین فازی تنک نشان داده است [11]، همچنین در کاربردهای گوناگون جهت انتخاب ویژگی بکاررفته است [4]. هدف این مقاله، بررسی تاثیر این روش روی مجموعه داده دیابت است. نهایتاً دسته­بند ANFIS به ویژگی­های انتخاب شده اعمال می­شود. جزئیات گام­های انتخاب ویژگی و اعمال دسته­بند در ادامه آمده است.



شکل 1: مراحل روش پیشنهادی

3-1- انتخاب ویژگی مبتنی بر انحنا

در این بخش، ابتدا انحنای منگر تعریف شده و سپس روش انتخاب ویژگی مبتنی بر آن شرح داده می­شود. انحنای منگر، با استفاده از شعاع دایره­­ی گذرنده از سه نقطه­ی $p\_{1}$، $p\_{2}$ و $p\_{3}$ ، انحنای سه نقطه را در فضای اقلیدسی $n$ بعدی اندازه­گیری می­کند. در شکل (2)، مفهوم انحنای منگر در فضای دو بعدی رسم شده است. با درنظرگرفتن فضای دو بعدی و با فرض سه نقطه­ی $p\_{1}$، $p\_{2}$ و $p\_{3}$ که خطی[[4]](#footnote-4) نیستند، انحنای منگر $MC$ روی $p\_{2}$ از رابطه­ی زیر اندازه­گیری می­شود:

|  |  |
| --- | --- |
| (1) | $$MC\left(p\_{1}, p\_{2}, p\_{3}\right)=\frac{1}{R}=\frac{2\sin(\left(φ\right))}{\left‖p\_{1},p\_{3}\right‖}$$ |
|   |  |

R شعاع دایره، $\left‖p\_{1},p\_{3}\right‖$ فاصله­ی اقلیدسی بین $p\_{1}$ و $p\_{3}$ و $φ$ زاویه­ی سمت $p\_{2}$ در مثلث تشکیل شده توسط $p\_{1}$، $p\_{2}$ و $p\_{3}$ است که از رابطه­ی زیر محاسبه می­شود:

|  |  |
| --- | --- |
| (2) | $$\cos(\left(φ\right))=\frac{\left‖p\_{1},p\_{2}\right‖^{2}+\left‖p\_{2},p\_{3}\right‖^{2}-\left‖p\_{1},p\_{3}\right‖^{2}}{2\left‖p\_{1},p\_{2}\right‖^{2}.\left‖p\_{2},p\_{3}\right‖^{2}}$$ |

$MC$ برای نقاط مرزی $p\_{1}$ و $p\_{3}$ قابل محاسبه نیست.

حال، هدف این است که با داشتن ماتریس $χϵR^{m×n} $ شامل $m$ داده­ی $n$ بعدی با برچسب مشخص $y$، ویژگی­هایی را مشخص کنیم که جهت هدف دسته­بندی مناسب­تر هستند. داده­ها نرمال شده هستند به نحوی که تمام مقادیر آن­ها بین 0 و 1 است. روش انتخاب ویژگی مبتنی بر انحنا شامل سه گام­ زیر است:

1. بازسازی داده­های دو بعدی: اولین گام، شامل تقسیم داده­های ورودی n بعدی در n فضای دو بعدی است، این فضاهای دوبعدی از ترکیب تمام صفات ورودی $F\_{i}^{'}(1\leq i\leq n)$ و خروجی به دست می­آید. سطوح دو بعدی با نماد $P\_{(F\_{i}^{'},y)}$ نمایش داده می­شود.
2. وزن­دهی ویژگی­ها: از انحنای منگر در سطوح دو بعدی جهت محاسبه­ی انحنای میانگین $F\_{i}^{'}$ استفاده می­شود. $P\_{(F\_{i}^{'},y)}$ شامل $m$ نقطه است. مقدار انحنای منگر برای $2\leq j\leq m-1$ نقطه با استفاده از رابطه­ی (1) محاسبه می­شود. مقدار انحنای ویژگی $F\_{i}^{'}$ ، $\hat{MC\_{F\_{i}^{' }}}$، بامیانگین­گیری مقدار انحنای منگر نقاط محاسبه شده بدست می­آید:

|  |  |
| --- | --- |
| (3) | $$\hat{MC\_{F\_{i}^{' }}}=\frac{1}{m-2}\sum\_{j=2}^{m-1}MC\_{m\_{j}}^{i} $$ |

$MC\_{m\_{j}}^{i}$ مقدار انحنای mمین داده در ویژگی $F\_{i}^{'}$ است. مقدار بیشتر $\hat{MC\_{F\_{i}^{' }}}$ نشان­دهنده­ی اهمیت بیشتر ویژگی $F\_{i}^{'}$ است.



**شکل 2:** انحنای منگر سه نقطه در فضای دوبعدی[4]

1. رتبه­دهی ویژگی­ها و انتخاب ویژگی: ویژگی­ها بر اساس مقدار $\hat{MC\_{F\_{i}^{' }}}$ مرتب شده و ویژگی­هایی که مقدار آن­ها از مقدار آستانه­ی از قبل مشخص شده بیشتر باشد انتخاب می­شود.

بعد از حذف ویژگی­ها، از دسته­بند ANFIS جهت دسته­بندی داده­های با تعداد ویژگی کمتر استفاده می­شود.

3-2- دسته­بند ANFIS

ANFIS یک سیستم ترکیبی است که در سال 1993 توسط Jang ارائه شد [12]. ANFIS یک سیستم استنتاج فازی است که در چارچوب شبکه­های عصبی مصنوعی (ANN) تطبیقی ارائه شده است. ایده­ی این روش ترکیب یادگیری شبکه عصبی با مزایای قوانین فازی است به نحوی که منعکس کننده­ی فکر و دانش انسان باشد. قوانین فازی قابلیت انعکاس فکر و توانایی استدلال انسان را در محیطهای فازی و توام با عدم قطعیت دارد. شبکه­های تطبیقی برای مدل کردن سیستم­های ناشناخته با داشتن ورودی و خروجی، مجموعه پارامترها و ساختار شبکه­ی بهینه فراهم می­سازد. ANFIS می­تواند تمام قوانین ممکن بر اساس ساختار ایجاد شده برای مساله را جهت کمک به افراد خبره ارائه کند و به همین دلیل می­توان از آن در ساخت یک سیستم خبره بهره گرفت. ANFIS شامل تعدادی گره[[5]](#footnote-5) است که به صورت مستقیم به یکدیگر متصل شده­اند، هر گره یک واحد پردازشی است و تمام یا برخی از گره­ها تطبیقی هستند. گره­های تطبیقی گره­هایی هستند که خروجی آن­ها مبتنی بر پارامترهایی است که قابل تغییر هستند. قانون یادگیری، مشخص کننده­ی نحوه­ی بروزرسانی این پارامترها جهت کمینه­سازی خطا است. مشخصه­ی مهم ANFIS، الگوریتم یادگیر ترکیبی آن است که دو روش نزول گرادیان و حداقل مربعات را به صورت توامان اعمال می­کند و همین ویژگی، عامل برتری آن نسبت به بسیاری از روش­های دیگر است.

ساختار ابتدایی سیستم استنتاج فازی به صورت مدلی است که ورودی­ها را به توابع عضویت (MF) مربوطه نگاشت می­کند. سپس توابع عضویت به قوانین و قوانین به مجموعه­ای از خروجی­ها نگاشت می­شوند. سرانجام خروجی­ها به تابع عضویت خروجی و تابع عضویت خروجی به یک مقدار خروجی نگاشت می­شود. ANFIS این مراحل را در چارچوب شبکه عصبی تطبیقی انجام می­دهد. در ادامه در ابتدا معماری ANFIS و سپس الگوریتم یادگیری آن شرح داده می­شود.

## معماری ANFIS

 جهت سادگی، فرض می­شود که سیستم استنتاج فازی دو ورودی و یک خروجی دارد. پایگاه قانون شامل دو قانون if-then فازی از نوع تاکاگی و سوگنو به صورت زیر است [13]:

|  |  |
| --- | --- |
|  | $$If x is A and y is B then z is f(x,y)$$ |

 A و B مجموعه­های فازی در بخش مقدم[[6]](#footnote-6) قانون و $f(x,y)$ معمولاً یک چندجمله­ای از متغیرهای $x$ و $y$ است اما سایر توابعی که بتوانند به صورت تقریبی خروجی سیستم را در ناحیه­ی فازی مشخص شده در بخش مقدم قانون توصیف کنند نیز می­توانند استفاده شوند. اگر $f(x,y)$ یک ثابت باشد یک مدل فازی سوگنو مرتبه صفر شکل داده می­شود که حالت خاصی از سیستم استنتاج فازی ممدانی [14] است. اگر $f(x,y)$ یک چندجمله­ای مرتبه 1 باشد، یک مدل فازی سوگنو مرتبه اول بدست می­آید. در یک سیستم استنتاج فازی سوگنو مرتبه اول با دو قانون، قوانین به صورت زیر بیان می­شوند:

|  |  |
| --- | --- |
|  | $If x is A1 and y is B1 then f1=p\_{1}x+q\_{1}y+r\_{1}$ |
|  | $If x is A2 and y is B2 then f2=p\_{2}x+q\_{2}y+r\_{2}$  |

در اینجا سیستم استنتاج فازی تاکاگی و سوگنو نوع 3 بکارگرفته شده است که در آن خروجی هر قانون ترکیب خطی متغیرهای ورودی به اضافه­ی یک عبارت ثابت است. ساختار ANFIS در شکل (3) نشان داده شده است. $x$ و $y$ ورودی­های شبکه هستند. به صورت خلاصه، اولین لایه، ورودی را به صورت فازی تبدیل می­کند، دومین لایه قوانین فازی را با استفاده از AND فازی محاسبه می­کند. سومین لایه، توابع عضویت را نرمال می­کند و چهارمین لایه بخش تالی قوانین فازی و آخرین لایه خروجی شبکه را نتیجه می­دهد. جزئیات مربوط به لایه­ها در ادامه آمده است.

لایه­ی 1 یا لایه­ی فازی: هر گره i در این لایه، گرهی با یک تابع گره تطبیقی است:

|  |  |
| --- | --- |
| (4) | $O\_{i}^{1}=μ\_{A\_{i}}\left(x\right), i=1,2$ |

که در آن $x$ ورودی به گره i و $A\_{i}$ متغیر زبانی[[7]](#footnote-7) مرتبط با این تابع گره بوده و $μ\_{A\_{i}}$ تابع عضویت $A\_{i}$ است. معمولاً $μ\_{A\_{i}}$ به یکی از شکل­های زیر انتخاب می­شود:

|  |  |
| --- | --- |
| (5) | $$μ\_{A\_{i}}\left(x\right)=\frac{1}{1+\left[\left(x-\frac{c\_{i}}{a\_{i}}\right)^{2}\right]\_{i}^{b\_{i}}}$$ |
| (6) | $μ\_{A\_{i}}\left(x\right)=exp⁡(-(\frac{x\_{i}-c\_{i}}{a\_{i}}))^{2}$ |

که در آن $x$ ورودی و $\left\{a\_{i},b\_{i},c\_{i}\right\}$ مجموعه پارامترهای اولیه[[8]](#footnote-8) هستند.

لایه­ی 2 یا لایه­ی ضرب: هر گره در این لایه یک گره ثابت است که شدت فعال شدن یک قانون، $w\_{i}$، را محاسبه می­کند. خروجی هر گره ضرب تمام سیگنال­های ورودی آن است و به صورت زیر محاسبه می­شود:

|  |  |
| --- | --- |
| (7) | $$O\_{i}^{2}=w\_{i}=μ\_{A\_{i}}\left(x\right)×μ\_{B\_{i}}\left(y\right), i=1,2$$ |

لایه­ی 3 یا لایه­ی Normalized: هر گره در این لایه یک گره ثابت است. گره $i$م نسبت شدت فعال شدن $i$مین قانون را به جمع شدت فعال شدن تمام قانون­ها محاسبه می­کند. خروجی گره $i$م شدت فعال شدن نرمال شده است:

|  |  |
| --- | --- |
| (8) | $$O\_{i}^{3}=\overbar{w}\_{i}=\frac{w\_{1}}{w\_{1}+w\_{2}} i=1,2$$ |

لایه­ی 4 یا لایه­ی غیرفازی­سازی: هر گره در این لایه یک گره تطبیقی با یک تابع گره است که به صورت زیر داده شده است:

|  |  |
| --- | --- |
| (8) | $$O\_{i}^{4}=\overbar{w}\_{i}.f\_{i}=w\_{i}\left(p\_{i}x+q\_{i}y+r\_{i}\right), i=1,2$$ |

$\overbar{w\_{i}}$ خروجی لایه­ی 3 و $\{p\_{i}, q\_{i},r\_{i}\}$ مجموعه پارامتر تالی است.

لایه­ی 5: این لایه از فقط یک گره ثابت تشکیل شده است که خروجی آن جمع سیگنال­های ورودی است و از رابطه­ی (7) بدست می­آید:

|  |  |
| --- | --- |
| (9) | $O\_{i}^{5}=overall output=\sum\_{i}^{}\overbar{w}\_{i}.f\_{i}=\frac{\sum\_{i}^{}w\_{i}f\_{i}}{\sum\_{i}^{}w\_{i}}$  |

## الگوریتم یادگیری

بر اساس ساختار ANFIS مشاهده می­شود که با داشتن مقادیر پارامترهای اولیه، خروجی نهایی به صورت ترکیب خطی پارامترهای تالی بیان می­شود. خروجی $f$ در شکل (3) به صورت زیر است:

|  |  |
| --- | --- |
| (10) | $f=\frac{w\_{1}}{w\_{1}+w\_{2}}f\_{1}+\frac{w\_{2}}{w\_{1}+w\_{2}}f\_{2}$$=\overbar{w\_{1}}f\_{1}+\overbar{w\_{2}}f\_{2}$$=\left(\overbar{w\_{1}}x\right)p\_{1}+\left(\overbar{w\_{1}}y\right)q\_{1}+\left(\overbar{w\_{1}}\right)r\_{1}+\left(\overbar{w\_{2}}x\right)p\_{2}+\left(\overbar{w\_{2}}y\right)q\_{2}+\left(\overbar{w\_{2}}\right)r\_{2}$  |

 $f$ تابعی خطی از پارامترهای تالی $\{p\_{1},q\_{1},r\_{1},p\_{2},q\_{2},r\_{2})$ است.

مجموعه پارامترهای ANFIS، پارامترهای تالی و پارامترهای تطبیقی هستند که در دو مرحله تخمین زده می­شوند: در اولین مرحله که مرحله­ی پیش­رو یا روبه جلو[[9]](#footnote-9) است، فرض بر این است که پارامترهای تطبیقی با الگوریتم خطای حداقل مربعات محاسبه می شوند. در مرحله­ی دوم که مرحله­ی رو به عقب[[10]](#footnote-10) است فرض می شود پارامترهای تالی ثابت هستند و پارامترهای تطبیقی با استفاده از الگوریتم نزول گرادیان بدست می­آیند. در این مرحله، سیگنال­های خطا که مشتق مربع خطا با توجه به خروجی هر گره است به صورت روبه عقب از لایه­ی خروجی به لایه­ی ورودی منتشر می­شود. بدین وسیله پارامترهای تطبیقی با استفاده از الگوریتم نزول گرادیان بروزرسانی می­شود. [15].



**شکل 3:** ساختار ANFIS نوع 3 [16]

1. ارزیابی

 مجموعه داده PIAM Indians Diabetes مجموعه داده­ای استاندارد است که در پژوهش­های مربوط به تشخیص بیماری دیابت مورد استفاده قرار می­گیرد. این مجموعه داده شامل داده­های از دست رفته[[11]](#footnote-11) است، لذا ابتدا شبیه [2] پر کردن مقادیر گمشده با استفاده از تکنیک جایگذاری چندگانه[[12]](#footnote-12) انجام می­شود. در نهایت، مجموعه داده شامل 753 داده خواهد بود. صفات این مجموعه داده در جدول 1 آمده است.

جدول 1- توصیف صفات مجموعه داده­ی PIMA Indian diabetes [17]

|  |  |  |
| --- | --- | --- |
| **ردیف** | نام صفت  | *توصیف* |
| 1 | Pregnancies | تعداد دفعات بارداری |
| 2 | Glucose | غلظت گلوکز گلاسما به مدت دو ساعت در آزمایش تحمل گلوکز خوراکی |
| 3 | BloodPressure | فشار خون دیاستولیک |
| 4 | SkinThickness | ضخامت پوست ماهیچه­ی سه سر |
| 5 | Insulin | سرم انسولین دو ساعته |
| 6 | BMI | شاخص توده­ی بدنی |
| 7 | DiabetesPedigreeFunction | عملکرد ارثی دیابت |
| 8 | Age | سن بر حسب سال  |
| 9 | Class Label | برچسب دسته ( 0 یا 1) |

داده­ها به صورت تصادفی به دو قسمت داده­ی آموزشی و داده­ی تست تقسیم می­شود. 75% داده­ها به عنوان داده­ی آموزشی در نظر گرفته شده است، مقدار انحنای منگر ویژگی­ها روی مجموعه داده­ی آموزشی مشخص شده است. نتیجه­ی اجرا روی 10 مجموعه داده­ی آموزشی تصادفی در شکل (4) نشان داده شده است. دو صفت شماره­ی 3 و 4 حذف می­شوند و دسته­بند ANFIS روی سایر صفات آموزش می­بیند. ملاک­های ارزیابی دقت، F1 و تشخیص­پذیری هستند که رابطه­ی آن­ها در ادامه آمده است:

|  |  |  |  |
| --- | --- | --- | --- |
| (11) |

|  |  |
| --- | --- |
| **Accuracy**$=\frac{TP+TN}{TP+TN+FP+FN}$**,** $F1=\frac{2×Precison×Recall}{Precison+Recall}$$Specificity=TN/(TN+FP)$ |  |

  |

TN، TP، FN و FP به ترتیب تعداد منفی­های درست، مثبت­های درست، منفی­های کاذب و مثبت­های کاذب هستند. روش پیشنهادی با حالتی که هیچ انتخاب ویژگی انجام نشده است مقایسه شده است. نتایج ارزیابی بر اساس ملاک­های دقت، F1 و تشخیص­پذیری[[13]](#footnote-13) در جدول 2 آمده و همانطور که نتایج نشان می­دهد روش پیشنهادی دقت بهتری در مقایسه با سایر روش­ها دارد.



نمودار 1- میانگین مقدار انحنای منگر صفات

جدول 2- مقایسه کارایی روش­ پیشنهادی با پژوهش­های مرتبط

|  |  |  |  |
| --- | --- | --- | --- |
| روش | دقت | F1 | تشخیص­پذیری |
| پیشنهادی | 1915/78 | 09/62 | 08/90 |
| بدون انتخاب ویژگی | 7128/77 | 86/60 | 43/89 |

**مراجع:**

|  |  |
| --- | --- |
| [1]  | M. Kirisci, Y. Hasan and S. M. Ubeydullah, "An ANFIS perspective for the diagnosis of type II diabetes," *Annals of Fuzzy Mathematics and Informatics 17, no. 2,* pp. 101-113, 2019.  |
| [2]  | R. Ramezani, M. Mansoureh and K. Seyedeh Malihe, "A novel hybrid intelligent system with missing value imputation for diabetes diagnosis," *Alexandria engineering journal 57, no. 3,* pp. 1883-1891, 2018.  |
| [3]  | z. Karimi and M. Farzinfar, "Estimation of power in combined cycle power plant using adaptive neuro-fuzzy inference system (in persian)," in *the first national conference of applied water and power industry*, 2020.  |
| [4]  | Z. Zuo, L. Jie and A. M. Noura, "Curvature-based Feature Selection with Application in Classifying Electronic Health Records," *arXiv preprint arXiv:2101.03581,* 2021.  |
| [5]  | J. P. Kandhasamy and B. S. J. P. C. S., "Performance analysis of classifier models to predict diabetes mellitus," *Procedia Computer Science 47,* pp. 45-51, 2015.  |
| [6]  | N. Yuvaraj and K. R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster," *Cluster Computing 22, no. 1 ,* pp. 1-9, 2019.  |
| [7]  | Z. Tafa, P. Nerxhivane and K. Bertran, "An intelligent system for diabetes prediction," in *In 2015 4th Mediterranean Conference on Embedded Computing (MECO)*, 2015.  |
| [8]  | F. Mercaldo, N. Vittoria and S. Antonella, "Diabetes mellitus affected patients classification and diagnosis through machine learning techniques," *Procedia computer science 112,* pp. 2519-2528, 2017.  |
| [9]  | G. Zahlmann, B. Kochner, I. Ugi, D. Schuhmann, B. Liesenfeld, A. Wegner, M. Obermaier and M. Mertz, "Hybrid fuzzy image processing for situation assessment," *IEEE Engineering in Medicine and Biology Magazine 19 (1),* pp. 76-83, 2000.  |
| [10]  | K. Polat and S. Güne¸s, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digital Signal Processing, 17,* pp. 702-710, 2007.  |
| [11]  | Y. Tan, P. S. Hubert, C. Fei, V. V. and Y. Longzhi, "Curvature-based sparse rule base generation for fuzzy rule interpolation," *Journal of Intelligent & Fuzzy Systems 36, no. 5,* pp. 4201-4214, 2019.  |
| [12]  | J.-S. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE transactions Sys. Man cybern. Syst. 23 (3),* pp. 665-685, 1993.  |
| [13]  | T. Takagi and S. M., "Derivation of fuzzy control rules from human operator’s control actions," in *Proc. IFAC Symp. Fuzzy Inform., Knowledge Representation and Decision Analysis*, 55-60, 1985.  |
| [14]  | E. Mamdani and A. S., " An experiment in linguistic synthesis with a fuzzy logic controller," *Int. J. Man-Mach. Stud. 7,* pp. 1-13, 1975.  |
| [15]  | S. Haykin, Neural Networks—A Comprehensive Foundation, New Delhi: Prentice-Hall of India Pvt, 2003.  |
| [16]  | M. Buragohain and M. Chitralekha, "A novel approach for ANFIS modelling based on full factorial design," *Applied soft computing 8, no. 1,* pp. 609-625, 2008.  |
| [17]  | N. Sambyal, S. Poonam and S. Rupali, "A review of statistical and machine learning techniques for microvascular complications in type 2 diabetes," *Current diabetes reviews 17, no. 2,* pp. 143-155, 2021.  |

An Expert System based on [ANFIS](https://ieeexplore.ieee.org/abstract/document/256541/) for Diabetes Diagnosis Using Curvature-based Fearure Selection

\*Zohre Karimi1

1 School of Engineering, Damghan University, Damghan, Iran

z.karimi@du.ac.ir

*Abstract*—Medical expert systems play an essential role in the early detection of diseases. The importance of these systems has become far greater than ever with the spread of corona virus and the allocation of a lot of time and energy to the medical staff. The focus of this paper is on diabetes diagnosis, which is usually implemented in expert systems by learning a classifier. According to machine learning literature, the classification performance is highly dependent on input features. Redundant features with high dimensionality can reduce the efficiency of the classifiers and increase the cost of training. In this paper, it is proposed to use a feature selection method based on Menger curvature to eliminate useless features. Due to the success of Adaptive neural-fuzzy inference system in the diabetes diagnosis, this classifier is applied to the selected features. The proposed method is applied to PIMA data set and the results is compared to the output of classifier without feature selection. The obtained results confirm the efficiency of the proposed method in the terms of accuracy, F1 and specificity.

Keywords— PIMA dataset, Expert system, Adaptive Neural Fuzzy Inference System, Diabetes Diagnosis, Menger curvature-based feature selection

1. Imputation [↑](#footnote-ref-1)
2. Retinopathy [↑](#footnote-ref-2)
3. Missing Values [↑](#footnote-ref-3)
4. Colinear [↑](#footnote-ref-4)
5. Node [↑](#footnote-ref-5)
6. Antecedent  [↑](#footnote-ref-6)
7. Linguistic Variable [↑](#footnote-ref-7)
8. Premise parameters [↑](#footnote-ref-8)
9. Forward [↑](#footnote-ref-9)
10. Backward [↑](#footnote-ref-10)
11. Missing [↑](#footnote-ref-11)
12. Multiple Imputation Analysis [↑](#footnote-ref-12)
13. Specificity [↑](#footnote-ref-13)