ORIGINAL ARTICLE

Predicting Type2 Diabetes Using Data Mining Algorithms

Amir Panah[1],
Samere Fallahpour[2]

[1] Lecturer, Department of Computer and Information Technology, Hadaf Institution of Higher Education, Sari, Iran
[2] MSc in Computer and Information Science, Mazandaran University of Medical Sciences, Sari, Iran

(Received August 24, 2019 ; Accepted October 11, 2020)

Abstract

Background and purpose: Today, information systems and databases are widely used and in order to achieve higher accuracy and speed in making diagnosis, preventing the diseases, and choosing treatments they should be merged with traditional methods. This study aimed at presenting an accurate system for diagnosis of diabetes using data mining and a heuristic method combining neural network and particle swarm intelligence.

Materials and methods: In this applied research, along with the training of the neural network, a particle swarm optimization algorithm was used to determine the weight of the optimal neural networks using RapidMiner Software on pima Indian Diabetes Dataset for 768 patients.

Results: The proposed algorithm was found to be in line with the real model. The highest accuracy, specificity, and sensitivity of the method, with 50 different tests, were 94.1%, 92.88%, and 92.12%, respectively.

Conclusion: In this study, average modeling error as a target function was minimized after a series of repetitions. By increase in initial population and number of replications, in addition to improving the accuracy of the proposed method, the sensitivity parameters and the positive predictive value ere improved. In fact, sensitivity and accuracy of the proposed method is better and higher than previous similar methods.

Keywords: data mining, diabetes, neural network, particle swarm optimization

J Mazandaran Univ Med Sci 2020; 30 (191): 22-30 (Persian).

*** Corresponding Author: Samere Fallahpour -** Mazandaran university of Medical Sciences, Sari, Iran
(E-mail: samere.fallahpour@gmail.com)

پژوهشی

مجله دانشگاه علوم پزشکی مازندران
دوره سی ام شماره 191 آذر سال 1399 (30-22)

مدل پیش بینی ابتلا به دیابت نوع 2 با استفاده از الگوریتم‌های داده‌کاوی

امیر پناه[1]
سامره فلاحپور[2]

چکیده

سابقه و هدف: استفاده گسترده از سیستم‌های اطلاعات و پایگاه‌های داده، ادغام آن را با شیوه‌های سنتی برای دستیابی به دقت و سرعت بالاتر جهت تشخیص و پیشگیری بیماری و انتخاب روش‌های درمان و تصمیم‌گیری‌ها به یک الزام تبدیل کرده است. این مطالعه با هدف ارائه یک سیستم دقیق برای تشخیص بیماری دیابت با استفاده از تکنیک داده کاوی و به کارگیری یک روش ابتکاری شامل ترکیب شبکه عصبی با الگوریتم هوش دسته جمعی ذرات، انجام پذیرفت.

مواد و روش‌ها: در این مطالعه کاربردی، همراه با آموزش شبکه عصبی از الگوریتم هوش دسته جمعی ذرات جهت تعیین بهینه‌تر اوزان شبکه عصبی با استفاده از نرم‌افزار رپیدماینر بر روی مجموعه داده pima مربوط به 768 بیمار درکشور هند استفاده گردید.

یافته‌ها: بررسی انجام شده نشان می دهد که الگوریتم پیشنهادی می تواند منطبق بر مدل واقعی باشد به طوری که بیش‌ترین مقدار دقت، ویژگی و حساسیت در روش پیشنهادی با تعداد 50 آزمایش مختلف، به ترتیب 94/1، 92/88، 92/12 درصد می باشد.

استنتاج: در روش پیشنهادی مدل پیش بینی دیابت نوع 2، متوسط خطای مدلسازی به عنوان تابع هدف بعد از یکسری تکرار کمینه شد با افزایش جمعیت اولیه و تعداد تکرارها علاوه بر افزایش دقت روش پیشنهادی باعث بهبود پارامترهای حساسیت، ویژگی پیش‌بینی مثبت نیز شد به‌طوری که حساسیت، دقت روش پیشنهادی نسبت به روش‌های مشابه که در گذشته بکار رفته بود، بهتر و بیش‌تر می‌باشد.

واژه های کلیدی: داده کاوی، دیابت، شبکه عصبی، هوش دسته جمعی ذرات

مقدمه

داده‌های پزشکی دارای ویژگی‌های منحصر به فردی هستند که برخی ازآن‌ها شامل ناهمگن بودن داده‌های پزشکی، موارد اخلاقی، قانونی می‌باشد. با وجود مشکلات و موانعی که در زمینه داده کاوی پزشکی وجود دارد، اهمیت داده‌های پزشکی و نقش کلیدی روش‌های تحلیل و به‌دست آوردن دانش مفید از این داده‌ها در سلامت بشر، منجر شده است که داده کاوی پزشکی تبدیل به یکی از مهم‌ترین حوزه‌های کاری داده کاوی شود و متخصصین فراوانی از سراسر دنیا، در زمینه رشد و تعالی روش‌ها و ابزارهای کسب دانش از این داده‌ها گام

مولف مسئول: سامره فلاح پور - ساری: میدان معلم، معاونت تحقیقات و فن آوری دانشگاه علوم پزشکی مازندران E-mail: : samere.fallahpour@gmail.com
1. مربی، گروه کامپیوتر و آی تی، موسسه آموزش عالی هدف، ساری، ایران
2. کارشناسی ارشد مهندسی نرم افزار، دانشگاه علوم پزشکی مازندران، ساری، ایران
تاریخ دریافت : 1398/6/2 تاریخ ارجاع جهت اصلاحات : 1398/10/7 تاریخ تصویب : 1399/7/20

بردارند(1-6،4). مهم‌ترین عاملی که موجب می‌شود مدیران در سازمان‌های پزشکی از داده‌کاوی استفاده کنند، شناخت این مطلب است که داده کاوی قابلیت تولید اطلاعاتی که بتواند برای تمام شرکای درگیر در صنعت پزشکی سودمند باشد، را دارا است. به عبارتی داده کاوی در حوزه پزشکی می‌تواند برای بیمارستان‌ها، کلینیک‌ها، پزشکان و بیماران مفید باشد. تاکنون فعالیت‌های کشف دانش زیادی در حوزه پزشکی انجام شده است و داده‌های بیولوژیکی افراد و بیماران توسط روش‌های مختلف داده کاوی مورد تجزیه و تحلیل قرار گرفته اند و روابط و الگوهای پنهان در میان انبوه داده‌ها کشف و استخراج شده است. یکی از بیماری‌های چند عاملی که غربالگری آن در جامعه از اهمیت خاصی برخوردار است دیابت نوع 2 است. عوامل خطرساز در ابتلا به دیابت نوع 2 اضافه وزن و چاقی، بی تحرکی یا کم تحرکی، رژیم غذایی با چربی بالا و فیبر کم، نژاد، سابقه فامیلی، سن، وزن کم هنگام تولد و غیره است. هر چه تعداد عوامل خطرساز در فرد بیش‌تر باشد بیش‌تر در معرض خطر ابتلا به دیابت نوع 2 قرار می‌گیرد(3). از طرفی کاوش و بررسی حجم زیادی از این اطلاعات، نیازمند استفاده از روش‌های موثر و کارآمد برای یافتن الگوهای مربوط در این اطلاعات می‌باشد که استفاده از تکنیک‌های مختلف داده کاوی به خصوص دسته‌بندی و الگوهای تکرار شونده می‌تواند کمک شایانی در این زمینه باشد(8،7). افزایش شیوع دیابت نوع 2 در همه‌ی دنیا به خصوص در کشورهای در حال توسعه از جمله ایران، نوعی اعلام خطر است. دیابت نوع 2 در بیش‌تر جوامع به یک اپیدمی تبدیل شده است و شواهد اپیدمیولوژی نشان می‌دهد که اگر اقدام‌های پیشگیرانه‌ی مؤثری انجام نشود، شیوع دیابت به طور جهانی افزایش خواهد یافت(9،5). براساس برآوردها، انتظار می‌رود در سال 2050 تعداد افراد دیابتی در دنیا به بیش از 330 میلیون نفر برسد که این تعداد دو برابر تعداد دیابتی‌ها در سال 2000 خواهد بود. همچنین، بسیاری از موارد جدید ابتلا به دیابت مربوط به کشورهای در حال توسعه است که به نظر می‌رسد خاورمیانه بیش‌ترین افزایش را در شیوع دیابت در سال 2030 خواهد داشت. تغییر عمده و سریع در سبک زندگی مردم این کشورها باعث افزایش شیوع چاقی و سایر عوامل خطرساز بیماری‌های غیرواگیر مانند فشارخون بالا و اختلال در چربی شده است که در سراسر دنیا به عنوان عمده‌ترین عوامل سبب شناختی مربوط به بروز دیابت نوع 2 شناخته شده‌اند. شناخت عوامل خطرساز مؤثر در بروز دیابت یک اقدام اساسی برای برنامه‌های پیشگیری از دیابت نوع 2 در هر جامعه‌ای است چرا که کاهش دادن این عوامل خطرساز باعث کاهش نرخ بروز دیابت نوع 2 خواهد شد. در این میان، یافتن معادل‌هایی برای تعیین اثر عوامل خطرساز و شدت ارتباط آن‌ها با ابتلا به دیابت، دارای اهمیت فراوان است. با توجه به اهمیت و بار فردی و اجتماعی این بیماری، لزوم شناسایی افراد در معرض خطر برای ابتلا به دیابت مشهود است. از این‌رو، نیاز به سیستمی که بتواند دقت بالا در تشخیص و پیش‌بینی این بیماری داشته باشد، احساس می‌شود. روش‌های مختلفی در داده‌کاوی جهت کشف دانش و یافتن الگوی پنهان وجود دارد که شبکه عصبی مصنوعی یکی از بهترین ابزارهای مدل‌سازی در داده کاوی می‌باشد. این الگوریتم‌ها از مجموعه ای از گره‌ها به نام نرون ساخته شده‌اند که هر گره ورودی و خروجی‌هایی با وزن خاص دارند. هر گره براساس تابعی خاص، محاسبه ساده‌ای انجام می‌دهد. بین گره‌ها اتصالاتی وجود دارد که براساس معماری شبکه مشخص می‌شوند. در کارهای گذشته الگوریتم مناسب و یا ترکیب الگوریتم‌های داده کاوی جهت تشخیص بیماری دیابت نوع 2 استفاده نشده و الگوریتم‌های به کار برده شده دقت تشخیص بسیار بالایی نداشته است و حتی بعضی از این روش‌ها برای افزایش دقت، زمان انجام کار را افزایش می‌دهند. در این مطالعه با آموزش شبکه عصبی از الگوریتم هوش دسته جمعی ذرات، یک سیستم دقیق برای تشخیص بیماری دیابت نوع 2 با

استفاده از تکنیک داده‌کاوی ارائه داده شد که همزمان با افزایش دقت و بالا بردن کارایی زمان مناسب، با به کارگیری یک روش ابتکاری شامل ترکیب شبکه عصبی با الگوریتم هوش دسته جمعی ذرات می‌باشد.

مواد و روش ها

در این مطالعه کاربردی، از دیتا ست pima استفاده شده است. این مجموعه داده شامل پیش‌بینی دیابت افراد هندی در طول مدت 5 سال می‌باشد. تعداد 768 داده با هشت ویژگی و ستون برچسب می‌باشد که 268 نفر بیمار و 500 نفر سالم می‌باشند(10).

مراحل الگوریتم پیشنهادی شامل موارد زیر است.

1) یک شبکه عصبی مصنوعی چند لایه با مشخصات زیر تعریف می‌شود.

- شبکه عصبی مصنوعی به کار رفته در این مطالعه فقط دارای یک خروجی می‌باشد که نشان‌دهنده وضعیت شخص از نظر سالم بودن یا بیماری است.
- تعداد لایه‌های شبکه عصبی:N_l
- تعداد گره‌های لایه i ام:N_i
- مقدار بایس گره i ام در لایه j ام:$B_{i,j}$
- تعداد ورودی‌های شبکه عصبی مصنوعی پیشنهادی با پارامتر n_f نشان داده می‌شود که به اندازه تعداد ویژگی‌های مجموعه داده pima می‌باشد (مجموعه داده به‌کار رفته در این مطالعه) که این ویژگی‌ها در زیر بیان شده است.
- تعداد دفعات بارداری
- غلظت گلوکز پلاسما در 2 ساعت در یک آزمایش تحمل گلوکز خوراکی
- فشار خون دیاستولیک (میلی‌متر جیوه)
- ضخامت پوست چین سه سر (میلی‌متر)
- سرم انسولین 2 ساعته
- شاخص توده بدنی (وزن در کیلوگرم/(قد در متر) ^2)
- عملکرد ارثی دیابت نوع 2
- سن (سال)

تعداد لایه های شبکه عصبی پیشنهادی و تعداد گره‌های لایه‌ی i ام شبکه عصبی به ترتیب با پارامتر $n_l n_i$ نشان داده می‌شود. در شبکه عصبی پیشنهادی هر گره یک بایاس دارد که با مقادیر اوزان ورودی به آن گره جمع می‌شوند. بایاس گره i ام در لایه j ام را با پارامتر $b_{i,j}$ نشان داده می‌شود. تعداد متغیرهایی که روش پیشنهادی، با آموزش شبکه عصبی تغییر می‌یابند با فرمول شماره 1 محاسبه می‌گردد.

(فرمول شماره 1):

$$v = \sum_{i \in nodes} \sum_{j \in layers} b_{ij} + \boldsymbol{n_f}.n_1 + \sum_{i=2}^{n} n_i n_{i+1}$$

2) جهت بهینه سازی اوزان شبکه عصبی مصنوعی یک تابع هدف که میانگین خطای مدلسازی است، از فرمول شماره 2 استفاده می‌شود.

(فرمول شماره 2):

$$\hat{e} = \frac{1}{n}\sum_{i=1}^{n}(f(x) - \hat{f}(x))^2$$

3) جهت یافتن کمینه تابع هدف به کمک الگوریتم ذرات نیاز است متغیرهای مسئله به درستی فرموله شود.

در لایه اول علاوه بر وزن‌های به‌کار رفته مقادیر بایاس نیز مطابق فرمول شماره 3 به وزن‌ها یا لایه‌ها اضافه شده است. در این رابطه W_1 مقدار ماتریس وزن‌ها و بایاس‌های جمع شده با آن است.

(فرمول شماره 3):

$$\boldsymbol{W_1} = \sum_{i \in layer1} \sum_{j \in layer2} w_{ij} + \boldsymbol{B_1}$$

در این رابطه، B_1 ماتریس بایاس لایه اول است که به شکل فرمول شماره 4 نشان داده می‌شود.

(فرمول شماره 4):

$$\boldsymbol{B_1} = [b_1 \quad \dots \quad b_n]$$

تعداد بایاس‌های رابطه بالا، برابر با تعداد گره های لایه اول یا n می‌باشد. در لایه دوم، جهت مشخص نمودن گره‌های فعال و غیر فعال می‌توان ماتریس گره را مطابق فرمول شماره 5 تعریف نمود.

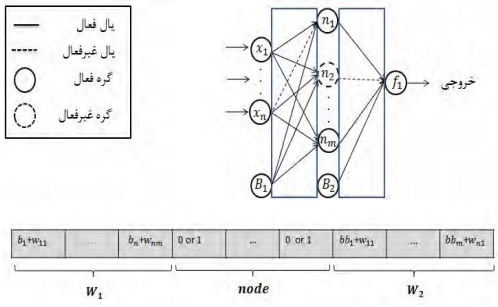
(فرمول شماره 5):
$$\boldsymbol{node} = \begin{bmatrix} \cdot \ or \ 1 \\ \dots \\ \cdot \ or \ 1 \end{bmatrix}_{m \times 1}$$

در ماتریس node که به تعداد گره‌های لایه دوم سطر دارد، در صورتی که گره فعال باشد مقدار آرایه متناظر با آن برابر 1 و در صورت غیر فعال بودن گره، مقدار آرایه متناظر با آن برابر 0 می‌باشد. ارتباط بین لایه دوم و سوم را می‌توان با ماتریس w_2 به شکل فرمول شماره 6 نشان داد. در این رابطه وزن‌های لایه دوم و سوم با مقادیر بایاس لایه دوم که با ماتریس B_2 نشان داده شده است، جمع می‌شوند. فرمول شماره 7 ماتریس B_2 را نشان داده است که شامل بایاس‌های لایه دوم است.

(فرمول شماره 6):
$$\boldsymbol{W_2} = \sum_{i \in layer2} \sum_{j \in layer3} w_{ij} + \boldsymbol{B_2}$$

(فرمول شماره 7):
$$\boldsymbol{B_2} = [bb_1 \quad \dots \quad bb_n]$$

می‌توان ماتریس‌های مورد نظر را به شکل یک آرایه خطی و مطابق تصویر شماره 1 نشان داد و از آن به عنوان جمعیت اولیه الگوریتم ذرات استفاده نمود.



تصویر شماره 1: فرموله سازی جمعیت اولیه در الگوریتم ذرات

آرایه‌ای که در تصویر شماره 1 به عنوان جمعیت اولیه مورد استفاده الگوریتم هوش ازدحامی ذرات استفاده می‌شود مطابق فرمول شماره 1 دارای متغیر مختلف می‌باشد.

4) در این مرحله تعدادی از بردارهای نشان داده در شکل شماره 1 به عنوان جمعیت اولیه الگوریتم ذرات در نظر گرفته می‌شود و شبکه عصبی مصنوعی را با داده‌های دیابت آموزش داده و با این جمعیت‌های اولیه آموزش می‌بیند و مراحل الگوریتم ذرات روی ذرات به اجراء گذاشته می‌شود. در هر مرحله بهترین ذره انتخاب می‌شود و در تکرار آخر بهترین ذره سراسری انتخاب شده و شبکه براساس این ذره سراسری مدلسازی می‌گردد. در مرحله ابتدایی الگوریتم هوش دسته جمعی ذرات، ذرات با موقعیت‌ها و سرعت‌های تصادفی ایجاد می‌شوند. در طی اجرای الگوریتم، موقعیت و سرعت هر ذره در مرحله t+1 ام از الگوریتم ذرات، از روی اطلاعات مرحله قبلی ساخته می‌شوند. اگر x_j مولفه‌ی j ام از بردار x باشد، آنگاه روابطی که سرعت و موقعیت ذرات را تغییر می‌دهند طبق رابطه شماره 1 و رابطه شماره 2 محاسبه می‌شود.

(رابطه شماره 1):
$$v_j^i(t+1) = wv_j^i(t) + c_1r_1\left(x_j^{ibest}(t) - x_j^i(t)\right) + c_2r_2\left(x_j^{gbest}(t) - x_j^i(t)\right)$$

(رابطه شماره 2):
$$x_j^i(t+1) = x_j^i(t) + v_j^i(t+1)$$

5) در این مرحله از داده های آزمون جهت ارزیابی روش پیشنهادی استفاده می شود.

معیارهای ارزیابی روش پیشنهادی

جهت ارزیابی روش پیشنهادی و مقایسه آن با روش‌های تشخیص بیماری دیابت نوع 2 سه معیار دقت، ویژگی و حساسیت در اکثر مطالعات بکار گرفته شده است. مقدار درصد دقت در بهترین و بدترین حالت به ترتیب 0 و 100 می باشد و نزدیکی به عدد 100 نشان دهنده دقت مناسب و خوب الگوریتم پیشنهادی است. در رابطه شماره3 نحوه محاسبه این معیار نشان داده شده است.

(رابطه شماره 3):

Accuracy= (TP+TN) / (TP+FP+TN+FN)

مقادیر به‌کار رفته در این رابطه به شرح ذیل است.

- TP[1]: افراد دیابتی که به درستی بیماری آن‌ها تشخیص داده شده است.
- TN[2]: افراد سالمی که به‌درستی سالم تشخیص داده می‌شوند.
- FP[3]: افرادی که به غلط دیابتی تشخیص داده می‌شوند.
- FN[4]: افرادی که به غلط سالم تشخیص داده می شوند.

جهت محاسبه معیار ویژگی و حساسیت می‌توان به ترتیب از رابطه شماره 4 و 5 استفاده نمود.

(رابطه شماره 4): Specificity=TN/(FP+TN)

(رابطه شماره 5): Sensitivity=TP/(TP+FN)

مقادیر معیارهای ویژگی و حساسیت در بهترین و بدترین حالت به ترتیب صد تا صفر درصد می‌باشند.

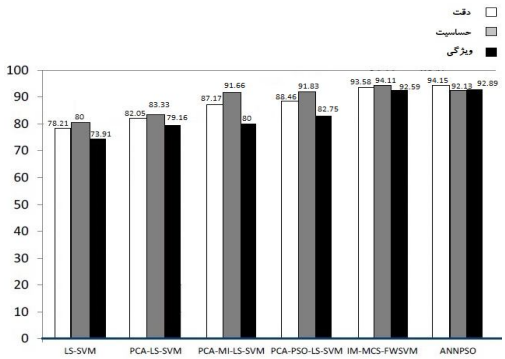
یافته ها

جهت پیاده‌سازی روش پیشنهادی حدود 90 درصد از داده‌های موجود در مجموعه داده pima را به‌عنوان داده آموزشی و مابقی داده‌ها را جهت ارزیابی و تست مورد استفاده قرار گرفت. داده‌های تست و یادگیری در هر اجرای برنامه به شکل تصادفی انتخاب شدند. تکرار الگوریتم ذرات به ترتیب200، 150، 100، 50 انتخاب گردید. همان‌طورکه در نمودارهای شماره 1 و2، مشخص شده است افزایش تکرار الگوریتم ذرات در روش پیشنهادی باعث افزایش عملکرد معیارهای دقت، حساسیت، ویژگی پیش‌بینی مثبت، پیش‌بینی منفی می‌شود. نتایج تحلیل این اشکال نشان می‌دهد در حالت کلی افزایش جمعیت اولیه دقت روش پیشنهادی را افزایش می‌دهد. جهت تحلیل و مقایسه دقت روش پیشنهادی با روش دیگر تشخیص بیماری دیابت، جمعیت اولیه و تعداد تکرار هرکدام 200 در نظرگرفته شد و جهت ارزیابی دقیق‌تر 50 آزمایش جداگانه انجام گرفت. در جدول شماره 1، روش پیشنهادی را با روش‌های مبتنی بر یادگیری ماشین با معیارهای دقت، حساسیت، ویژگی مورد مقایسه و ارزیابی قرار گرفت. اطلاعات جدول شماره 1 نشان می‌دهد، روش پیشنهادی از نظر معیار دقت و ویژگی از سایر روش‌های مندرج در این جدول عملکرد بهتری دارد. نمودار مقایسه‌ای روش پیشنهادی با روش‌های ذکر شده در نمودار شماره 1 نشان داده است.

جدول شماره 1: مقادیر پارامتر های بکار رفته در روش پیشنهادی براساس درصد

روش	دقت (درصد)	حساسیت (درصد)	ویژگی (درصد)
LS-SVM	78/2	80	73/9
PCA-LS-SVM	82	83/3	79/1
PCA-MI-LS-SVM	87/1	91/6	80
PCA-PSO-LS-SVM	88/4	91/8	82/7
MI-MCS-FWSVM	93/5	94/1	92/5
ANNPSO	94/1	92/1	92/8

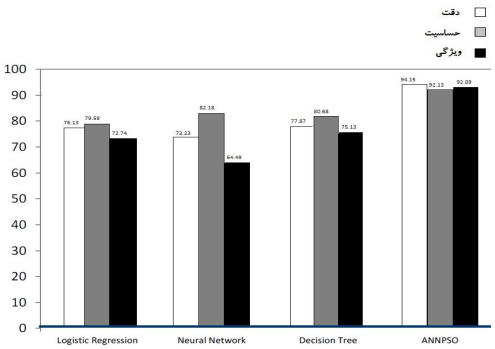
LS: Logistic Regression
SVM: Support Vector Machine
PCA: Principal Component Analysis
MI: Mutual Information
PSO: particle swarm optimization
MCS: Modified Cuckoo Search
FWSVM: Feature Weighted Support Vector Machines
ANNPSO: artificial neural network - particle swarm optimizatio



نمودار شماره 1: مقایسه عملکرد روش پیشنهادی با چند روش مبتنی بر یادگیری ماشین

1. TP: True Positive
2. TN: True Negative
3. FP: False Positive
4. FN: False Negative

روش پیشنهادی درنمودارشماره 2 در معیارهای دقت، حساسیت، ویژگی با سه تکنیک داده‌کاوی (رگرسیون، شبکه عصبی مصنوعی و در تصمیم‌گیری) مورد مقایسه قرار گرفته است. نتایج مقایسه شکل موردنظر نشان می‌دهد که روش پیشنهادی در سه معیار ارزیابی دقت، ویژگی و حساسیت نسبت به تکنیک‌های رگرسیون، شبکه عصبی مصنوعی و درخت تصمیم‌گیری عملکرد بهتری دارد.



نمودار شماره 2: مقایسه عملکرد روش پیشنهادی با روش رگرسیون، شبکه عصبی و درخت تصمیم گیری

بحث

در این مطالعه براساس پرونده بیماران دیابتی نوع 2 و اطلاعات ارزشمند موجود در این رکوردها، یک سیستم مبتنی بر شبکه عصبی مصنوعی با یادگیری به کمک هوش دسته جمعی ارائه شد و براساس این داده‌های آموزشی، بیماری دیابت نوع2 در افراد متقاضی، فقط با دو آزمایش بالینی غلظت گلوکز پلاسما 2 ساعت در آزمایش تحمل گلوکز خوراکی و سرم 2 ساعته انسولی تشخیص داده شود.

در مطالعه‌ای که کاوور و همکاران به پیش‌بینی بیماری دیابت نوع 2 با استفاده از مجموعه داده pima پرداخته‌اند، نشان داده شد، سطح سرم انسولین مهم‌ترین ویژگی در افراد دیابتی است، اگر مقدار انسولین بالای 800 باشد نشان می‌دهد فرد مبتلا به دیابت است. براساس نتایج تمام الگوریتم‌ها به جز الگوریتم پیشنهادی کاوور و همکاران دارای نرخ دقت کم‌تر از 78 درصد می‌باشند(11).

آنانسیاپادمانابهان و همکاران با استفاده از Naïve Bayes روش تشخیص رتینوپاتی دیابتی را ارائه داده‌اند که به پیش‌بینی بیماری در مراحل اولیه کمک می‌کند. این مطالعه بر روی مجموعه داده pima انجام شد و به دقت بالایی دست یافته است(12).

عسگرنژاد و همکاران در بررسی پیش‌بینی اولیه دیابت بر روی مجموعه داده pima را با استفاده از روش‌های مختلف داده کاوی، به دقت 84/35 درصد رسیدند(13).

در مطالعه‌ای که دوی و همکاران به افزایش دقت پیش‌بینی بیماری دیابت نوع 2 پرداخته‌اند، روش طبقه‌بندی Naïve Bayes ،SVM (Support Vector Machine) مورد بررسی قرارگرفت. در مطالعه مورد نظر از روش‌های پیش پردازش ترکیبی جهت جایگزینی مقادیر گمشده استفاده شد و در نهایت بهینه‌سازی انتخاب با استفاده از الگوریتم ژنتیک بر روی دیتاست Pima انجام گرفت(14). هووانگ و همکاران در مطالعه خود، با استفاده از الگوریتم‌های درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، شبکه بیزین به پیش‌بینی بیماری دیابت نوع 2 پرداخته‌اند و با استفاده از این روش به دقت بالای 60 درصد رسیدند(15).

بالاکریشنان و همکاران نیز برای پیش‌بینی بیماری دیابت، از الگوریتم‌های Naïve Bayes و SVM استفاده کردند. آن‌ها با کار بر روی دیتاست pima و با استفاده از نرم‌افزار وکا، به بررسی دقت الگوریتم‌های داده‌کاوی پرداخته و یک روش انتخاب ویژگی برای پیدا کردن یک زیرمجموعه ویژگی مطلوب پیشنهاد دادند که دقت طبقه‌بندی naïve bayes و svm را بالا برده است. تحلیل‌های خود را در نمودار سطح زیر منحنی مقایسه کرده که در نهایت الگوریتم SVM به دقت بالاتری رسید(16).

کراتی ساکنا و همکاران یک رویکرد ادغامی بین روش SVM، knn(K Nearest Neighbor) برای تشخیص بیماری دیابت نوع 2 بر روی مجموعه داده pima را

ارائه دادند و به این نتیجه رسیدند که knn نیز در این مجموعه داده جهت تشخیص بیماری دیابت نوع 2 جز موثرترین الگوریتم‌های هوش مصنوعی می‌باشد و توانسته به میانگین دقت 66 درصد برسد(17).

در مطالعه‌ای که پاوار و همکاران به بررسی جامع در مورد تشخیص دیابت نوع 2 در بهداشت و درمان پرداختند، تحلیل‌های خود را در نمودار سطح زیر منحنی مقایسه کردند و نتایج نشان داد که الگوریتم SVM به دقت بالاتری برای تشخیص رسیده است(18).

لنگری‌زاده و همکاران در مطالعه خود، با استفاده از شبکه عصبی به پیش‌بینی تولد نوزاد نارس در مادران باردار پرداختند که نتایج نشان داد، استفاده از شبکه پرسپترون چند لایه برای پیش‌بینی نتیجه زایمان از نظر تولد نوزاد ترم یا نوزاد نارس در مادران باردار از طریق فناوری‌های کمک باروری می‌تواند در پیشگیری از عوارض تولد نوزاد نارس کمک‌کننده باشد(19). در روش پیشنهادی شبکه عصبی مصنوعی به کمک هوش دسته جمعی ذرات بهترین اوزان یک شبکه عصبی مصنوعی را به کمک داده آموزشی مدلسازی کرده و یک مدل پیش‌بینی دیابت را ارائه داد که متوسط خطای مدلسازی به‌عنوان تابع هدف بعد از یکسری تکرار کمینه شد. در واقع از آن‌جا که دقت روش پیشنهادی به انتخاب مناسب اوزان شبکه عصبی مصنوعی بستگی دارد، لذا با انتخاب مناسب اوزان مسئله متوسط خطای مدل‌سازی کمینه شد و استفاده از الگوریتم هوش دسته جمعی، باعث شد در هر تکرار متوسط کمینه خطای مدل‌سازی کاهش یابد. همچنین با افزایش تعداد ذرات به کار رفته در روش پیشنهادی شبکه عصبی مصنوعی دقیق‌تر آموزش یافت و وزن‌های بهینه‌تری نسبت به جمعیت‌های کم‌تر به‌دست آمد. علاوه بر این با افزایش جمعیت اولیه و تعداد تکرارها علاوه بر افزایش دقت روش پیشنهادی باعث بهبود پارامتر های حساسیت و ویژگی نیز شد. انتظار می‌رود مدل ارائه شده در این مطالعه در مراحل بعدی جهت استفاده در برنامه‌های غربالگری جهت تعیین میزان خطر ابتلا و پیشرفت بیماری مورد استفاده قرار گیرد. همچنین کمبود داده به منظور آموزش بیش‌تر مدل، داده‌های گمشده در نمونه‌ها از محدودیت‌های این مطالعه می‌باشد که به‌طور حتم مدل‌های ارائه شده در این مطالعه را می‌توان بهبود بخشید. هر چند مدل الگوریتم به‌دست آمده دارای دقت بالایی است اما با جمع‌آوری داده‌های جدید می‌توان مدل را مجدد آموزش و دقت آن را افزایش داد(20).

References

1. Pardalos MP, Tomaino V, Xanthopoulos P. Optimization and data mining in medicine. Top 2009; 17(2): 215.
2. Wager KA, Lee FW, Glaser JP. Health Care Information Systems: A Practical Approach for HealthCare Managment. 2th ed. San Francisco: Jossey-Bass Inc; 2009.
3. Zimmet PZ. Diabetes epidemiology as a tool to trigger diabetes research and care. Diabetologia 1999; 42(5): 499-518.
4. Tan J. Medical Informatics. Concepts, Methodologies, Tools and Applications. Hershey: IGI Global; 2008.
5. Zimmet P. Globalization, coca-colonization and the chronic disease epidemic: can the Doomsday scenario be averted? J Intern Med 2000; 247(3): 301-310.
6. Siuly S, Li Y, Zhang Y. Improving Prospective Performance in MI Recognition: LS-SVM with Tuning Hyper Parameters, EEG Signal Analysis and Classification. New York: Springer; 2017.
7. Mahdizadeh H, Barani A. Clinical Data Mining: An Overview of Data Mining Techniques in Diabetes. IJDLD 2016; 15(4): 225-236.

8. Amin Ul Haq, Jian Ping Li, Jalaluddin Khan, Muhammad Hammad Memon, Shah Nazir, Sultan Ahmad, et al. Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data. Sensors 2020; 20(9): 2649.
9. Loannis Kavakiotis, OlghaTsave, Athanasios, Salifoglou, NicosMaglaveras, LoannisVlahavas, Loanna Chouvarda. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J 2017; 15: 104-116.
10. Pima diabetes. Available at: htpp://archive. ics.uci.edu/ml/machine-learning-databases/ pima-indians-diabetes 1990.
11. Kaur G, Chhabra A. Improved J48 classification algorithm for the prediction of diabetes. Int J Comput Appl 2014; 98(22): 107-126.
12. Ananthapadmanabhan K, Parthiban G. Prediction of chances-diabetic retinopathy using data mining classification techniques. Indian J Sci Technol 2014; 7(10): 1498-1503.
13. Asgarnezhad R, Shekofteh M, Zamani F. Improving diagnosis of diabetes mellitus using combination of preprocessing techniques. J Theor Appl 2017; 95(13): 2889-2895.
14. Devi MR, Shyla JM. Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus. Int J Appl Eng Res 2016; 11(1): 727-730.
15. Huang GM, Huang KY, Lee TYi, Weng J. An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients. BMC Bioinformatics 2015; 16(suppl1): S5.
16. Balakrishnan S, Ramaraj N, Savarimuthu N, Samikannu R. SVM ranking with backward search for feature selection in type II diabetes databases. IEEE International Conference on Systems, Man and Cybernetics; 2008 Oct 12-15; Singapore; 2008.
17. Saxena K, Khan Z, Singh Sh. Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm. IJCST 2014; 2(4): 36-43.
18. Pawar S, Sikchi S. An Extensive Survey on Diagnosis of Diabetes Mellitus in Healthcare. Proceedings of the 2th International Conference on Data Engineering and Communication Technology. Singapore, Springer; 2017.
19. Langarizadeh M, Ghazi Saeedi M, Karam Niay Far M, Hoseinpour M. Predicting Premature Birth in Pregnant Women via Assisted Reproductive Technologies using Neural Network. JHA 2016; 18(62): 42-51 (Persian).
20. Karegowda AG, Manjunath AS, Jayaram MA. Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes. Int J Soft Comput 2011; 2(2): 15-23.
21. Jahani M, Rezaeenour J, Mahdavi M, Hadavandi E. Proposing a Model for Predicting Diabetes Patients by Neural Network. Journal of Health Administration 2017; 20(67): 24-35 (Persian).