

Predicting Type2 Diabetes Using Data Mining Algorithms

Amir Panah¹,
Samere Fallahpour²

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)

مدل پیش بینی ابتلا به دیابت نوع 2 با استفاده از الگوریتم های داده کاوی

امیر پناه¹سامره فلاحپور²

چکیده

سابقه و هدف: استفاده گسترده از سیستم های اطلاعات و پایگاه های داده، ادغام آن را با شیوه های سنتی برای دستیابی به دقت و سرعت بالاتر جهت تشخیص و پیشگیری بیماری و انتخاب روش های درمان و تصمیم گیری ها 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 روش ها: در این مطالعه کاربردی، همراه با آموزش شبکه عصبی از الگوریتم هوش دسته جمعی ذرات جهت تعیین بهینه تر اوزان شبکه عصبی با استفاده از نرم افزار ریپد ماینر بر روی مجموعه داده pima مربوط به 768 بیمار در کشور هند استفاده گردید.

یافته ها: بررسی انجام شده نشان می دهد که الگوریتم پیشنهادی می تواند منطبق بر مدل واقعی باشد به طوری که بیش ترین مقدار دقت، ویژگی و حساسیت در روش پیشنهادی با تعداد 50 آزمایش مختلف، به ترتیب 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 پیش بینی مثبت نیز شد به طوری که حساسیت، دقت روش پیشنهادی نسبت به روش های مشابه که در گذشته بکار رفته بود، بهتر و بیش تر می باشد.

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

مقدمه

و به دست آوردن دانش مفید از این داده ها در سلامت بشر، منجر شده است که داده کاوی پزشکی تبدیل به یکی از مهم ترین حوزه های کاری داده کاوی شود و متخصصان فراوانی از سراسر دنیا، در زمینه رشد و تعالی روش ها و ابزارهای کسب دانش از این داده ها گام

داده های پزشکی دارای ویژگی های منحصر به فردی هستند که برخی از آن ها شامل ناهمگن بودن داده های پزشکی، موارد اخلاقی، قانونی می باشد. با وجود مشکلات و موانعی که در زمینه داده کاوی پزشکی وجود دارد، اهمیت داده های پزشکی و نقش کلیدی روش های تحلیل

E-mail: samere.fallahpour@gmail.com

مؤلف مسئول: سامره فلاح پور - ساری: میدان معلم، معاونت تحقیقات و فن آوری دانشگاه علوم پزشکی مازندران

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 (1-4، 6). مهم ترین عاملی که موجب می شود مدیران در سازمان های پزشکی از داده کاوی 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 می تواند برای بیمارستان ها، کلینیک ها، پزشکان و بیماران مفید باشد. تاکنون فعالیت های کشف دانش زیادی در حوزه پزشکی انجام شده است و داده های بیولوژیکی افراد و بیماران توسط روش های مختلف داده کاوی مورد تجزیه و تحلیل قرار گرفته اند و روابط و الگوهای پنهان در میان انبوه داده ها کشف و استخراج شده است. یکی از بیماری های چند عاملی که غربالگری آن در جامعه از اهمیت خاصی برخوردار است دیابت نوع 2 است. عوامل خطر ساز 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 خطر ساز در فرد بیشتر باشد بیش تر در معرض خطر ابتلا به دیابت نوع 2 قرار می گیرد (3). از طرفی کاوش و بررسی حجم زیادی از این اطلاعات، نیازمند استفاده از روش های موثر و کارآمد برای یافتن الگوهای مربوط در این اطلاعات می باشد که استفاده از تکنیک های مختلف داده کاوی به خصوص دسته بندی و الگوهای تکرار شونده می تواند کمک شایانی در این زمینه باشد (7، 8). افزایش شیوع دیابت نوع 2 در همه ی دنیا به خصوص در کشورهای در حال توسعه از جمله ایران، نوعی اعلام خطر است. دیابت نوع 2 در بیش تر جوامع به یک اپیدمی تبدیل شده است و شواهد اپیدمیولوژی نشان می دهد که اگر اقدام های پیشگیرانه ی مؤثری انجام نشود، شیوع دیابت به طور جهانی افزایش خواهد یافت (5، 9). براساس برآوردها، انتظار می رود در سال 2050 تعداد افراد دیابتی در دنیا به بیش از 330 میلیون نفر برسد که این تعداد دو برابر تعداد دیابتی ها در سال 2000 خواهد بود. همچنین، بسیاری از موارد جدید

ابتلا به دیابت مربوط به کشورهای در حال توسعه است که به نظر می رسد خاورمیانه بیش ترین 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 خطر ساز بیماری های غیر واگیر مانند فشارخون بالا و اختلال در چربی شده است که در سراسر دنیا به عنوان عمده ترین عوامل سبب شناختی مربوط به بروز دیابت نوع 2 شناخته شده اند. شناخت عوامل خطر ساز مؤثر در بروز دیابت یک اقدام اساسی برای برنامه های پیشگیری از دیابت نوع 2 در هر جامعه ای است چرا که کاهش دادن این عوامل خطر ساز باعث کاهش نرخ بروز دیابت نوع 2 خواهد شد. در این میان، یافتن معادلهایی برای تعیین اثر عوامل خطر ساز و شدت ارتباط آن ها 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 این رو، نیاز به سیستمی که بتواند دقت بالا در تشخیص و پیش بینی این بیماری داشته باشد، احساس می شود. روش های مختلفی در داده کاوی جهت کشف دانش و یافتن الگوی پنهان وجود دارد که شبکه عصبی مصنوعی یکی از بهترین ابزارهای مدل سازی در داده کاوی می باشد. این الگوریتم ها از مجموعه ای از گره ها به نام نرون ساخته شده اند که هر گره ورودی و خروجی هایی با وزن خاص دارند. هر گره براساس تابعی خاص، محاسبه ساده ای انجام می دهد. بین گره ها اتصالاتی وجود دارد که براساس معماری شبکه مشخص می شوند. در کارهای گذشته الگوریتم مناسب و یا ترکیب الگوریتم های داده کاوی جهت تشخیص بیماری دیابت نوع 2 استفاده نشده و الگوریتم های به کار برده شده دقت تشخیص بسیار بالایی نداشته است و حتی بعضی از این روش ها برای افزایش دقت، زمان انجام کار را افزایش می دهند. در این مطالعه با آموزش شبکه عصبی از الگوریتم هوش دسته جمعی ذرات، یک سیستم دقیق برای تشخیص بیماری دیابت نوع 2 با

- عملکرد ارثی دیابت نوع 2
- سن (سال)

تعداد لایه های شبکه عصبی پیشنهادی و تعداد گره های لایه ی i ام شبکه عصبی به ترتیب با پارامتر $n_i n_i$ نشان داده می شود. در شبکه عصبی پیشنهادی هر گره یک بایاس دارد که با مقادیر اوزان ورودی به آن گره جمع می شوند. بایاس گره i ام در لایه z ام را با پارامتر b_{iz} نشان داده می شود. تعداد متغیرهایی که روش پیشنهادی، با آموزش شبکه عصبی تغییر می یابند با فرمول شماره 1 محاسبه می گردد.

(فرمول شماره 1):

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

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

$$\hat{e} = \frac{1}{n} \sum_{i=1}^n (f(x) - \hat{f}(x))^2 \quad \text{(فرمول شماره 2):}$$

(3) جهت یافتن کمینه تابع هدف به کمک الگوریتم ذرات نیاز است متغیرهای مسئله به درستی فرموله شود. در لایه اول علاوه بر وزن های به کار رفته مقادیر بایاس نیز مطابق فرمول شماره 3 به وزن ها یا لایه ها اضافه شده است. در این رابطه w_1 مقدار ماتریس وزن ها و بایاس های جمع شده با آن است.

$$w_1 = \sum_{i \in \text{layer } 1} \sum_{j \in \text{layer } 2} w_{ij} + B_1 \quad \text{(فرمول شماره 3):}$$

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

$$B_1 = [b_1 \quad \dots \quad b_n] \quad \text{(فرمول شماره 4):}$$

استفاده از تکنیک داده کاوی 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 می باشد.

مواد و روش ها

در این مطالعه کاربردی، از دیتا ست pima استفاده شده است. این مجموعه داده شامل پیش بینی دیابت افراد هندی در طول مدت 5 سال می باشد. تعداد 768 داده با هشت ویژگی و ستون برجسب می باشد که 268 نفر بیمار و 500 نفر سالم می باشند (10).

مراحل الگوریتم پیشنهادی شامل موارد زیر است.

(1) یک شبکه عصبی مصنوعی چند لایه با مشخصات زیر تعریف می شود.

- شبکه عصبی مصنوعی به کار رفته در این مطالعه فقط دارای یک خروجی می باشد که نشان دهنده وضعیت شخص از نظر سالم بودن یا بیماری است.

- تعداد لایه های شبکه عصبی: N_i

- تعداد گره های لایه i ام: N_{iz}

- مقدار بایاس گره i ام در لایه z ام: B_{iz}

- تعداد ورودی های شبکه عصبی مصنوعی پیشنهادی با پارامتر n_f نشان داده می شود که به اندازه تعداد ویژگی های مجموعه داده pima می باشد (مجموعه داده به کار رفته در این مطالعه) که این ویژگی ها در زیر بیان شده است.

- تعداد دفعات بارداری

- غلظت گلوکز پلاسما در 2 ساعت در یک آزمایش تحمل گلوکز خوراکی

- فشار خون دیاستولیک (میلی متر جیوه)

- ضخامت پوست چین سه سر (میلی متر)

- سرم انسولین 2 ساعته

- شاخص توده بدنی (وزن در کیلوگرم / (قد در متر)²)

استفاده می شود مطابق فرمول شماره 1 دارای متغیر مختلف می باشد.

(4) در این مرحله تعدادی از بردارهای نشان داده در شکل شماره 1 به عنوان جمعیت اولیه الگوریتم ذرات در نظر گرفته می شود و شبکه عصبی مصنوعی را با داده های دیابت آموزش داده و با این جمعیت های اولیه آموزش می بیند و مراحل الگوریتم ذرات روی ذرات به اجراء گذاشته می شود. در هر مرحله بهترین ذره انتخاب می شود و در تکرار آخر بهترین ذره سراسری انتخاب شده و شبکه براساس این ذره سراسری مدل سازی می گردد. در مرحله ابتدایی الگوریتم هوش دسته جمعی ذرات، ذرات با موقعیت ها و سرعت های تصادفی ایجاد می شوند. در طی اجرای الگوریتم، موقعیت و سرعت هر ذره در مرحله $t+1$ ام از الگوریتم ذرات، از روی اطلاعات مرحله قبلی ساخته می شوند. اگر x_i مولفه ی i ام از بردار x باشد، آنگاه روابطی که سرعت و موقعیت ذرات را تغییر می دهند طبق رابطه شماره 1 و رابطه شماره 2 محاسبه می شود.

(رابطه شماره 1):

$$v_j^i(t+1) = w v_j^i(t) + c_1 r_1 (x_j^{ibest}(t) - x_j^i(t)) + c_2 r_2 (x_j^{gbest}(t) - x_j^i(t))$$

$$x_j^i(t+1) = x_j^i(t) + v_j^i(t+1) \quad \text{(رابطه شماره 2):}$$

(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 روش های 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 شماره 3 نحوه محاسبه این معیار نشان داده شده است.

تعداد بایاس های رابطه بالا، برابر با تعداد گره های لایه اول یا n می باشد. در لایه دوم، جهت مشخص نمودن گره های فعال و غیر فعال می توان ماتریس گره را مطابق فرمول شماره 5 تعریف نمود.

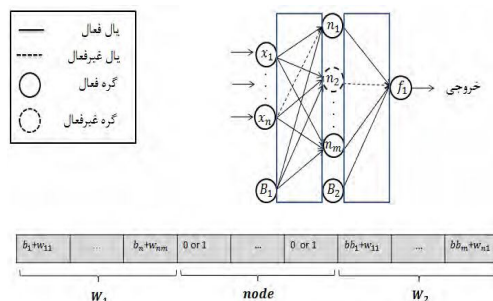
$$node = \begin{bmatrix} 0 \text{ or } 1 \\ \dots \\ 0 \text{ or } 1 \end{bmatrix}_{m \times 1} \quad \text{(فرمول شماره 5):}$$

در ماتریس node که به تعداد گره های 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 می باشد. ارتباط بین لایه دوم و سوم را می توان با ماتریس w_2 به شکل فرمول شماره 6 نشان داد. در این رابطه وزن های لایه دوم و سوم با مقادیر بایاس لایه دوم که با ماتریس B_2 نشان داده شده است، جمع می شوند. فرمول شماره 7 ماتریس B_2 را نشان داده است که شامل بایاس های لایه دوم است.

$$w_2 = \sum_{i \in layer1} \sum_{j \in layer2} w_{ij} + B_2 \quad \text{(فرمول شماره 6):}$$

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

می توان ماتریس های 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

آرایه ای که در تصویر شماره 1 به عنوان جمعیت اولیه مورد استفاده الگوریتم هوش ازدحامی ذرات

(رابطه شماره 3):

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{TN} + \text{FN})$$

مقادیر به کار رفته در این رابطه به شرح ذیل است.

TP¹: افراد دیابتی که به درستی بیماری آن‌ها تشخیص داده شده است.

TN²: افراد سالمی که به درستی سالم تشخیص داده می‌شوند.

FP³: افرادی که به غلط دیابتی تشخیص داده می‌شوند.

FN⁴: افرادی که به غلط سالم تشخیص داده می‌شوند.

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

(رابطه شماره 4): $\text{Specificity} = \text{TN} / (\text{FP} + \text{TN})$

(رابطه شماره 5): $\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$

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 به عنوان 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 ترتیب 50، 100، 150، 200، انتخاب گردید. همان‌طور که در نمودارهای شماره 1 و 2، 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

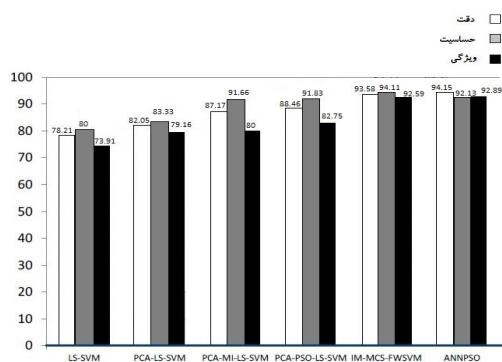
1. TP: True Positive
2. TN: True Negative
3. FP: False Positive
4. FN: False Negative

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 هر کدام 200 در نظر گرفته شد و جهت ارزیابی دقیق تر 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 پارامترهای بکار رفته در روش پیشنهادی

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 optimization



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

کاوور و همکاران دارای نرخ دقت کم تر از 78 درصد می باشند (11).

آنانسا پادمانابهان و همکاران با استفاده از Naïve Bayes روش تشخیص رتینوپاتی دیابتی را ارائه داده اند که به پیش بینی بیماری در مراحل اولیه کمک می کند. این مطالعه بر روی مجموعه داده pima انجام شد و به دقت بالایی دست یافته است (12).

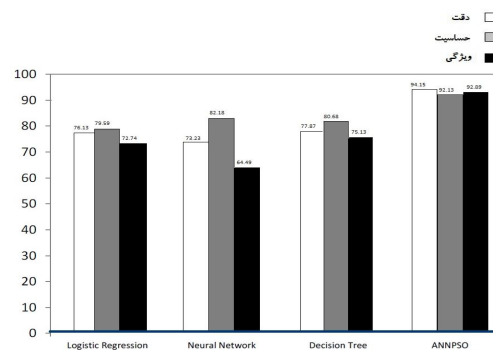
عسگر نژاد و همکاران در بررسی پیش بینی اولیه دیابت بر روی مجموعه داده pima با استفاده از روش های مختلف داده کاوی، به دقت 84/35 درصد رسیدند (13).

در مطالعه ای که دوی و همکاران به افزایش دقت پیش بینی بیماری دیابت نوع 2 پرداخته اند، روش طبقه بندی Naïve Bayes، SVM (Support Vector Machine) مورد بررسی قرار گرفت. در مطالعه مورد نظر از روش های پیش پردازش ترکیبی جهت جایگزینی مقادیر گمشده استفاده شد و در نهایت بهینه سازی انتخاب با استفاده از الگوریتم ژنتیک بر روی دیتاست Pima انجام گرفت (14). هووانگ و همکاران در مطالعه خود، با استفاده از الگوریتم های درخت تصمیم، جنگل تصادفی، ماشین بردار پشتیبان، شبکه بیزین به پیش بینی بیماری دیابت نوع 2 پرداخته اند و با استفاده از این روش به دقت بالای 60 درصد رسیدند (15).

بالا کریشان و همکاران نیز برای پیش بینی بیماری دیابت، از الگوریتم های Naïve Bayes و SVM استفاده کردند. آن ها با کار بر روی دیتاست pima و با استفاده از نرم افزار وکا، به بررسی دقت الگوریتم های داده کاوی پرداخته و یک روش انتخاب ویژگی برای پیدا کردن یک زیرمجموعه ویژگی مطلوب پیشنهاد دادند که دقت طبقه بندی naïve bayes و svm را بالا برده است. تحلیل های 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

روش پیشنهادی در نمودار شماره 2 در معیارهای دقت، حساسیت، ویژگی با سه تکنیک داده کاوی (رگرسیون، شبکه عصبی مصنوعی و در تصمیم گیری) مورد مقایسه قرار گرفته است. نتایج مقایسه شکل مورد نظر نشان می دهد که روش پیشنهادی در سه معیار ارزیابی دقت، ویژگی و حساسیت نسبت به تکنیک های رگرسیون، شبکه عصبی مصنوعی و درخت تصمیم گیری 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 داده های آموزشی، بیماری دیابت نوع 2 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 مطالعه ای که کاوور و همکاران به پیش بینی بیماری دیابت نوع 2 با استفاده از مجموعه داده pima پرداخته اند، نشان داده شد، سطح سرم انسولین مهم ترین ویژگی در افراد دیابتی است، اگر مقدار انسولین بالای 800 باشد نشان می دهد فرد مبتلا به دیابت است. براساس نتایج تمام الگوریتم ها به جز الگوریتم پیشنهادی

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 مدل سازی کمینه شد و استفاده از الگوریتم هوش دسته جمعی، باعث شد در هر تکرار متوسط کمینه خطای مدل سازی کاهش یابد. همچنین با افزایش تعداد ذرات به کار رفته در روش پیشنهادی شبکه عصبی مصنوعی دقیق تر آموزش یافت و وزن های بهینه تری نسبت به جمعیت های کم تر به دست 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 می رود مدل ارائه شده در این مطالعه در مراحل بعدی جهت استفاده در برنامه های غربالگری جهت تعیین میزان خطر ابتلا و پیشرفت بیماری مورد استفاده قرار گیرد. همچنین کمبود داده به منظور آموزش بیش تر مدل، داده های گمشده در نمونه ها از محدودیت های این مطالعه می باشد که به طور حتم مدل های ارائه شده در این مطالعه را می توان بهبود بخشید. هر چند مدل الگوریتم به دست آمده دارای دقت بالایی است اما با جمع آوری داده های جدید می توان مدل را مجدداً آموزش و دقت آن را افزایش داد (20).

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

در مطالعه ای که پاوار و همکاران به بررسی جامع در مورد تشخیص دیابت نوع 2 در بهداشت و درمان پرداختند، تحلیل های خود را در نمودار سطح زیر منحنی مقایسه کردند و نتایج نشان داد که الگوریتم SVM به دقت بالاتری برای تشخیص رسیده است (18).
لنگری زاده و همکاران در مطالعه خود، با استفاده از شبکه عصبی به پیش بینی تولد نوزاد نارس در مادران باردار پرداختند که نتایج نشان داد، استفاده از شبکه پرسپترون چند لایه برای پیش بینی نتیجه زایمان از نظر تولد نوزاد ترم یا نوزاد نارس در مادران باردار از طریق فناوری های کمک باروری می تواند در پیشگیری از عوارض تولد نوزاد نارس کمک کننده 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 پیش بینی دیابت را ارائه داد که متوسط خطای مدلسازی به عنوان تابع هدف بعد از یکسری تکرار

References

- Pardalos MP, Tomaino V, Xanthopoulos P. Optimization and data mining in medicine. Top 2009; 17(2): 215.
- Wager KA, Lee FW, Glaser JP. Health Care Information Systems: A Practical Approach for HealthCare Managment. 2th ed. San Francisco: Jossey-Bass Inc; 2009.
- Zimmet PZ. Diabetes epidemiology as a tool to trigger diabetes research and care. Diabetologia 1999; 42(5): 499-518.
- Tan J. Medical Informatics. Concepts, Methodologies, Tools and Applications. Hershey: IGI Global; 2008.
- Zimmet P. Globalization, coca-colonization and the chronic disease epidemic: can the Doomsday scenario be averted? J Intern Med 2000; 247(3): 301-310.
- 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.
- Mahdizadeh H, Barani A. Clinical Data Mining: An Overview of Data Mining Techniques in Diabetes. IJDL 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. Ioannis Kavakiotis, Olgha Tsave, Athanasios, Salifoglou, Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda. Machine Learning and Data Mining Methods in Diabetes Research. *Comput Struct Biotechnol J* 2017; 15: 104-116.
10. Pima diabetes. Available at: <http://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).