Journal of
Modern Medical Information Sciences

Summer 2022. Vol 8. Issue 2

Research Paper

Proposing a Model for Diagnosing the Type 2 Diabetes Using a Self-Organizing Genetic Algorithm

Fatemeh Ahouz[1], *Amin Golabpour[2], Abdolhosseain Shakibaeenia[3]

1. Department of Computer Engineering, School of Engineering, Behbahan Khatam Alanbia University of Technology, Behbahan, Iran.
2. Department of Health Informatics Technology, School of Allied Medical Sciences, Shahroud University of Medical Sciences, Shahroud, Iran.
3. Department of Computer Engineering, Dezful Branch, Islamic Azad University, Dezful, Iran.

Use your device to scan and read the article online

Citation Ahouz F, Golabpour A, Shakibaeenia A. [Proposing a Model for Diagnosing the Type 2 Diabetes Using a Self-Organizing Genetic Algorithm (Persian)]. Journal of Modern Medical Information Sciences. 2022; 8(2):168-183. https://doi.org/10.32598/JMIS.8.2.1

https://doi.org/10.32598/JMIS.8.2.1

CC BY NC

Article Info:

Received: 30 Jul 2021

Accepted: 08 May 2022

Available Online: 01 Jul 2022

Key words:
Knowledge discovery, Fuzzy logic, Medical informatics, Data mining, Classification

ABSTRACT

Objective Building clinical decision support models to automatically extract knowledge from data helps physicians in early diagnosis of disease. Interpretability of the diagnostic rules of these models for understanding how they make decisions and increasing confidence in their output is a key indicator in determining their efficacy.

Methods In this retrospective study, an automated hybrid rule extraction model is proposed for type 2 diabetes. In order to evaluate the model, the PIMA Diabetes dataset including 768 records and 9 variables was used. After removing the missing and outlier data in the data preprocessing stage, a proposed fuzzy-genetic hybrid model was implemented using MATLAB software to extract the rules. A self-organizing chromosomal structure was used to eliminate the complexity of setting genetic algorithm operators and facilitate the re-implementation of the model in other applications.

Results The accuracy of the proposed model on the PIMA dataset was 79.05%. This accuracy was obtained by two fuzzy rules, each of which contained only two independent variables. In addition, two single diagnostic rules for diabetic and non-diabetic individuals were presented with accuracy of 70.83% and 81.48%, respectively. The number of pregnancies, body mass index, diastolic blood pressure, diabetes pedigree function, plasma glucose concentration, and triceps skinfold thickness were the most effective factors in having or not having diabetes in the extracted rules.

Conclusion The proposed model with high accuracy and interpretability is quite suitable in producing an accurate and highly interpretable set of rules as well as single rules for diagnosing diabetes or absence of diabetes. Due to its self-organizing ability, it can also be used for other binary classification purposes.

* ***Corresponding Author:***
Amin Golabpour ,PhD.
Address: Department of Health Informatics Technology, School of Allied Medical Sciences, Shahroud University of Medical Sciences, Shahroud, Iran.
Tel: +98 (23) 32395054
E-mail: a.golabpour@shmu.ac.ir

Journal of
Modern Medical Information Sciences

Extended Abstract

Introduction

In recent years, fuzzy systems have been successfully used in different fields of science, especially medical sciences [1, 2]. They use linguistic rules to describe systems that are easily interpreted and checked by the user [3, 4]. Therefore, one of their applications of interest is in clinical decision support systems, where the discovery of rules hidden in the data and the interpretability of these rules are of great importance. The extraction of these rules with two indicators of accuracy and high interpretability helps specialists in increasing the accuracy and speed of disease diagnosis [5-7].

In the studies that have been conducted so far, attention to the extraction of single rules with high positive predictive value (PPV) and negative predictive value (NPV) has been neglected [8-10]. Due to the interest of specialists in diagnostic rules that are capable of quick evaluation and are easy to remember, and because many medical data sets, including the Pima diabetes dataset, contain clinical parameters resulting from tests, it is important to design a model for extracting single rules with high efficiency in terms of PPV and NPV. In this paper, we aim to propose a hybrid genetic-fuzzy classification system that automatically extracts the rules hidden in the data. Then, by evaluating each of the extracted rules, we provide the best single rule for diagnosis of disease and absence of disease. In addition, a new self-organizing chromosomal structure is proposed to eliminate the effect of the selection of genetic algorithm operators on the efficiency of the model. The Pima diabetes dataset was finally used to evaluate the proposed model.

Methods

In this study, the rule base of retrospective Mamdani fuzzy systems is designed using PIMA data. One of the problems of PIMA dataset is the presence of missing values and outliers. In this study, KNNi method was used to remove missing values and K-means was applied to remove outliers [2, 9]. At the beginning, there are no rules in the rule base and no membership functions are assigned to the fuzzy variables. They are generated and optimized using the genetic algorithm [2]. To eliminate the effect of the type of mutation and recombination operators on the efficiency of the model and reduce the time of setting the parameters of the genetic algorithm by trial and error, a new chromosomal structure was proposed which, while producing the fuzzy rule base, provides the best mutation and recombination operators among the existing methods for each dataset. In this proposed chromosomal structure, as generations pass and the results converge, the recombination operator corresponding to individuals with the highest fitness is selected more than others. In this way, the optimal recombination operator for the examined data set is automatically selected. The same process is done for the mutation operator. Then, each rule was evaluated on the dataset and their accuracy was measured. Finally, the best single rules with the highest PPV for detecting people with diabetes and the highest NPV for people without diabetes were determined. Afterwards, the diagnostic rules of having and not having the disease were combined and a set of two diagnostic rules was presented as the output of the model.

Results

The result of implementing the proposed model was the presentation of 81 single rules. Among the rules for diagnosing diabetes, those with PPV >70% were selected, and among the rules for diagnosing the absence of diabetes, those with NPV >80% were selected. Four single rules were finally determined which are listed in Table 1. Table 2 presents the result of combining the best single rules and the output of the proposed model.

Table 3 compares the efficiency of the proposed method with some existing methods in terms of the number of rules selected in the final rule base, the number of linguistic terms to describe each variable, the fuzzy mem-

Table 1. The best single diagnostic rules extracted by the proposed model on the PIMA diabetes dataset

Row	Rule	Classification Rate (%)
1	If Pregnancies is high and Tri Fold Thick is normal Then Diabetic	70.83
2	If Tri Fold Thick is high and BMI is normal and Then Diabetic	70.83
3	If Pregnancies is very low and PG Concentration is very low Then not-Diabetic	81.48
4	If Diastolic BP is very low and DP Function is very low Then not-Diabetic	81.48

Table 2. The set of rules extracted by the proposed model on the PIMA diabetes dataset

Row	Rule	%		
		Accuracy	PPV	NPV
1	If Pregnancies is high and Tri Fold Thick is normal Then Diabetic If Pregnancies is very low and PG Concentration is very low Then not-Diabetic	79.05	70.83	81.48
2	If Tri Fold Thick is high and BMI is normal and Then Diabetic If Pregnancies is very low and PG Concentration is very low Then not-Diabetic	79.05	70.83	81.48
3	If Tri Fold Thick is high and BMI is normal and Then Diabetic If Diastolic BP is very low and DP Function is very low Then not-Diabetic	79.05	70.83	81.48
4	If Pregnancies is high and Tri Fold Thick is normal Then Diabetic If Diastolic BP is very low and DP Function is very low Then not-Diabetic	79.05	70.83	81.48

bership function, the total number of conditions in the rules, the number of records used to build the model, and accuracy.

Discussion

In this study, due to the importance of extracting precise rules and their interpretability in clinical decision support system, a rule extraction model using a hybrid genetic-fuzzy algorithm with high accuracy was presented for determining the most compact set of rules and benchmarked on the PIMA diabetes dataset. In addition, Also, to avoid the complexity of setting genetic algorithm parameters and remove their effect on model efficiency, a new chromosomal structure with automatic adjustment of mutation and recombination operators was presented. The best set of rules according to the two criteria of interpretability and high accuracy, was the set with 2 rules, 4 sets of fuzzy terms per each antecedent, and an average length of 2 per each rule, which achieved 79.05% accuracy. The used membership function was a symmetric triangular function, which makes it easier for human users to understand the concepts due to the constant width of all functions.

Based on the rule length, number of rules, and membership functions, which are the most important indicators of the interpretability of rules by human users [10-13], the proposed model could provide an effective rule set with high interpretability to distinguish diabetic and non-diabetic people including only one diagnostic rule for diabetic people and one diagnostic rule for non-diabetic people.

Conclusion

Rule extraction from datasets in the field of medical diagnosis is an important area for knowledge discovery. Fuzzy systems are known as a popular tool in this field. For the automatic design of fuzzy systems from the data,

Table 3. Comparing the efficiency and interpretability of the proposed method with the methods presented in other studies on the PIMA diabetes dataset

References	Number of Rules	Number of Terms	Membership Function	Rule Length	Number of Records in the Database	Accuracy (%)
[2]	7	3	Symmetric triangle	15	590	77.12
[14]	3	7	-	72	732	78.03
[12]	5	5	Asymmetric trapezoid	6	768	85.10
[1]	3	4	Asymmetric Gaussian	6	768	77
[9]	1	-	-	10	590	79.32
[10]	53	8	Asymmetric triangle	-	768	73.83
[13]	19	5	Asymmetric triangle	91	768	84
Proposed model	2	4	Symmetric triangle	4	590	79.05

Journal of
Modern Medical Information Sciences

the genetic algorithm has shown a high ability. In this study, using a genetic-fuzzy hybrid method, a self-organizing rule extraction model was proposed for extracting diagnostic rules. The PIMA diabetes dataset was used to evaluate the proposed model. Based on this model, the best single diagnostic rules for people without and with diabetes were presented.

Based on the results, it can be concluded that the proposed model with accuracy of 79.05, PPV of 70.83 and NPV of 81.41% can be used as a promising general model in medical data classification and diagnosis. The small number of rules and their shortness are among the important features of the proposed rules, which can be easily evaluated and remembered by experts, along with quick implementation.

One of the limitations of the study was the use of public data and not evaluation of the proposed rules by clinical experts. It is suggested that the proposed model be implemented on local data and each of the extracted rules be evaluated by clinical experts.

Ethical Considerations

Compliance with ethical guidelines

In this study, no experiment on human or animal samples were conducted. Therefore, there as no need for obtaining ethical code.

Funding

This study was extracted from a research project (Code: 400-01-00) in Behbahan Khatam Alanbia University of Technology. It was not funded by any organization.

Authors' contributions

Conceptualization: Amin Golabpour; Writing: Fatemeh Ahouz; Methodology, model implementation and analysis: All authors

Conflicts of interest

The authors declared no conflict of interest.

This Page Intentionally Left Blank

مقاله پژوهشی

یک روش استخراج قانون تشخیصی دیابت نوع ۲با استفاده از الگوریتم ژنتیک خودسازمان‌دهنده

فاطمه آهوز[۱]، *امین گلاب‌پور[۲] ، عبدالحسین شکیبایی‌نیا[۳]

۱. کارشناسی ارشد هوش مصنوعی، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه صنعتی خاتم الانبیاء بهبهان، بهبهان، ایران
۲. دکترای تخصصی انفورماتیک پزشکی، گروه فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم‌پزشکی شاهرود، شاهرود، ایران.
۳. دانشجوی دکتری ، سیستم های نرم افزاری، گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه آزاد اسلامی، دزفول، ایران.

Use your device to scan and read the article online

Citation Ahouz F, Golabpour A, Shakibaeenia A. [Proposing a Model for Diagnosing the Type 2 Diabetes Using a Self-Organizing Genetic Algorithm (Persian)]. Journal of Modern Medical Information Sciences. 2022; 8(2):168-183. https://doi.org/10.32598/JMIS.8.2.1

https://doi.org/10.32598/JMIS.8.2.1

اطلاعات مقاله:
تاریخ دریافت: ۰۸ مرداد ۱۴۰۰
تاریخ پذیرش: ۱۸ اردیبهشت ۱۴۰۱
تاریخ انتشار: ۱۰ تیر ۱۴۰۱

چکیده

هدف ساخت مدل‌های کمک تصمیم‌یار پزشکی جهت استخراج خودکار دانش از داده‌ها به پزشکان در تشخیص زودهنگام بیماری کمک می‌کند. تفسیرپذیری قوانین اسستنتاجی این مدل‌ها جهت درک نحوه تصمیم‌گیری آن‌ها راجع به داده‌ها و افزایش اعتماد به خروجی مدل، یک شاخص ضروری در تعیین کارایی آن‌هاست.

روش‌ها در این پژوهش یک مطالعه گذشته‌نگر است جهت استخراج قوانین تشخیصی دیابت نوع ۲. داده‌های این مطالعه، مجموعه داده عمومی پیما شامل ۷۶۸ رکورد و ۹ ویژگی است که در سال ۱۴۰۰ استخراج شد. پس از حذف گمشدگی و داده‌های پرت در مرحله پیش‌پردازش داده‌ها، جهت استخراج قوانین یک مدل ترکیبی پیشنهادی فازی‌ژنتیک با استفاده از نرم‌افزار متلب ارائه شد. جهت حذف پیچیدگی تنظیم عملگرهای الگوریتم ژنتیک و تسهیل اجرای مجدد مدل در کاربردهای دیگر، ساختار کروموزومی خودسازمان‌دهنده‌ای پیشنهاد شده است.

یافته‌ها ارزیابی مدل پیشنهادی روی مجموعه داده پیما به صحت ۷۹/۰۵ درصد دست یافت. این صحت توسط ۲ قانون فازی که هرکدام فقط شامل ۲ متغیر مستقل است به دست آمده است. همچنین برای تشخیص افراد دارای دیابت و فاقد آن قانون‌های تشخیصی منفرد به ترتیب با صحت ۷۰/۸۳ و ۸۱/۴۸ درصد ارائه شده است. مهم‌ترین عوامل مؤثر بر ابتلا و عدم ابتلا به دیابت در این قوانین تعداد دفعات بارداری، شاخص توده بدنی، فشار خون، سابقه خانوادگی، غلظت گلوکز پلاسما و ضخامت پوست چین سه سر تعیین شدند.

نتیجه‌گیری روش پیشنهادی در تولید مجموعه و همچنین قوانین منفرد تشخیص بیماری یا عدم بیماری با صحت و قابلیت تفسیر بسیار بالا در کاربردهای پزشکی کاملاً مناسب بوده و به دلیل خودسازمانده بودن قابلیت تکرار در سایر کاربردهای دسته‌بندی دوکلاسی را دارد.

کلیدواژه‌ها:
استخراج دانش، منطق فازی، انفورماتیک پزشکی، داده‌کاوی، دسته‌بندی

* نویسنده مسئول:
دکتر امین گلاب‌پور
نشانی: شاهرود، دانشگاه علوم‌پزشکی شاهرود، دانشکده پیراپزشکی، گروه فناوری اطلاعات سلامت.
تلفن: ۳۲۳۹۵۰۵۴ (۲۳) ۹۸+
پست الکترونیکی: a.golabpour@shmu.ac.ir

مقدمه

در سال‌های اخیر، سیستم‌های فازی به‌طور موفقیت‌آمیزی در حیطه‌های متفاوت علوم، مانند مهندسی، علوم اجتماعی، علوم اقتصادی و به‌خصوص علوم‌پزشکی استفاده شده است [۱, ۲]. قوانین فازی به‌واسطه قابلیت خوانایی بسیار بالا، تفسیر آسان توسط انسان و فراهم کردن دیدگاهی راجع به دانش تعبیه‌شده در سیستم‌های دسته‌بندی، بین کاربران محبوبیت زیادی پیدا کرده‌اند [۳, ۴]. آن‌ها از قوانین زبانی برای توصیف سیستم‌ها استفاده می‌کنند که به‌آسانی کاربر تفسیر و بررسی‌شان می‌کند [۵,۶]. ازاین‌رو یکی از کاربردهای مورد توجه آن‌ها در سیستم‌های حمایتی تصمیم‌گیری پزشکی است. جایی که کشف قوانین نهفته در داده‌ها و تفسیرپذیری این قوانین دارای اهمیت بالایی است. استخراج قوانین مؤثر موجود در داده‌ها با ۲ شاخص صحت و تفسیرپذیری بالا در افزایش دقت و سرعت تشخیص بیماری به متخصصان کمک می‌کند [۷-۹].

یکی از مهم‌ترین عوامل در طراحی سیستم‌های فازی، تولید قوانین فازی است. برای این کار ۲ راهکار اصلی وجود دارد. در روش اول قوانین را فرد خبره تولید می‌کند. این روش به‌خصوص در مسائل کنترلی با تعداد ورودی کم به کار می‌رود [۵]. روش دوم، تولید خودکار قوانین با استفاده از تکنیک‌های نوروفازی، روش‌های خوشه‌بندی و الگوریتم‌های تکاملی است [۱۰، ۱۱].

بیماری دیابت احتمال مرگ زودرس را افزایش داده و علت اصلی بیماری‌های دیگر، ازجمله نارسایی کلیه، سکته قلبی و مغزی و نابینایی است [۱۲، ۱۳]. یکی از مجموعه داده‌های پرکاربرد در ارزیابی سیستم‌های تشخیصی دیابت، مجموعه داده دیابت پیما[1] است؛ دلیل انتخاب این مجموعه توسط محققان، در کنار چالش‌برانگیز بودن این مجموعه داده، به‌واسطه وجود مقادیر گمشده و داده‌های پرت، شیوع بسیار بالا و آمار خاص بیماری دیابت در میان سرخپوستان پیما آرزونا است [۱۴]. در این بخش برخی از کارهای صورت‌گرفته روی این مجموعه داده بررسی می‌شود.

در سال ۲۰۱۹ یک پژوهش با استفاده از مدل ترکیبی فازی‌ژنتیک جهت دسته‌بندی دیابت روی دیتاست پیما ارائه شده است. نویسندگان از الگوریتم ژنتیک جهت تعیین پارامترهای بهینه این تابع و تولید قوانین استفاده کرده‌اند. صحت مدل پیشنهادی را با استفاده از روش اجرای اعتبارسنجی متقاطع، ۵ فولد برابر با ۸۵/۸۲ درصد گزارش دادند. در این مقاله تعداد قوانین تولیدشده و ویژگی‌های منتخب در قوانین ذکر نشده است [۱۵].

1. PIMA

در دسته‌بندی فازی مبتنی بر قانون[2] از داده‌های پزشکی ارائه شده است. جهت تولید مجموعه قوانین بهینه، یافتن بهترین پارامترها برای توابع عضویت و انتخاب متغیرها از الگوریتم بهینه‌سازی چندهدفه NSGA II[3] استفاده شده است. نویسندگان نوآوری مقاله را در معرفی معیارهایی، جهت اندازه‌گیری تفسیرپذیری سیستم‌های دسته‌بندی فازی، نمایشی مستقل از کدگذاری خاص از پایگاه قوانین و ارائه عملگرهای ژنتیکی برای پردازش نمایش ارائه‌شده عنوان کردند. بهترین مقدار گزارش‌شده از اجرای مدل به ازای تقسیم مجموعه داده به‌صورت ۱:۹ دارای صحت ۸۱/۵ درصد شامل ۵ قانون، ۳ ویژگی و ۵ مجموعه فازی (ترم‌های زبانی) و ۱/۲ ویژگی به ازای هر قانون است [۳].

در میان کارهای صورت‌گرفته، در توجه به استخراج قوانین منفرد با ارزش اخباری مثبت[4] و منفی[5] بالا غفلت شده است [۱۴-۱۶]. با توجه به علاقه متخصصان به قوانین ساده و منفرد تشخیصی که قابلیت ارزیابی سریع و سادگی در به خاطر سپردن را دارند و همچنین از آنجا که بسیاری از مجموعه داده‌های پزشکی، ازجمله مجموعه داده دیابت پیما شامل پارامترهای بالینی حاصل انجام آزمایش است، طراحی مدلی جهت استخراج قوانین منفرد با کارایی بالا برحسب ارزش اخباری مثبت و منفی امری ضروری است. ازاین‌رو در این مقاله، یک سیستم دسته‌بندی ترکیبی ژنتیک‌فازی ارائه می‌شود که به‌طور خودکار از روی داده‌ها قوانین نهفته در آن‌ها را استخراج می‌کند. سپس با ارزیابی هریک از قوانین استخراج‌شده، بهترین قانون منفرد تشخیصی برای هریک از حالت‌های بیماری و عدم بیماری را ارائه می‌دهد. به‌علاوه، جهت حذف تأثیر انتخاب عملگرهای الگوریتم ژنتیک روی کارایی مدل، یک ساختار کروموزومی خودسازمان‌دهنده جدید پیشنهاد شده است. جهت ارزیابی مدل پیشنهادی از مجموعه داده دیابت پیما استفاده شده است.

مواد و روش‌ها

در این مطالعه، طراحی پایگاه قانون سیستم‌های فازی ممدانی گذشته‌نگر با استفاده از داده‌های داده عمومی پیما بررسی می‌شود. محل انجام پژوهش دانشگاه صنعتی خاتم‌الانبیای بهبهان در سال ۱۴۰۰ بود. طراحی سیستم‌های منطق فازی یکی از محبوب‌ترین مسائل رام‌نشدنی است که در آن از محبوب‌ترین و جدیدترین الگوریتم‌های تکاملی استفاده می‌شود [۱۷]. این مسئله شامل پارامترهایی، ازجمله شکل و مکان مجموعه‌های فازی، مقدم‌ها و تالی‌های پایگاه قانون فازی و سایر پارامترهای استراتژی مانند تجمیع، استلزام و روش‌های دیفازی‌سازی می‌شود. در این مقاله، طراحی پایگاه قانون سیستم‌های فازی ممدانی از داده‌ها بررسی

2. Fuzzy Rule-Based Classification Systems (FRBCSs)
3. Non-dominated Sorting Genetic Algorithm II
4. Positive Predictive Value (PPV)
5. Negative Predictive Value (NPV)

می‌شود. تصویر شماره ۱ توابع عضویت مورد استفاده جهت توصیف متغیرهای مجموعه داده را نشان می‌دهد. در ابتدای الگوریتم هیچ قانونی در پایگاه قوانین و هیچ تابع عضویتی منتسب به متغیرهای فازی نیست. سپس، قوانین و توابع عضویت ازطریق الگوریتم ژنتیک تولید و بهینه می‌شوند [۲]. تصویر شماره ۲ فلوچارت مدل پیشنهادی را نشان می‌دهد.

الگوریتم ژنتیک[۶]

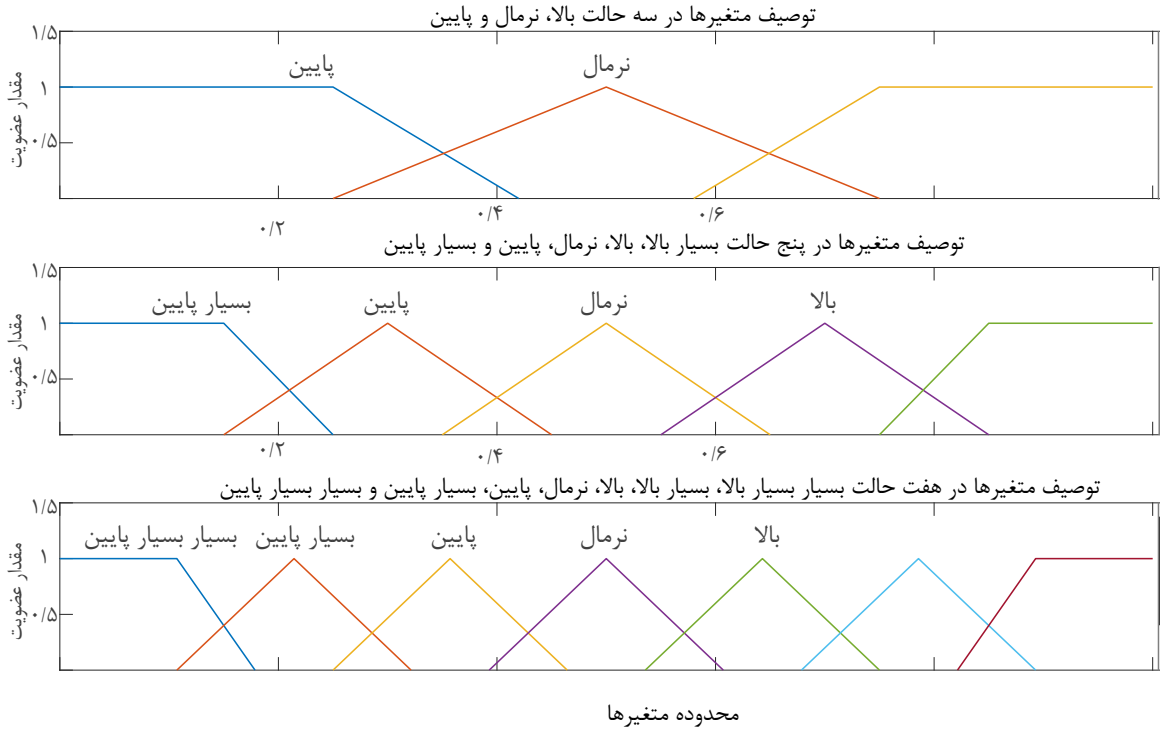
الگوریتم ژنتیک محبوب‌ترین الگوریتم بهینه‌سازی متاهیوریستیک مبتنی بر تکامل است [۱۷]. در حوزه‌های متفاوت، الگوریتم ژنتیک برای حل مسائل بهینه‌سازی چندهدفه و چندمتغیره کارایی مناسبی از خود نشان داده است [۱۸]. از الگوریتم ژنتیک برای انتخاب قانون استفاده می‌شود زیرمجموعه‌ای از قوانین اولیه که نرخ دسته‌بندی بالایی دارند، انتخاب شوند [۱۹]. الگوریتم ژنتیک دارای چند گام اساسی است: تعریف ساختار کروموزومی، تولید جمعیت اولیه، تعیین تابع هدف، انتخاب، بازترکیب و جهش [۱۶].

مدل پیشنهادی

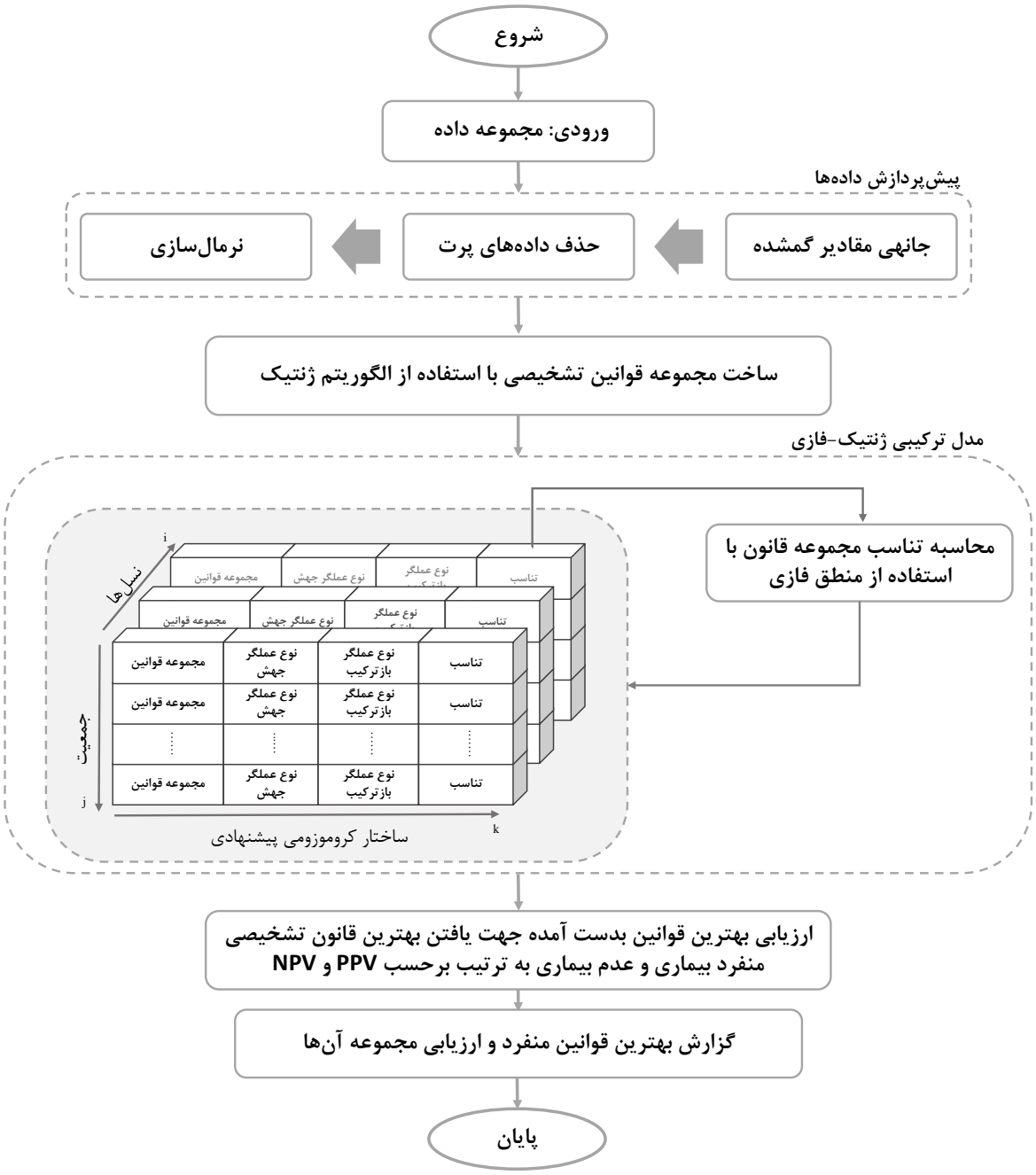
در این مقاله یک سیستم دسته‌بندی ترکیبی ژنتیک‌فازی با هدف دست‌یابی به صحت بالا و تولید مجموعه قوانین فشرده و با تفسیرپذیری بسیار بالا طراحی شده است. مدل پیشنهادی از مجموعه داده دیابت پیما استفاده شده است. این مجموعه دارای ۷۶۸ رکورد شامل ۲۶۸ فرد دارای دیابت و ۵۰۰ فرد فاقد دیابت است. برای هر رکود ۹ ویژگی شامل برچسب کلاس ذخیره شده است. یکی از مشکلات این مجموعه داده وجود مقادیر گمشده و داده‌های پرت است. در این برای جانهی مقادیر گمشده از روش KNNi و حذف داده‌های پرت از K-means استفاده شد [۱۵ , ۲]. پس از پیش‌پردازش تعداد رکودها به ۵۹۰ رکورد کاهش یافت که شامل ۲۱۲ فرد دارای دیابت و ۳۷۸ فرد سالم بود. جهت تعیین ارتباط بین متغیرهای وابسته و مستقل از تست آماری تی استفاده شد. در آزمون آماری تی برای بررسی رابطه معناداری متغیرها و داشتن دیابت استفاده و مقدار ۰/۰۵ درصد به‌عنوان سطح معناداری در نظر گرفته شد

در این پژوهش بهترین مجموعه قوانین تشخیصی برای دیابت توسط الگوریتم ژنتیک استخراج می‌شود. تک‌تک قوانین روی مجموعه داده ارزیابی و صحت آن‌ها محاسبه می‌شود. سپس بهترین قوانین منفرد با بالاترین ارزش اخباری مثبت برای تشخیص افراد دارای دیابت و بالاترین ارزش اخباری منفی برای افراد فاقد دیابت تعیین می‌شود. درنهایت پس از تعیین بهترین قوانین منفرد، قوانین تشخیصی ابتلا و عدم ابتلا به بیماری ترکیب و مجموعه‌ای شامل ۲ قانون تشخیصی به‌عنوان خروجی مدل ارائه می‌شود.

6. Genetic Algorithm (GA)



تصویر ۱. توابع عضویت فازی توصیف‌کننده ویژگی‌های مجموعه داده‌ها



تصویر ۲. فلوچارت مدل پیشنهادی جهت استخراج قوانین تشخیصی بیماری و عدم بیماری.

ازآنجاکه تولید مجموعه قوانین فازی یک مسئله رام‌نشدنی[7] است، امکان تولید تمام قوانین و سپس انتخاب از میان آن‌ها امکان‌پذیر نیست. اگر تعداد قوانین را ۲ در نظر بگیریم و فرض کنیم هر ویژگی تنها با ۳ مجموعه فازی و یک حالت بی‌اثر نشان داده شود، برای مجموعه داده پیما با ۸ ویژگی در هر قانون و ۱ برچسب کلاس دوحالته، ۲۸۲×۱۶۸۴ حالت متفاوت وجود خواهد داشت. مدت‌زمان اجرای داخلی‌ترین حلقه در شبیه‌سازی کامپیوتری ۰/۰۰۳۲ ثانیه است که با توجه به تعداد دفعات تکرار ۶۳۶ روز به طول خواهد انجامید. به همین دلیل جهت تولید و انتخاب مجموعه قوانین بهینه فازی از الگوریتم ژنتیک استفاده شده است. در مدل پیشنهادی، تمام مجموعه داده به‌عنوان مجموعه آموزش استفاده می‌شود، زیرا مجموعه داده استاندارد بوده و در سیستم‌های مبتنی بر قانون مجاز به تقسیم مجموعه داده به‌صورت آموزش و تست نیستیم. همچنین نمی‌توان از اجرای اعتبارسنجی متقاطع استفاده کرد، زیرا هر بار مجموعه قوانین جدیدی تولید خواهد شد.

7. Non-deterministic Polynomial (NP)

پارامترهای الگوریتم ژنتیک

یکی از مؤلفه‌های الگوریتم ژنتیک عملگرهای بازترکیب و جهش هستند که تعیین نوع این عملگرها اغلب به‌صورت سعی و خطا انجام می‌شود. انتخاب این ۲ عملگر تأثیر بسیاری روی نتیجه خواهد داشت، زیرا رفتار الگوریتم ژنتیک با روابط اکتشافی و بهره‌وری تعیین می‌شود که به ترتیب متأثر از عملگرهای جهش و بازترکیب است [۵]. در این مقاله، جهت حذف تأثیر نوع عملگرهای جهش و بازترکیب بر کارایی مدل و کاهش زمان تنظیم پارامترهای الگوریتم ژنتیک به روش سعی و خطا، یک ساختار کروموزومی جدید پیشنهاد می‌شود که ضمن تولید پایگاه قانون فازی، بهترین عملگر جهش و بازترکیب از میان روش‌های موجود را برای هر مجموعه داده تعیین می‌کند. ایده کار به این صورت است که در جمعیت اولیه به‌طور تصادفی به هر کروموزوم یک عملگر بازترکیب و جهش انتساب داده می‌شود. سپس زمانی که کروموزوم‌ها توسط عملگر انتخاب برای عمل ترکیب انتخاب می‌شوند، عملگر بازترکیبی که مربوط به والدی با بالاترین تناسب است برای تولید فرزندان اعمال می‌شود. با گذشت نسل‌ها و همگرایی نتایج، عملگر بازترکیب مربوط به افراد، با بالاترین تناسب بیشتر از سایرین انتخاب می‌شود. به این ترتیب عملگر بازترکیب بهینه برای مجموعه داده موردبررسی به‌طور خودکار انتخاب می‌شود. همین روند برای عملگر جهش نیز صادق است.

ساختار کروموزومی در تصویر شماره ۳ نشان داده شده است. این کروموزوم شامل ژن‌هایی برای تعیین ترم‌های زبانی برای هر متغیر در مجموعه داده، ژنی برای تعیین نوع عملگر جهش، یک ژن برای تعیین نوع عملگر بازترکیب و یک ژن برای مقدار تناسب کروموزوم است. اندازه کروموزوم برحسب تعداد قوانین و تعداد ویژگی‌های موجود در مجموعه داده متفاوت است. اگر تعداد قوانین موردنظر در پایگاه قانون فازی r باشد و مجموعه داده دارای n ویژگی باشد، اندازه کروموزوم از فرمول شماره ۱ به دست خواهد آمد.

1. $chroSize=r\times n+3$

محدوده ژن‌های 1 الی chroSize-2 مجموعه قوانین پایگاه داده را مشخص می‌کنند که در دسته‌های n تایی تقسیم شده‌اند. مقادیر این ژن‌ها مجموعه‌های فازی برای ترم‌های زبانی متغیرها را تعیین می‌کنند. به‌طور نظری هر متغیر فازی می‌تواند تعداد زیادی مجموعه فازی داشته باشد و هرکدام تابع عضویت خودشان را داشته باشند، اما معمولاً برای هر متغیر فازی، ۳، ۵، ۷ یا ۹ مجموعه استفاده می‌شود [۵]. با داشتن شماره ژن، تفسیر مقدار ژن مدنظر در کروموزوم که با Gene(k) نشان داده می‌شود، با فرمول شماره ۲ مشخص می‌شود.

$$2.\ Gene(k)=\begin{cases} fitnessValue \\ crossoverType \\ mutationeType \\ yi,\ where\ J=k/n \\ Ai,\ m,\ where\ J=k/n+1,\ m=k-(J-1)\times n \end{cases}$$

if k= chroSize

if k= chroSize-1

if k= chroSize-2

if k is divisable by n otherwise

Ai, m, where J=k/n+1, m= k-(J-1)×n

که در آن yl معرف تالی قانون Lام و Al, m مجموعه فازی برای ویژگی mام از قانون ام را مشخص می‌کند. الگوریتم ژنتیک ازطریق تکرار مراحل در میان نسل‌ها راه‌حل بهینه را می‌یابد. هر نسل شامل چندین کروموزوم است.

جمعیت اولیه ۳۰۰ فرد و شرط توقف الگوریتم رسیدن به تعداد نسل ۱۰۰۰ انتخاب شده است (این مقدار در منابعی ذکر شده است [۱، ۸]. جهت کاهش تأثیر انتخاب‌های تصادفی، هر اجرای الگوریتم ۱۰ دور تکرار می‌شود. نرخ بازترکیب ۰/۸ و نرخ جهش ۰/۵ انتخاب شده است. تناسب هر فرد برابر با صحت قانون فازی موجود در کروموزوم است که توسط سیستم خبره فازی محاسبه می‌شود. جمعیت اولیه به‌صورت تصادفی‌یکنواخت ایجاد می‌شود.

پس از ایجاد هر فرد، اعتبار کروموزوم بررسی می‌شود و کروموزوم هایی که دارای برچسب کلاسی یکسان در تمام قوانین هستند، به‌عنوان نامعتبر شناخته می‌شوند. در صورت یکی بودن تمام برچسب‌ها به‌طور تصادفی یکی از آن‌ها تغییر می‌کند. همچنین اگر حالت تمام ویژگی‌ها «بی‌اثر» باشد نیز کروموزوم نامعتبر است. در صورتی که کمتر از ۲ ویژگی دارای حالت غیربی‌اثر باشد، به‌طور تصادفی الگوریتم مقدار یکی از ویژگی‌های بی‌اثر را به حالتی به‌جز حالت بی‌اثر تغییر می‌دهد.

عملگر بازترکیب از میان ۱۲ عملگر شامل بازترکیب سه‌والدی، گسسته، اکتشافی، هموار، میانگین، نیمه‌یکنواخت، یکنواخت اکتشافی، تک‌نقطه‌ای، دونقطه‌ای، یکنواخت، درهم‌ریزی و جایگزین کاهش‌یافته و عملگر جهش از میان ۸ عملگر شامل جهش مکمل، یکنواخت، درجی، معکوس‌سازی مرکزی، کریپ، جابه‌جایی، تراورز و عملگر جهش تروا‌س انتخاب شده است. این عملگرها متداول‌ترین عملگرهایی مورداستفاده در کاربردهای ژنتیک هستند [۱۵]. تابع انتخاب در این مقاله تابع‌های رنک‌بندی و تصادفی است. تابع رنک‌بندی به نحوی پیاده‌سازی شده است که در صورت برابر بودن تناسب افراد، اولویت با فردی باشد که در مجموعه قوانین خود تعداد حالت‌های «بی‌اثر» بیشتری داشته باشد. در ابتدا ۶۰ درصد افراد با تابع تصادفی و ۴۰ درصد

تصویر ۳. ساختار کروموزوم پیشنهادی.

باقی‌مانده توسط تابع رنک‌بندی انتخاب می‌شوند. با گذشت نسل‌ها، میزان افرادی که به‌طور تصادفی برای عملیات جهش یا بازترکیب انتخاب می‌شوند هر ۱۰ نسل به میزان ۵ درصد کاهش می‌یابد تا درصد انتخاب با تابع رنک‌بندی افزایش یابد تا به این ترتیب شانس انتخاب افرادی که صحت بالاتری دارند بیشتر شود. اما این مقدار هیچ‌گاه زیر ۰/۱ نخواهد شد.

پیاده‌سازی روش پیشنهادی در نرم‌افزار متلب ۲۰۱۷ و تحلیل آماری توسط نرم‌افزار SPSS نسخه ۱۶ اجرا شده است.

مقایسه کارایی الگوریتم و روش‌های موجود

به‌دلیل اینکه تمامی کارهای موردبررسی در این مقاله از تمام ۷۶۸ رکورد مجموعه داده پیما استفاده کردند و وجود داده‌های پرت را نادیده گرفتند و امکان پیاده‌سازی روش‌های ذکرشده به دلیل مشخص نبودن تمام پارامترها یا عدم دسترسی به سخت افزارهای ذکرشده در آن مقالات میسر نیست، شرایط اجرای الگوریتم‌ها متفاوت بوده و امکان مقایسه صحت سیستم‌ها فراهم نیست. اما مجموعه قوانین به‌دست‌آمده به‌لحاظ قابلیت تفسیرپذیری با مجموعه قوانین به‌دست‌آمده در این مقاله مقایسه می‌شود.

یافته‌ها

جدول شماره ۱ توصیف متغیرهای مجموعه داده را برحسب آماره‌های توصیفی میانگین، انحراف معیار و بازه تغییرات در ۲ گروه افراد فاقد و دارای دیابت نشان می‌دهد. همچنین نتایج آزمون آماری تی برای بررسی رابطه معناداری متغیرها و داشتن دیابت، استفاده و مقدار ۰/۰۵ درصد به‌عنوان سطح معناداری در نظر گرفته شد.

بهترین مجموعه قوانین به‌دست‌آمده حاصل از اجرای روش پیشنهادی، در حالتی که تعداد مجموعه قوانین از ۲ تا ۷ قانون متغیر است و برای هر ویژگی مستقل در مجموعه داده به ترتیب ۳، ۵ و ۷ ترم زبانی در نظر گرفته شده است، در جدول شماره ۲ نشان داده شده است.

پس از به دست آمدن بهترین مجموعه قوانین، تمام ۸۱ قانون منفرد منتخب توسط الگوریتم پیشنهادی، به‌صورت منفرد روی مجموعه داده پیما ارزیابی و صحت هر قانون محاسبه شد. سپس در میان قوانینی که جهت تشخیص ابتلا به دیابت بودند، قوانین با ارزش اخباری مثبت بالاتر از ۷۰ درصد و در میان قوانینی که جهت تشخیص عدم ابتلا به دیابت بودند قوانین با ارزش اخباری منفی بالاتر از ۸۰ درصد انتخاب شدند. نتیجه این بررسی ۴ قانون ذکرشده در جدول ۳ است.

جدول ۱. اطلاعات آماری مجموعه داده پیما بعد از پیش‌پردازش [۲] (Sig= ۰/۰۰)

ردیف	نام ویژگی	مخفف	میانگین±انحراف معیار		بازه مقادیر (حداقل–حداکثر)	
			دارای دیابت	فاقد دیابت	دارای دیابت	فاقد دیابت
۱	تعداد دفعات بارداری	Prg	۳/۶۰±۴/۹۴	۲/۹۹±۳/۲۷	(۰–۱۵)	(۱۳–۰)
۲	غلظت گلوکز پلاسما ۲ ساعته	PG	۱۲۵/۷۶±۳۷/۰۸	۲۸/۱۰±۸۴/۸۰	(۱۹۹–۴۴)	(۱۹۳–۲۳)
۳	فشار خون دیاستولیک (میلی‌متر جیوه)	BP	۱۲/۶۲±۶۳/۲۸	۱۲/۵۲±۵۷/۸۸	(۱۰۲–۳۰)	(۱۰۲–۲۵)
۴	ضخامت پوست چین سه سر (میلی‌متر)	TrFT	۷/۴۲±۲۶/۸۷	۸/۸۷±۲۱/۷۹	(۵۳–۸)	(۵۱–۱)
۵	انسولین سرم	SrIns	۷۸/۳۰±۱۸۱/۰۱	۶۳/۵۵±۱۱۲/۱۹	(۵۷۵–۲۲)	(۴۷۹–۱)
۶	شاخص توده بدنی (وزن در کیلوگرم / (قد در متر)۲)	BMI	۸/۰۷±۲۳/۰۰	۸/۲۸±۱۷/۲۴	(۴۸–۷)	(۴۱–۰)
۷	سابقه خانوادگی دیابت نوع ۲	DP	۰/۳۱±۰/۵۱	۰/۲۵±۰/۴	(۱/۸۹ –۰/۰۹)	(۱/۷۳ –۰/۰۸)
۸	سن (سال)	Age	۱۰/۹۱±۳۶/۸۳	۱۱/۴۵±۳۰/۶۷	(۷۰–۲۱)	(۷۲–۲۱)

جدول ۲. بهترین مجموعه قوانین بدست آمده با تعداد قوانین و تعداد ترم‌های توصیفی متفاوت مدل پیشنهادی روی مجموعه داده دیابت پیما

قوانین	حالات: ۳	۵ حالته	۷ حالته
دو قانون	- اگر شاخص توده بدنی و سابقه خانوادگی دیابت نوع ۲ پایین باشند(سالم). - اگر فشار خون و شاخص توده بدنی بالا باشند(بیمار).	- اگر فشار خون و سابقه خانوادگی در دیابت نوع دو پایین باشد(سالم). - اگر تعداد دفعات بارداری بسیار بالا و سن پایین باشد(بیمار).	- اگر فشار خون و شاخص توده بدنی بسیار بسیار پایین باشند(سالم). - اگر غلظت گلوکز پلاسما و سابقه خانوادگی دیابت نوع دو بسیار بسیار بالا باشند(بیمار)
سه قانون	- اگر ضخامت پوست چین سه‌سر و انسولین سرم پایین باشند(سالم). - اگر ضخامت پوست چین سه‌سر و شاخص توده بدنی نرمال باشند(سالم). - اگر سابقه خانوادگی در دیابت نوع دو و سن بالا باشند(بیمار).	- اگر ضخامت پوست چین سه‌سر و سن بسیار پایین باشند(سالم). - اگر ضخامت پوست چین سه‌سر و سن پایین باشند(سالم). - اگر تعداد دفعات بارداری بسیار پایین و غلظت گلوکز پلاسما بسیار بالا باشد(بیمار)	- اگر ضخامت پوست چین سه‌سر و سن بسیار بسیار پایین باشند(سالم). - اگر تعداد دفعات بارداری و ضخامت پوست چین سه‌سر بسیار پایین باشند(سالم). - اگر فشار خون بسیار بالا و ضخامت پوست چین سه‌سر بالا باشد. بیمار)
چهار قانون	- اگر فشار خون و ضخامت پوست چین سه‌سر پایین باشند(سالم). - اگر فشار خون بالا و سن پایین باشد(سالم). - اگر تعداد دفعات بارداری و فشار خون پایین باشند(سالم). اگر غلظت گلوکز پلاسما نرمال و سابقه خانوادگی دیابت نوع دو بالا باشند(بیمار).	- اگر تعداد دفعات بارداری و انسولین سرم پایین باشند(سالم) - اگر تعداد دفعات بارداری نرمال و غلظت گلوکز پلاسما بالا باشد(سالم) - اگر انسولین سرم و شاخص توده بدنی پایین باشند(سالم) - گر غلظت گلوکز پلاسما و شاخص توده بدنی بالا باشند(بیمار).	- اگر شاخص توده بدنی و سن هر دو بسیار بسیار پایین باشند(سالم). - اگر غلظت گلوکز پلاسما و سابقه خانوادگی دیابت نوع دو بسیار پایین باشند(سالم). - اگر غلظت گلوکز پلاسما و سابقه خانوادگی دیابت نوع دو پایین باشند(سالم). - اگر فشار خون بسیار پایین و شاخص توده بدنی نرمال باشد(بیمار)
پنج قانون	- اگر شاخص توده بدنی و سابقه خانوادگی در دیابت نوع دو بالا باشند(سالم) - اگر غلظت گلوکز پلاسما و سابقه خانوادگی در دیابت نوع دو پایین باشند(سالم) - اگر غلظت گلوکز پلاسما و ضخامت پوست چین سه‌سر نرمال باشند(سالم) - اگر غلظت گلوکز پلاسما و انسولین سرم پایین باشند(سالم) - اگر غلظت گلوکز پلاسما بالا و فشار خون نرمال باشد(بیمار).	- اگر تعداد دفعات بارداری و سن پایین باشند(سالم). - اگر فشار خون و شاخص توده بدنی پایین باشند(سالم). اگر تعداد دفعات بارداری و انسولین سرم بسیار پایین باشند(سالم). اگر تعداد دفعات بارداری بالا و ضخامت پوست چین سه‌سر نرمال باشد(بیمار). اگر غلظت گلوکز پلاسما و فشار خون بسیار بالا باشند(بیمار).	- اگر تعداد دفعات بارداری و فشار خون بسیار بسیار پایین باشند(سالم). - اگر انسولین سرم و سن بسیار پایین باشند(سالم). - اگر تعداد دفعات بادارى پایین و غلظت گلوکز پلاسما نرمال باشد(سالم). - اگر غلظت گلوکز پلاسما و انسولین سرم بسیار بسیار پایین باشند(سالم). - اگر تعداد دفعات بارداری و غلظت گلوکز پلاسما بسیار بالا باشند(بیمار).
شش قانون	- اگر غلظت گلوکز پلاسما و انسولین سرم پایین باشند(سالم). - اگر تعداد دفعات بارداری بالا و سابقه خانوادگی دیابت نوع دو پایین باشد(سالم) - اگر ضخامت پوست چین سه‌سر و سن پایین باشند(سالم) اگر فشار خون نرمال و سابقه خانوادگی دیابت نوع دو پایین باشد(سالم) - اگر شاخص توده بدنی نرمال و سابقه خانوادگی دیابت نوع دو بالا باشد(بیمار) - اگر انسولین سرم و شاخص توده بدنی بالا باشند(بیمار)	- اگر غلظت گلوکز پلاسما بسیار پایین و سابقه خانوادگی در دیابت نوع دو پایین باشد(سالم). - اگر تعداد دفعات بارداری پایین و غلظت گلوکز پلاسما نرمال باشد(سالم). - اگر فشار خون و شاخص توده بدنی بسیار پایین باشند(سالم). - اگر ضخامت پوست چین سه‌سر و سابقه خانوادگی در دیابت نوع دو بسیار پایین باشند(سالم). - اگر ضخامت پوست چین سه‌سر بالا و شاخص توده بدنی نرمال باشد(بیمار). اگر فشار خون و ضخامت پوست چین سه‌سر بسیار بالا باشد(بیمار).	- اگر فشار خون و سن بسیار بسیار پایین باشند(سالم) - اگر فشار خون بسیار بالا باشد و سن پایین باشد(سالم) - اگر ضخامت پوست چین سه‌سر و شاخص توده بدنی بسیار بسیار پایین باشند(سالم). -اگر ضخامت پوست چین سه‌سر و سن بسیار پایین باشند (سالم). اگر تعداد دفعات بارداری و ضخامت پوست چین سه‌سر بسیار بالا باشد(بیمار). اگر شاخص توده بدنی بسیار بسیار پایین باشد و سن بالا باشد(بیمار).
هفت قانون	- اگر غلظت گلوکز پلاسما و ضخامت پوست چین سه‌سر پایین و سابقه خانوادگی در دیابت نوع دو بالا باشد(سالم). - اگر شاخص توده بدنی نرمال و سن پایین باشد(سالم). - اگر تعداد دفعات بارداری و سابقه خانوادگی در دیابت نوع دو پایین باشند(سالم). - اگر فشار خون و ضخامت پوست چین سه‌سر پایین باشند(سالم). - اگر فشار خون پایین و انسولین سرم بالا باشد(بیمار). - اگر تعداد دفعات بارداری و انسولین سرم بالا باشند(بیمار). - اگر تعداد دفعات بارداری و ضخامت پوست چین سه‌سر نرمال باشند(بیمار).	- اگر غلظت گلوکز پلاسما بالا باشد و شاخص توده بدنی بسیار پایین باشد(سالم). - اگر سابقه خانوادگی در دیابت نوع دو بسیار پایین باشد و سن پایین باشد(سالم). - اگر تعداد دفعات بارداری و سابقه خانوادگی در دیابت نوع دو بسیار پایین باشند(سالم). - اگر انسولین سرم و سن بسیار پایین باشند(سالم). - اگر غلظت گلوکز پلاسما بسیار پایین، فشار خون نرمال و سابقه خانوادگی در دیابت نوع دو بسیار بالا باشد(سالم). - اگر شاخص توده بدنی و سابقه خانوادگی در دیابت نوع دو بسیار بالا باشند(بیمار). - اگر تعداد دفعات بارداری پایین باشد و انسولین سرم بسیار بالا باشد(بیمار).	- اگر تعداد دفعات بارداری بسیار بسیار پایین و غلظت گلوکز پلاسما بسیار پایین باشد(سالم). - اگر سابقه خانوادگی در دیابت نوع دو و سن بسیار پایین باشند(سالم). - فشار خون بسیار پایین و سابقه خانوادگی در دیابت نوع دو بسیار بسیار پایین باشد(سالم). - اگر تعداد دفعات بارداری و شاخص توده بدنی بسیار پایین باشند(سالم). - اگر ضخامت پوست چین سه‌سر، شاخص توده بدنی و سن بسیار پایین باشند(سالم). - اگر تعداد دفعات بارداری بسیار پایین و انسولین سرم نرمال باشد(بیمار). - اگر غلظت گلوکز پلاسما بسیار پایین و انسولین سرم بسیار بسیار بالا باشد(بیمار).

جدول ۳. بهترین قوانین تشخیصی منفرد روی مجموعه داده دیابت پیما

ردیف	قانون	نرخ دسته‌بندی
۱	اگر تعداد دفعات بارداری بالا و ضخامت پوست چین سه سر نرمال باشد آن‌گاه فرد مبتلا است	۷۰/۸۳ درصد
۲	اگر ضخامت پوست چین سه سر بالا و شاخص توده بدنی نرمال باشد، آن‌گاه فرد مبتلا است	۷۰/۸۳ درصد
۳	اگر تعداد دفعات بارداری بسیار بسیار پایین و غلظت گلوکز پلاسما بسیار پایین باشد، آن‌گاه فرد مبتلا نیست	۸۱/۴۸ درصد
۴	فشار خون بسیار پایین و سابقه خانوادگی دیابت نوع ۲ بسیار بسیار پایین باشد، آن‌گاه فرد مبتلا به دیابت نیست	۸۱/۴۸ درصد

جدول۴. مجموعه قوانین تشخیصی مدل پیشنهادی روی مجموعه داده دیابت پیما.

ردیف	قانون	درصد		
		صحت	ارزش اخباری مثبت	ارزش اخباری منفی
۱	اگر تعداد دفعات بارداری بالا و ضخامت پوست سه سر نرمال باشد، آن‌گاه فرد مبتلا به دیابت است اگر تعداد دفعات بارداری بسیار بسیار پایین و غلظت گلوکز پلاسما بسیار پایین باشد، آن‌گاه فرد مبتلا به دیابت نیست	۷۹/۰۵	۷۰/۸۳	۸۱/۴۸
۲	اگر ضخامت پوست چین سه سر بالا و شاخص توده بدنی نرمال باشد، آن‌گاه فرد مبتلا است اگر تعداد دفعات بارداری بسیار بسیار پایین و غلظت گلوکز پلاسما بسیار پایین باشد، آن‌گاه فرد مبتلا به دیابت نیست	۷۹/۰۵	۷۰/۸۳	۸۱/۴۸
۳	اگر ضخامت پوست چین سه سر بالا و شاخص توده بدنی نرمال باشد، آن‌گاه فرد مبتلا است اگر فشار خون بسیار پایین و سابقه خانوادگی دیابت نوع ۲ بسیار بسیار پایین باشد، آن‌گاه فرد مبتلا به دیابت نیست	۷۹/۰۵	۷۰/۸۳	۸۱/۴۸
۴	اگر تعداد دفعات بارداری بالا و ضخامت پوست سه سر نرمال باشد آنگاه فرد مبتلا به دیابت است اگر فشار خون بسیار پایین و سابقه خانوادگی دیابت نوع ۲ بسیار بسیار پایین باشد آنگاه فرد مبتلا به دیابت نیست	۷۹/۰۵	۷۰/۸۳	۸۱/۴۸

قانون شماره ۱ و ۲ از مجموعه قوانین انتخاب‌شده با ۵ حالت برای هر ویژگی و قوانین ۳ و ۴ از مجموعه قوانین با ۷ حاالت برای هر ویژگی بهترین عملکرد را از میان ۸۱ قانون داشته‌اند. با توجه به مقادیر ذکرشده در تصویر شماره ۲ برای ترم‌های زبانی هر متغیر و با توجه به قانون شماره ۱، اگر تعداد دفعات بارداری بالا باشد، علی‌رغم نرمال بودن ضخامت پوست چین سه سر، شخص مبتلا به دیابت است. براساس قانون شماره ۲، اگر ضخامت پوست چین سه سر بالا باشد، علی‌رغم نرمال بودن شاخص توده بدنی شخص دارای دیابت است. براساس قانون شماره ۳، اگر تعداد دفعات بارداری بسیار پایین باشد و قند خون نیز بسیار پایین باشد، شخص دارای دیابت نیست. براساس قانون شماره ۴، فردی که دارای عدم سابقه خانوادگی در دیابت است و دارای فشار خون بسیار پایین است، مبتلا به دیابت نیست. جدول شماره ۴ نتیجه حاصل از ترکیب بهترین قوانین منفرد و خروجی مدل پیشنهادی را ارائه می‌دهد.

جدول ۵. مقایسه کارایی و تفسیرپذیری روش پیشنهادی با روش‌های بررسی شده در پیشینه روی مجموعه داده دیابت PIMA

رفرنس	تعداد قوانین	تعداد ترم‌ها	تابع عضویت	طول قانون	تعداد رکوردهای مجموعه داده	صحت (درصد)
[۲]	۷	۳	مثلثی متقارن	۱۵	۵۹۰	۷۷/۱۲
[۲۰]	۳	۷	گزارش نشده	۷۲	۷۳۲	۷۸/۰۳
[۳]	۵	۵	ذوزنقه‌ای نامتقارن	۶	۷۶۸	۸۵/۱۰
[۱۰]	۳	۴	گوسی نامتقارن	۶	۷۶۸	۷۷/۰۰
[۱۵]	۱	–	–	۱۰	۵۹۰	۷۹/۳۲
[۱۶]	۵۳	۸	مثلثی نامتقارن	گزارش نشده	۷۶۸	۷۳/۸۳
[۱۹]	۱۹	۵	مثلثی نامتقارن	۹۱	۷۶۸	۸۴/۰۰
مدل پیشنهادی	۲	۴	مثلثی متقارن	۴	۵۹۰	۷۹/۰۵

مقایسه با سایر روش‌ها

جدول شماره ۵ مقایسه کارایی روش پیشنهادی را برحسب تعداد قوانین انتخاب‌شده در پایگاه قانون نهایی، تعداد ترم‌های زبانی برای توصیف هر متغیر، تابع عضویت فازی، مجموع تعداد شرط‌ها در قوانین، تعداد رکوردهای مورداستفاده برای ساخت مدل و صحت با برخی از روش‌های موجود و بررسی‌شده در پیشینه مطالعه نشان می‌دهد.

هرچند پژوهشی با ارائه تنها یک قانون کوچک‌ترین مجموعه قانون تشخیصی را ارائه می‌دهد [۱۵]، اما مجموعه قانون مدل پیشنهادی تفسیرپذیری بالاتری دارد. در زیر قانون ارائه‌شده، علی‌رغم کارایی مناسب با توجه به روابط ضربی و توانی در ویژگی‌ها، جهت استفاده توسط کاربر انسانی مناسب نیست.

IF ([BMI+0.003×[DP FUNCTION×(SERUM INS × BMI + AGE × AGE + (PG CONCENTRATION× PREGNANCIES) × PREGNANCIES)]+0.65×PG CONCENTRATION]>111) Then Diabetic else not-Diabetic

بحث

در این مقاله به دلیل اهمیت استخراج قوانین دقیق و تفسیرپذیری آن‌ها در سیستم‌های همیار پزشکی، یک مدل استخراج قانون با استفاده از الگوریتم ترکیبی ژنتیک‌فازی و با هدف تعیین فشرده‌ترین مجموعه قوانین با صحت بالا ارائه و روی مجموعه داده دیابت پیما ارزیابی شد. همچنین جهت جلوگیری از پیچیدگی تنظیم پارامترهای الگوریتم ژنتیک و حذف آن‌ها بر کارایی مدل، یک ساختار کروموزومی جدید با قابلیت تنظیم خودکار عملگرهای جهش و بازترکیب ارائه شد. بهترین مجموعه قوانین برحسب ۲ معیار قابلیت تفسیرپذیری و صحت بالا، مجموعه قانونی با ۲ قانون، ۴ مجموعه ترم فازی به ازای هر مقدم در قانون و متوسط طول ۲ به ازای هر قانون بود که به صحت ۷۹/۰۵ درصد دست یافت. تابع عضویت به‌کاررفته برای مجموعه متغیرهای زبانی تابع مثلثی متقارن است که به دلیل ثابت بودن عرض تمامی توابع درک مفاهیم را توسط کاربر انسانی ساده‌تر می‌کند.

نتیجه ارزیابی روش پیشنهادی روی مجموعه داده دیابت پیما نشان می‌دهد که روش پیشنهادی قادر به تولید مجموعه قوانین فشرده همراه با صحت بالاست. در میان ۸۱ قانون تولیدشده ۴ قانون منفرد ذکرشده در جدول شماره ۳ بالاترین کارایی را داشته‌اند. این قوانین موارد زیر را نشان می‌دهند:

• با توجه به قانون‌های ۱ و ۴، تعداد دفعات بارداری عامل مهمی در ابتلا یا عدم ابتلا به دیابت است. به‌نحوی‌که تعداد دفعات بارداری پایین احتمال عدم ابتلا به دیابت و تعداد دفعات بارداری بالا شانس ابتلا به دیابت را به‌شدت افزایش می‌دهد. تأثیر پایین بودن تعداد دفعات بارداری در عدم ابتلا و بالا بودن آن در ابتلا به دیابت بیان شده است [۱۵، ۲۰، ۲۱].

• با توجه به قوانین ۱ و ۲، ضخامت پوست چین سه سر بالا نشانه مهمی از ابتلا به دیابت است. این ویژگی در هر ۲ قانون مربوط به تشخیص ابتلا به دیابت انتخاب شده است. همچنین این قانون در مطالعات مختلف بیان شده است [۲۰، ۲۱].

• با توجه به قانون شماره ۲، شاخص توده بدنی بالا علامتی از ابتلا به دیابت است. همچنین این قانون در مطالعات مختلف بیان شده است [۳، ۹، ۱۰، ۲۰].

• با توجه به قانون شماره ۳، قند خون پایین نشانه سلامت شخص است. تأثیر ویژگی قند در مطالعات مختلفی بیان شده است [۴، ۹، ۱۰، ۱۵، ۲۰].

• با توجه به قانون شماره ۴ فشار خون پایین نشانه عدم ابتلا به دیابت است. ارتباط فشار خون پایین و عدم ابتلا به دیابت در برخی مطالعات بیان شده است [۹، ۲۰، ۲۱].

• با توجه به قانون شماره ۴، عدم وجود دیابت در سابقه خانوادگی شخص عامل مهم دیگری در عدم ابتلا به دیابت است [۹، ۱۰، ۱۵، ۲۰، ۲۱].

نتایج آزمون تی جهت بررسی ارتباط معنادار متغیرها با متغیر وابسته که در جدول شماره ۱ آمده است نشان‌دهنده وجود ارتباط میان تمام متغیرهای مستقل با متغیر وابسته است. بااین‌حال، در قوانین به‌دست‌آمده توسط روش پیشنهادی در هیچ قانونی تمام متغیرها انتخاب نشده‌اند. این امر نشان می‌دهد ارتباط میان متغیرها خطی نبوده و درنتیجه با روش‌های ساده آماری قابل‌کشف نیستند. درنتیجه نیاز به استفاده از روش‌های یادگیری ماشین و داده‌کاوی است.

با توجه به نتایج حاصله که شامل معیارهای طول قوانین، تعداد قوانین و توابع عضویت است این معیارها از مهم‌ترین شاخص‌های تفسیرپذیری قوانین توسط کاربر انسانی هستند [۱۹، ۱۰، ۱۶، ۳] مدل پیشنهادی شامل یک قانون تشخیصی افراد بیمار و یک قانون تشخیصی افراد فاقد بیماری است که تفسیرپذیری بالا ارائه می‌دهد. همچنین با توجه به نتایج به‌دست‌آمده از ارزیابی مدل روی مجموعه داده دیابت پیما مدل پیشنهادی با دست‌یابی به صحت ۷۹/۰۵، ارزش اخباری مثبت ۷۰/۸۳ و ارزش اخباری منفی ۸۱/۴۱ درصد ضمن کارایی مناسب و قابلیت اعتماد بالا به نتایج قوانین تشخیصی، می‌تواند به‌عنوان یک مدل عمومی امیدبخش در سایر کاربردهای دسته‌بندی داده‌های پزشکی مورد استفاده قرار بگیرد.

نتیجه‌گیری

استخراج قوانین از مجموعه داده‌ها در کاربردهای تشخیص پزشکی، یک حوزه مهم در کشف دانش است؛ سیستم‌های فازی به جهت تولید قوانین قابل‌تفسیر توسط انسان، به‌عنوان یک ابزار محبوب در این کاربرد شناخته می‌شوند. جهت طراحی خودکار سیستم‌های فازی از روی داده‌ها، الگوریتم ژنتیک توانایی بالایی از خود نشان داده است. به همین دلیل در این مقاله یک سیستم ژنتیک‌فازی استخراج قانون خودسازمانده برای کاربردهای تشخیصی پزشکی پیشنهاد شد. جهت ارزیابی مدل پیشنهادی از مجموعه داده دیابت پیما استفاده شد. ضمن ارائه مجموعه قوانین تشخیصی، بهترین قوانین منفرد تشخیصی افراد فاقد و مبتلا به دیابت ارائه شد. تعداد کم قوانین وکوتاه بودن آن‌ها ازجمله ویژگی‌های مهم قوانین پیشنهادی است که درکنار پیاده‌سازی سریع، می‌تواند توسط متخصصان به‌آسانی ارزیابی و به خاطر سپرده شود.

از محدودیت‌های پژوهش استفاده از داده‌های عمومی و عدم ارزیابی قوانین پیشنهادی توسط متخصصین بالینی است. هرچند با مقالات مشابه، قوانین ارزیابی شده‌اند. پیشنهاد می‌شود مدل پیشنهادی روی داده‌های محلی اجرا شده و هریک از قوانین استخراجی را متخصصین ارزیابی کنند.

ملاحظات اخلاقی

پیروی از اصول اخلاق پژوهش

با توجه به اینکه در این مطالعه از مجموعه داده دیابت پیما که یک مجموعه داده عمومی است استفاده شده است نیازی به اخذ کد اخلاق وجود نداشت.

حامی مالی

این مقاله مستخرج از طرح پژوهشی (کد: ۰۰-۱-۰۰-۴۰) در دانشگاه صنعتی خاتم الانبیاء بهبهان بوده است.

مشارکت نویسندگان

ایده مقاله: امین گلاب‌پور؛ نگارش مقاله و بحث: فاطمه آهوز؛ پیاده‌سازی مدل، روش، و ارزیابی و درصد همکاری: تمام نویسندگان.

تعارض منافع

بنابر اظهار نویسندگان این مقاله تعارض منافع ندارد.

References

[1] Ishibuchi H, Nojima Y, Kuwajima I. Genetic rule selection as a post-processing procedure in fuzzy data mining. Int Symp Evolv Fuzzy Syst; 2006. [DOI:10.1109/ISEFS.2006.251149]

[2] Ahouz F, Golabpour A. A novel compact rule extractor based on genetic-fuzzy algorithm. Paper presented at: 10th International Conference on Computer and Knowledge Engineering (ICCKE); 29-30 October 2020, Mashhad, Iran. [DOI:10.1109/ICCKE50421.2020.9303613]

[3] Gorzałczany MB, Rudziński F. Interpretable and accurate medical data classification–a multi-objective genetic-fuzzy optimization approach. Expert Syst Appl. 2017; 71:26-39. [DOI:10.1016/j.eswa.2016.11.017]

[4] Mitra S, Hayashi Y. Neuro-fuzzy rule generation: Survey in soft computing framework. IEEE Trans Neural Netw. 2000; 11(3):748-68. [PMID]

[5] Shi Y, Eberhart R, Chen Y. Implementation of evolutionary fuzzy systems. IEEE Trans Fuzzy Syst. 1999;7(2):109-19. [DOI:10.1109/91.755393]

[6] Ahouz F, Golabpour A. A novel structure of highly interpretable fuzzy rules extraction. Front Health Inform. 2021; 10(1):53. [DOI:10.30699/fhi.v10i1.253]

[7] Shortliffe EH, Wiederhold G, Fagan LM, Perreault LE. Medical Informatics: Computer Applications in Health Care and Biomedicine. New York: Springer; 2013. [Link]

[8] Sujatha R, Ephzibah EP, Dharinya S, Uma Maheswari G, Mareeswari V, Pamidimarri V. Comparative study on dimensionality reduction for disease diagnosis using fuzzy classifier. Int J Eng Technol. 2018; 7(1):79-84. [DOI:10.14419/ijet.v7i1.8652]

[9] Seera M, Lim CP. A hybrid intelligent system for medical data classification. Expert Syst Appl. 2014; 41(5):2239-49. [DOI:10.1016/j.eswa.2013.09.022]

[10] Chang X, Lilly JH. Evolutionary design of a fuzzy classifier from data. IEEE Trans Syst Man Cybern B Cybern. 2004; 34(4):1894-906. [PMID]

[11] GaneshKumar P, Rani C, Devaraj D, Victoire TAA. Hybrid ant bee algorithm for fuzzy expert system based sample classification. IEEE ACM Trans Comput Biol Bioinformatics. 2014; 11(2):347-60. [PMID]

[12] Pinto CMA, Carvalho ARM. Diabetes mellitus and TB co-existence: Clinical implications from a fractional order modelling. Appl Math Model. 2019; 68:219-43. [DOI:10.1016/j.apm.2018.11.029]

[13] Abedini S, Jomehpour S, Fallahi S, Ghanbarnejad A, Nikparvar M. [The effect of virtual education of cardiovascular risk factors on the knowledge of general physicians (Persian)]. J Mod Med Inf Sci. 2020; 6(1):31-6. [DOI:10.29252/jmis.6.1.32]

[14] Schulz LO, Chaudhari LS. High-risk populations: The pimas of Arizona and Mexico. Curr Obes Rep. 2015; 4(1):92-8. [PMID]

[15] Ahouz F, Sadehvand M, Golabpour A. Extracting rules for diagnosis of diabetes using genetic programming. Int J Health Stud. 2019; 5(3):23-32. [DOI:10.22100/ijhs.v5i3.691]

[16] Feng TC, Li THS, Kuo PH. Variable coded hierarchical fuzzy classification model using DNA coding and evolutionary programming. Appl Math Model. 2015; 39(23-24):7401-19. [DOI:10.1016/j.apm.2015.03.004]

[17] Singh S, Singh S, Banga VK. Design of fuzzy logic system framework using evolutionary techniques. Soft Comput. 2020; 24(6):4455-68. [DOI:10.1007/s00500-019-04207-9]

[18] López-Campos JA, Segade A, Fernández JR, Casarejos E, Vilán J. Behavior characterization of visco-hyperelastic models for rubber-like materials using genetic algorithms. Appl Math Model. 2019; 66:241-55. [DOI:10.1016/j.apm.2018.08.031]

[19] Mansourypoor F, Asadi S. Development of a reinforcement learning-based evolutionary fuzzy rule-based system for diabetes diagnosis. Comput Biol Med. 2017; 91:337-52. [PMID]

[20] Tan CH, Tan MS, Chang SW, Yap KS, Yap HJ, Wong SY. Genetic algorithm fuzzy logic for medical knowledge-based pattern classification. J Eng Sci Technol. 2018; 13:242-58. [Link]

[21] Vaishali R, Sasikala R, Ramasubbareddy S, Remya S, Nalluri S. Genetic algorithm based feature selection and MOE Fuzzy classification algorithm on Pima Indians Diabetes dataset. International Conference on Computing Networking and Informatics (ICCNI); 29-30 October 2017; Lagos, Nigeria. [DOI:10.1109/ICCNI.2017.8123815]