JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Vol. 9, Issue.9, September 2021 E-ISSN: 2347-2693

Early-Stage Diabetes Risk Detection Using Data Mining Techniques With Particle Swarm Optimization

Sanat Kumar Sahu

Department of Computer Science, Govt. Kaktiya P.G. College Jagdalpur (C.G.), India

Author's Mail Id:sanat.kosa1@gmail.com

DOI: https://doi.org/10.26438/ijcse/v9i9.6365 | Available online at: www.ijcseonline.org

Received: 18/Sept/2021, Accepted: 20/Sept/2021, Published: 30/Sept/2021

***Abstract*—** In this study feature Selection technique (FST) namely Particle Swarm Optimization (PSO) is used to optimize the features of diabetes datasets. There are different types of classifiers that give low performances. So we need an FST to combined classifier may be required for best results. We used FST to improve the overall performance of the classification model. Classification of diabetes dataset classifier C4.5 and Support Vector Machine (SVM) is applied. The selected feature of diabetes is applied to classifiers and a comparative study was conducted. The experimental outcome reveals that the C4.5 is performed better with selected features compared to other models.

***Keywords*—** Classification, C4.5, feature Selection technique, Particle Swarm Optimization, Support Vector Machine

I. Introduction

Intelligence techniques can capture human beings and collective knowledge extending a knowledge base with the help of machine learning techniques[1]. It does knowledge detection or discovering principles from unknown patterns in data sets by using data mining (DM). In the healthcare field, newly found knowledge may be used by health workers and medical professionals to enhance diagnostic accuracy, improve medical processes quality, and reduce prescription side effects. It also aspires to suggest less expensive treatment options [2]. The classification of healthcare data in the present scenario is very important and has gained the attention of medical researchers in previous years [3]. Also, optimizing features solves the scalability problem and improves the performance of classification models by eliminating redundant, irrelevant, or noisy features from large datasets. The great interest of diabetic disease is considered, which is a serious health problem in the world is and a comparative analysis of FST like PSO were carried out based on the performance of C4.5 and SVM classification for the classification of diabetes disease risks.

II. Related Work

Many researchers have been working in the field of computer-based disease diagnosis systems. These diseases include lungs cancer, breast cancer, heart diseases, thyroid diseases, and other diseases occurring in human beings. In this paper we have reviewed the diabetic datasets paper already done.

Altamimi (2020) **[4]** worked on the classification of diabetes in children detection datasets using the algorithms like Naïve Bayes (NB), Random Forest(RF), Decision Tree (DT) and Support vector machine(SVM). The maximum accuracy received by SVM as compared to other methods used. Pethunachiyar (2020) **[5]** to classify the diabetic dataset they used Logistic Regression(LR), Neural Network(NN), NB and SVM. The SVM with polynomial kernel obtained the highest accuracy compared to other methods. Kaur et al.(2019) **[6]** used the SVM method to classify the diabetic dataset. The SVM obtained the 75.3% maximum accuracy. Kumari et al. (2013) **[7]** worked to classify the Pima Indian Diabetes Dataset. They utilized the classification algorithm SVM. The SVM obtained a maximum of 78% accuracy. Soliman et al. (2014) **[8]** used the classification as SVM and feature optimization as PSO. They obtained a maximum of 97.83% accuracy

III. Methodology

The research work started with the study of background information of classification for diabetic disease by using Data Mining (DM) method. In the present research work the C4.5, SVM have used proposed and PSO has been used as FST.

A. Support Vector Machine

Support Vector Machine (SVM) is a well-known data mining classification method. In machine learning, SVM is a supervised learning method with correlated learning algorithms that study the data used for regression and classification. An SVM [9] is a new method for the classification of each linear and nonlinear data. This decision is based on the idea of planes which define decision boundaries. A decision plane is different between the set of various class membership objectives.

B.C4.5 Trees

C4.5 [10], [11]could be a methodology utilizes to provide a decision tree developed by Ross Quinlan.C4.5 is a related expansion of Quinlan's former ID3 algorithm.

C. Feature Selection Technique

Feature selection is a crucial technique to explain the problem of dimensionality in machine learning by choosing appropriate and non-redundant features [12], [13]. Feature Selection Technique (FST) is also called the feature optimization technique which helps to remove the unrelated feature subset from the original feature space[14].

D. Particle Swarm optimization

In DM, a meta-heuristic is a high-level approach designed to fix, generate or select a heuristic, specifically providing an appropriate solution to an optimization problem; but it has limited computing capability [15]. PSO is a population-based stochastic optimization technique developed by Dr. Eberhart and Dr. Kennedy in 1995 and inspired by bird flocking or fish schooling social behavior [16], [17].

IV. RESULTS AND DISCUSSION

The Dataset of the diabetic is downloaded by the UCI Machine Learning website. This dataset has a total of 17 variables with appearing as a binary class. In the preliminary experiments, we have used the 10fold cross-validation method.

Table 1: Shows the feature selected by PSO

Attribute Selected by PSO	Total Attributes
$a_1,a_2,a_4,a_5,a_8,a_9,a_{10},a_{11},a_{12},a_{13},a_{14},a_{16}$	12

The list of attributes selected by PSO is shown in the above table. The PSO has selected a total of 12 attributes out of a total of 17 attributes.

Table 2: Accuracy of Classification

Name of Classifiers	FST	Accuracy
C4.5	WFST	95.96
PSO-C4.5	PSO-J48	96.15
SVM-RBF Kernel	WFST	87.5
PSO-SVM-RBF Kernel	PSO-J48	89.62

Table 3: Sensitivity of Classification

Name of Classifiers	FST	Sensitivity
C4.5	WFST	95
PSO-C4.5	PSO-J48	97.19
SVM-RBF Kernel	WFST	85.63
PSO-SVM-RBF Kernel	PSO-J48	86.88

Table 4: Specificity of Classification

Name of Classifiers	FST	Specificity
C4.5	WFST	97.5
PSO-C4.5	PSO-J48	94.5
SVM-RBF Kernel	WFST	90.5
PSO-SVM-RBF Kernel	PSO-J48	94

Table 5: F1-Score of Classification

Name of Classifiers	FST	F1-SCORE
C4.5	WFST	96.66
PSO-C4.5	PSO-J48	96.88
SVM-RBF Kernel	WFST	89.4
PSO-SVM-RBF Kernel	PSO-J48	91.15

The above tables 2, 3, 4 and 5 Shows the different performances values of proposed models.

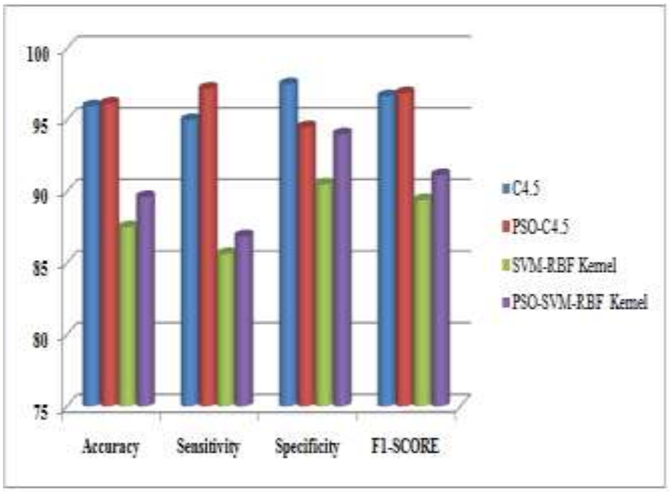

Figure 1: Comparative Bar chart of Proposed Models

Figure 1 Comparative study shows that after applying the FST the performance of the classifier's accuracy is increased compared to the WFST. The highest accuracy is obtained by the proposed PSO-C4.5 model compared to all models.

Table 6: Accuracy difference after and before FST

Name of Classifiers	WFST	FST-PSO	Difference
C4.5	95.96	96.15	0.19
SVM-RBF Kernel	87.5	89.62	2.12

The accuracy difference in FST and WFST proposed model is shown the above table 3. It shows the classifiers with the selected subset of features gives higher accuracy compare to all features of the diabetic dataset.

V. CONCLUSION AND FUTURE SCOPE

This research was created to aid in the identification of diabetes. There are some duplicate and low useful features in the databases that were removed. These features are decreasing the classifier's performance and the system's processing time. So we used the FST like PSO. The results demonstrated that employing FST using PSO has enhanced classification accuracy. The obtained findings demonstrate that the suggested method's functionality is extremely successful when compared to other results obtained and is very promising for pattern recognition applications.

Suitable other possible FST, optimization techniques, and different kinds of tools that can be applied to obtain better results in the proposed model.

REFERENCES

[1] S. K. Sahu and P. K. Chandrakar, "Classification of Chronic Kidney Disease with Genetic Search Intersection Based Feature," in Advances in Intelligent Systems and Computing 1122, **vol. 1, pp. 11–21, 2020.**

[2] S. M. Alzahani, A. Althopity, A. Alghamdi, B. Alshehri, and S. Aljuaid, "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction," Eng. Technol. Publ., **vol. 2, no. 4, pp. 310–315, 2014.**

[3] B. O. Eriksen and O. C. Ingebretsen, "In chronic kidney disease staging the use of the chronicity criterion affects prognosis and the rate of progression," Kidney Int., **vol. 72, no. 10, pp. 1242–1248, 2007.**

[4] A. M. Altamimi, "Performance Analysis of Supervised Classifying Algorithms to Predict Diabetes in Children," J. Xi'an Univ. Archit. Technol., vol. XII, no. III, **pp. 2010–2017, 2020.**

[5] G. A. Pethunachiyar, "Classification of diabetes patients using kernel based support vector machines," in 2020 International Conference on Computer Communication and Informatics, ICCCI 2020, **pp. 22–25, 2020.**

[6] H. Kaur and G. Kaur, "Prediction of Diabetes Using Support Vector Machine," Int. J. Res. Eng. Appl. Manag., **vol. 05, no. 02, pp. 470–473, 2019.**

[7] A. Kumari and R. Chitra, "Classification Of Diabetes Disease Using Support Vector Machine," Int. J. Eng. Res. Appl., **vol. 3, no. 2, pp. 1797–1801, 2013.**

[8] O. S.Soliman and E. AboElhamd, "Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine," Int. J. Comput. Trends Technol., **vol. 8, no. 1, pp. 38–44, 2014.**

[9] S. Sathyanarayana and S. Amarappa, "Data classification using Support vector Machine (SVM), a simplified approaCH," Int. J. Electron. Comput. S cience Eng. Vol. 3, Number 4, ISSN-2277-1956, **pp. 435–445, 2014.**

[10] J. R. Quinlan, "Improved Use of Continuous Attributes in C4 . 5," **vol. 4, no. 1996, pp. 77–90, 2006.**

[11] A. K. Shrivas and S. K. Sahu, "Classification of Chronic Kidney Disease using Combination Feature Selection Techniques and Classifiers," **vol. 7, no. 3, pp. 114–117, 2019.**

[12] N. Boodhun and M. Jayabalan, "Risk prediction in life insurance industry using supervised learning algorithms," Complex Intell. Syst., no. **March, 2018.**

[13] M. Dash and H. Liu, "Feature selection for classification," Intell. Data Anal., **vol. 1, no. 3, pp. 131–156, 1997.**

[14] P. Verma, V. K. Awasthi, and S. K. Sahu, "An Ensemble Model With Genetic Algorithm for Classification of Coronary Artery Disease," Int. J. Comput. Vis. Image Process., **vol. 11, no. 3, pp. 70–83, 2021.**

[15] L. Bianchi, M. Dorigo, L. Maria, and W. J. Gutjahr, "A survey on metaheuristics for stochastic combinatorial optimization," **pp. 239–287, 2009.**

[16] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," **pp. 1942–1948, 1995.**

[17] Y. Shi and R. Eberhart, "A Modified Particle Swarm Optimizer," **pp. 69–73, 1998.**

AUTHORS PROFILE

Dr. Sanat Kumar Sahu is working as an Assistant Professor in the Department of Computer Science, Govt. Kaktiya PG College, Jagdalpur (Bastar) Chhattisgarh, India. He has more than 11 years teaching Experience. His area of interest includes soft computing, machine learning, and data mining.. He has more than 22 research paper in national and international journals.