TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 8195 - 8199

Classification of Diabetes Dataset using KNN Classifier and Attribute Selection through Bees Algorithm

V. Karunakaran, Computer Science and Engineering, Karunya Institute of Technology and Sciences, Coimbatore.
C. Sorna Chandra Devadass, Civil Engineering, Samskruti College of Engineering and Technology, Hyderabad.
V. Rajasekar, Computer Science and Engineering, SRM Institute of Science and Technology, Chennai.

Article Info
Volume 83
Page Number: 8195 - 8199
Publication Issue:
May - June 2020

Article History
Article Received*: 19 November 2019*
Revised*: 27 January 2020*
Accepted*: 24 February 2020*
Publication*: 18 May 2020*

Abstract:

Diabetes is considered as a one of the chronic disease it cause an increase of sugar in the human blood. The prime objective of this research work is provides better classification accuracy through reduced set of attributes. In this work, first the classification is carried out with entire dataset, and then second the classification is carried out through reduced dataset i.e., with reduced number of attributes. The proposed system has evaluated through three performance metrics such as Detection Rate (DR), Accuracy Rate (AR) and False Positive Rate (FPR). The proposed system provides good AR, DR and FPR what we have obtained through entire dataset.

Keywords: *KNN Classifier, Diabetes dataset, Bees algorithm, Feature Selection, Classification.*

1. Introduction

Classification strategies were broadly used in medical fields for classifying the data into different number of classes based on several constraints. Normally the human body has ability to produces hormone insulin. If the person is suffered from diabetes illness, the human body losses the ability to producing hormone insulin due to this reason, some changes should be happened in the human body, first one is metabolism of carbohydrates became abnormal and second one raises the level of glucose in the blood. The primary objective of this research work will provides better classification accuracy. Lot of techniques was available in the practices. Deepti Sisodia et al. designed a model for predicting diabetes with maximum accuracies. The experiment was conducted with three well known classifiers such as, 1. NB (Navie Bayes) 2. SVM (Support Vector Machine) and 3. DT (Decision Tree), among the three classifiers authors claimed that Naïve Bayes algorithm provides better classification accuracies [1]. Gopinath et al. made comparative study on different classification algorithm for diabetes dataset. Finally authors claimed KNN and naïve bayes classifier provides better classification accuracy [2]. Rajesh kumar et al. proposed whale optimization algorithm used for selecting finest features and they used back propagation neural network classifier for classification. The experiment result shows the proposed system provides better accuracy, less execution time compared to the existing system such as Particle Swarm Optimization (PSO) and Whale Optimization Algorithm [3]. Aishwarya Iyer et al. investigated the performance of Decision tree and Navie bayes classifiers for diabetes dataset [4]. Aruna kumari et al. proposed ensemble based neural network for diabetes classification. In this paper, the proposed work is divided into three phases. In the first phase, the attribute selection is carried out through neighbouring search techniques. In the second phase, apply attribute ranking model to selected attributes to obtain best subset of attributes. In the third phase, finally the best subset of attributes were trained and classified by neural network classifiers.

May – June 2020
ISSN: 0193-4120 Page No. 8195 - 8199

The proposed system is compared with the following methods such as adaptive fuzzy, modified particle swarm optimization and etc. The proposed system provides better classification accuracies [5]. Dilip Kumar Choubey et al. proposed diabetes classification using Naïve Bayes classifier and attribute selection is carried out through Genetic Algorithm. The proposed system is compared with several existing methods and the proposed system provides better classification accuracy with minimum computation cost and maximizes the ROC. Fatima Bekaddour et al. evaluated the performance of several Meta heuristic algorithms for diabetes diagnosis. In this paper, the following Meta heuristic algorithms were evaluated such as Particle Swarm Optimization (PSO), Firefly Optimization Algorithm (FOA) and Homogeneity Based Algorithm. In terms of accuracy Homogeneity Based Algorithm (HBA) provides better result and in terms of computational cost Firefly Optimization Algorithm is best [6]. Khyati K. Gandhi et al. evaluated the performance of Support Vector Machine classifier through diabetes dataset. The proposed system is separated into two phases. In the first phase feature selection is carried out through F-score method and K means clustering algorithm. In the second phase classification is carried out through SVM classifier for reduced features. The performance of Support Vector Machine is evaluated based on Sensitivity rate, Accuracy rate and Specificity rate and authors claimed that, the work carried out through feature selection method and data normalization will improve the performance of SVM classifier [7]. Rahmat Zolfaghari proposed ensemble of Back Propagation Neural Network and Support Vector Machine for diabetes diagnosis. Several methods were compared with the proposed system and the proposed system provides better classification accuracies [8]. Nithyapriya et al. analyzed various classification techniques to predict diabetes. In this paper the following classifiers were analyzed such as 1.Support Vector Machine (SVM), 2. J48 and 3. Naïve Bayes (NB) and the authors claimed the Support Vector Machine (SVM) provides better classification accuracies [9].

The rest of the paper is planned as follows: Section 2 offers Introduction to Attribute Selection Method and Bees Algorithm. Section 3 offers an Overall Architecture of the Proposed System. Section 4 offers a finding optimal subset attributes using Bees Algorithm. Section 5 offers experimental results and finally section 6 offers conclusion.

2. Introduction to Attribute Selection and Bees Algorithm

Attribute selection is a task of identifying the attributes which are contributing more information during classification. Mainly attribute selection methods were classified into 3 categories such as Filter method (FM), Wrapper method (WM) and Embedded method (EM). Filter Method a selection criterion is based on evaluation function and not based on classification algorithms. The selection criterion of wrapper method is purely depends upon on the classification accuracy what have been obtained through classification algorithms. Embedded method, the selection criterion of attribute is based on combination of filter method and wrapper method. The proposed work is belongs to wrapper method, so the selection criterion is based on classification algorithms. Bees algorithm was developed by Pham, Ghanbarzadeh and et al. in the year of 2005 and the algorithm belongs to population based strategies. The algorithm will mimics the food foraging behaviour of honey bee colonies. The Bees algorithm is belongs to global optimization algorithm and the algorithm will help to run away from local optima and surely it will reach a global optima so that the bee's algorithm surely arrives a global solution. The basic version of the algorithm will performs a neighborhood search operation combined with global search. The template of basic version of bee's algorithm is shown in figure 1.

May – June 2020
ISSN: 0193-4120 Page No. 8195 - 8199

1. Initialise population with random solutions.
2. Evaluate fitness of the population.
3. While (stopping criterion not met) //Forming new population.
4. Select elite bees.
5. Select sites for neighbourhood search.
6. Recruit bees for selected sites (more bees for best e sites) and evaluate fitness.
7. Select the fittest bee from each patch.
8. Assign remaining bees to search randomly and evaluate their fitness.
9. End While.

Figure 1: Basic version of Bees algorithm.

3. Overall Architecture of Proposed System

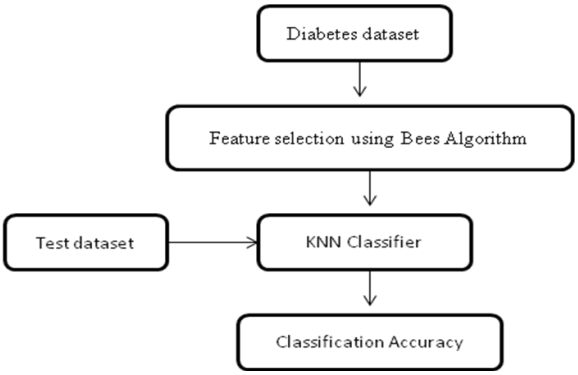

Figure 2: Overall architecture of proposed system

This work consists of two phases and input is a diabetes dataset. In the first phase, we are applying bee's algorithm for identifying finest subset of attributes from the entire attributes. In the second phase, the reduced representation of dataset is obtained from the first phase is classified by using KNN classifier. The proposed system is appraised in terms of false positive rate, accuracy rate and detection rate. The proposed system is compared with entire or original dataset i.e., without feature selection and the proposed system provides good AR, DR and FPR and the overall architecture of the proposed system is shown in figure 2.

4. Feature selection using Bees Algorithm

The work consists of two phase, Now we see a first phase, attribute selection is carried out through bees algorithm and the bees algorithm is fit into population-based search algorithm and it also belongs to global optimization. In this work, input is a diabetes dataset and it consists of eight attributes, all the eight attributes are not contributing equal information. Some of the attributes were contributing additional information and some of the attributes were contributing very small information. In this work, we are using bees algorithm and decision tree classifier for identifying which are the attributes are contributing more information. Diabetes dataset consists of eight attributes. In our experiment selected subset size of attributes is considered as a five and unselected subset size of attributes is considered as three. Now we see how to find the finest subset of attributes by using Bees algorithm.

Step 1: Initialize the population with random solutions.

Create N number of random solutions and the Decision Tree classifier is used for calculating the fitness value for each random solution. Among the solutions, which solution is having finest fitness value then we will assume that solution is a finest subset of attributes. Two subset is considered in this work 1. Selected subset of attributes and 2. Unselected subset of attributes. In this work, consider the selected subset size is five and unselected subset size is three. Intersection of unselected subset attributes and selected subset of attributes is null.

Step 2: While stopping criteria is not met, then goto step 3, otherwise goto step 6.

Step 3: Evaluate an each initial solution with fitness value

Arrange N random solution in descending order based on fitness value computed in step 1. Example: S1, S2……SN.

Step 4: Neighborhood operations

Take first N/2 solutions from the step 3. Apply exchange operator between them to create new subset.

Example:

TEST Engineering & Management

May – June 2020
ISSN: 0193-4120 Page No. 8195 - 8199

Now we see how to create a new subset. Randomly choose one attribute from the selected subset of attributes and unselected subset of attributes and exchange them.

S1: Selected subset of attributes. (A stands for attribute)

A 1	**3**	6	7	8

Randomly select one attribute from the selected subset of attributes. Example selected attribute is A3.

S1: Unselected subset of attributes.

2	A4	**5**

Randomly select one attributes from the unselected subset of attributes. Example selected attribute is A5.

New subset is

New subset of S1: Selected subset of attributes.

A1	A5	A6	A7	A8

New Subset of S1: Unselected subset of attributes.

A2	A3	A4

This process will continue until N/2 times. If any new solution provides finest fitness value than best subset of attributes. Assign the grace solution as a best subset of attributes.

Step 5: Take remaining population and create a random solution among the population. Compute the fitness value for each solution. If any new solution provides finest fitness value than best subset of attributes. Assign the grace solution as a best subset of attributes.

Step 6: Return best subset of attributes.

Stopping criteria is based on number of iteration. Number of iteration is a user defined value.

5. Experimental Results

In this work, Input is a diabetes dataset. It consists of 8 features and 768 instances. The dataset consists of the following attributes 1. Pregnancies, 2. Glucose, 3. Blood Pressure, 4. Skin Thickness, 5. Insulin, 6. BMI, 7. Diabetes Pedigree Function and 8. Age. All the attributes are not contributing equal information. Some of the attributes are contributing additional information and some of the attributes are contributing less information. In our work, we used bee's algorithm to identifying those attributes which are contributing more information. Those attributes are retained in the training dataset remaining attributes are eliminated from the training dataset. How to identify the best subset of features using bee's algorithm is already explained in section 4. In this section, we see how our proposed system provides better FPR, AR and DR what have been obtained through entire dataset.

Fitness function is used in our work is,

$$Fitness = \alpha * DR + \beta * (1 - FPR) \quad \ldots\ldots (1)$$

Our research work is belongs to maximization problem, so we will give more importance to detection rate so we will have alpha value as a 0.7 and we will have false positive rate as a 0.3.

DR= (No. of diabetes instance classified as diabetes/Total no. of diabetes instances in the dataset)* 100. …… (2)

FPR= (No. of normal instances that are incorrectly classified as diabetes/Total no. of normal instances in the dataset)*100. …… (3)

AR= (No. of instances correctly classified/ Total no. of instances in test dataset) *100. (4)

Criteria 1: Original dataset is given to the K-Nearest Neighbor classifiers for classification. FPR, AR and DR are shown in table 1 and graph representation is shown in figure 3.

Criteria 2: Apply feature selection using bees algorithm to the original dataset then reduced representation of dataset are given to the K-Nearest Neighbor classifier for classification. FPR, AR and DR are shown in table 1and graph representation is shown in figure 3.

May – June 2020
ISSN: 0193-4120 Page No. 8195 - 8199

Table 1: FPR, DR and DR of diabetes dataset using K-Nearest Neighbour classifier.

Criteria's	FPR	DR	AR
Criteria 1	5.93±0.41	86.28±0.19	88.98±0.45
Criteria 2	5.02±0.74	87.89±0.31	90.18±0.62

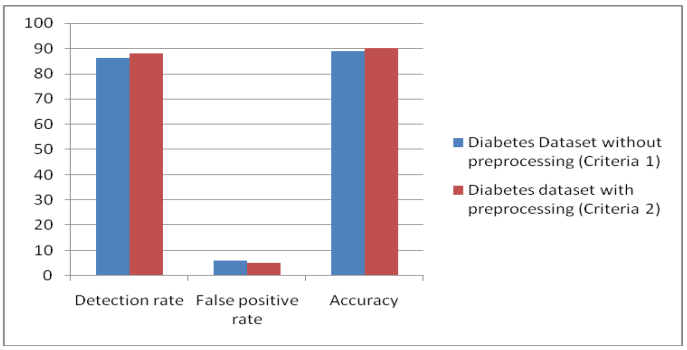

Figure 3: FPR, DR and AR using K-Nearest Neighbor classifier for diabetes dataset.

6. Conclusion

In this paper, the work is split into two phases. First phase attribute selection is carried out through bees algorithm and obtain a reduced representation of data set. In the second phase, reduced representation of dataset is given to the KNN classifier for classification. The performance of proposed system is evaluated in terms of DR, FPR and AR is result is shown in table 1. The experimental result shows clearly the proposed system provides better result in all aspects when compared to what we have obtained through entire dataset.

7. References

[1] Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, 132, 1578-1585.

[2] Gopinath, M. P., & Murali, S. (2017). Comparative study on Classification Algorithm for Diabetes Data set. International Journal of Pure and Applied Mathematics, 117(7), 47-52.

[3] Rajeshkumar, J., & Kousalya, K. (2017). Diabetes data classification using whale optimization algorithm and backpropagation neural network. International Research Journal of Pharmacy, 8(11), 219-222.

[4] Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv:1502.03774.

[5] G L Aruna Kumari, Padmaja P, Jaya Suma G (2020). ENN-Ensemble based Neural Network method for Diabetes Classification. International Journal of Engineering and Advanced Technology, 9(3), 576-579.

[6] Bekaddour, F., Rahmoune, M. B., Salim, C., & Hafaifa, A. Performances study of different metaheuristics algorithm for diabetes diagnosis, Conference Paper in Lecture Notes in Computer Science · May 2017.

[7] Gandhi, K. K., & Prajapati, N. B. (2014). Diabetes prediction using feature selection and classification. International journal of advance Engineering and Research Development, 1(05).

[8] Zolfaghari, R. (2012). Diagnosis of diabetes in female population of pima indian heritage with ensemble of bp neural network and svm. Int. J. Comput. Eng. Manag, 15, 2230-7893.

[9] Devi, M. R., & Shyla, J. M. (2016). Analysis of various data mining techniques to predict diabetes mellitus. International journal of applied engineering research, 11(1), 727-730.