Disease Prediction Model for Diabetes Mellitus Using Ant Miner – Genetic Algorithm

Dr.M.Durgadevi [1]

[1] Assistant Professor(Sr.G),SRM Institute of Science and Technology, Vadapalani Campus, Chennai

ABSTRACT

The two notable evolutionary computational and swarm based approach used for combinatorial optimization problems which have a discrete pursuit space are Ant Miner algorithm (AM) and genetic algorithm (GA). The feature selection process selects a feature subset and then processes the data with chosen features to learn the algorithm. A classification model gets executed in the prediction stage using the last chosen feature. The hybrid algorithm (AM-GA) combines the distinctive features of ants and genetic approach to optimize the attribute reduction. By hybridization, the space intricacy is decreased by eliminating the stagnant behavior of ants and the time complexity is reduced by the global search mechanism adopted by the genetic algorithm. The performance of the proposed approach has been evaluated on the PIMA Indian Type II diabetic datasets taken from the UCI machine learning repository. The experimental results prove that the proposed approach has selected the best possible features for achieving the highest rate of classification accuracy.

Keywords

Ant Miner, Genetic Algorithm, Type II Diabetes, Feature Selection

Article Received: 10 August 2020, Revised: 25 October 2020, Accepted: 18 November 2020

Introduction

The data mining enhances the procedure of decision-making by exploring patterns and tendency in a massive quantity of composite data.. For e.g., A hospital database may contain knowledge about the patients from which one can easily identify which kind of patient has the highest probability of developing a particular disease. The healthcare organizations which make use of data mining techniques are highly equipped to meet its long-term requirements. The process of knowledge discovery in databases is considered to be both iterative and interactive. The preprocessing step involves the preparation of data and the post processing stage validates and refines the extracted knowledge such that it should be accurate and intelligible. In disease prediction model it is highly obligatory that the knowledge discovered should have a high predictive accuracy [1]. The comprehensibility of the knowledge discovered is facilitated by the representation of the 'IF – THEN' rules. The interestingness is determined by how well the model is potentially useful for the user.

In literature there are several ways of representing the knowledge among which logical conjunction plays a significant role. The association rule learning exploits the relation among the parameter value pairs of a dataset in the form of 'if-then' rules. The 'if' portion is called as the antecedent of rule and the 'then' portion known as the consequent of rule, that demonstrates the degree of rule vagueness (does not comprise any values in general).High level of abstraction can be reached with the help of if-then rules which will be comprehensible for most of the users.

IF $cond_1$ AND $cond_2$ ….AND $cond_n$

THEN pred

The rule antecedent has the following form,

(1) A_1 O val_1 (2) A_1 O A_5

An attribute can be compared to its particular value in the domain or it can also be compared with the other attributes. 'O' represents the operators used for comparison {=,>,<}. Rule-based classification is a conventional data mining procedure that is used in extensive medical diagnosis models [2]. The rules stored in the rule base strongly influences the efficiency of classification. The rule sets filtered from the data mining methods will go through optimization by the use of heuristic or meta-heuristic approaches for enhancing the quality of the rule base. In this paper, the evolutionary algorithms have been used for determining the rules with respect to the medical database. The main advantage of using evolutionary algorithm is that they perform a global search mechanism (stochastic search) as they are capable of working with a large population of candidate rules.

The problem of feature selection is to select a feature subset of size's'(s<S) from a large set 'S' with a high rate of accuracy. To improve the

performance of the system the irrelevant attributes must be removed from the real world datasets. There are many approaches [3] proposed for feature selection among which the population based algorithms like the ant colony optimization and genetic algorithms are considered to be optimal as these methods work by using knowledge from previous iterations.

SURVEY OF RESEARCH WORK

Though there are a lot of software's available for storing the medical data, the size of the rule-set is too large to maintain resulting in problems in the server [4]. An artificial intelligent technique plays a vital role in analyzing this non-systematic medical data. The data mining technology integrates techniques from various fields for analysis of large volumes of data.

The performance of the data mining technique can be improved by cooperative effort of humans and computers [5]. Prediction (finding unknown patterns) and description (describing the existing patterns) are the two major goals of data mining. In [6], the different life threatening diseases in the field of medical prediction with respect to various statistical and mathematical approaches has been summarized in brief. The Fuzzy C means approach has outperformed the support vector machine in the prediction of type II diabetes [7].In [8],the fuzzy hierarchical model has been used as an application of computational intelligence for detecting diabetes at an early stage. This fuzzy approach was analyzed on real time datasets taken from the Jakarta hospital in Indonesia. Fasting Plasma glucose concentration is the main attribute in diabetic analysis.

A study has been performed in Korean adults (males and females) for analyzing and comparing the various anthropometric measures using logistic regression and naïve bayes classifier to detect the high risk of type II diabetes [9].Rule based approach using support vector machines using an ensemble of classifiers has been used as an application for diagnosing diabetes which has resulted in high precision and recall [10].The performance of various classifier models like J48 Decision Tree, K-Nearest Neighbors, and Random Forest, Support Vector Machines were compared to classify patients with diabetes mellitus[11].The decision tree and J48 classifier has achieved high rate of accuracy compared to the other approaches. Most of the studies in literature have concluded that Indians are more prone to diabetes. It is estimated that nearly 44 lakh Indian people are not aware that they are diabetic [12]. Hence, the main objective of the proposed technique is to forecast the future outcomes with a set of variables which can aid the experts in identifying the risk of developing a particular disease.

The Hybrid AM-GA

The AM and GA individually suffer from the problem of stagnation and delayed convergence. To overcome this, the AM and GA have been combined together to generate the best initial set of solutions from the search space such that the poor solutions are worn out and only the best solutions are given as input to for better optimization. Algorithm I and II represents the steps of AM and GA for feature selection and classification.

The PIMA Indian Type II diabetic dataset was obtained from the UCI machine learning repository. This dataset is a huge database of 768 instances (women of age 21) with 8 predictor attributes. The dataset consists of one target attribute which indicates whether the particular patient has been diagnosed with diabetes (1) or not (0). The ultimate aim of the proposed work is to build a machine learning model to predict the occurrence of type II diabetes. The problem is identified as a classic supervised binary classification.

The real world data may not be clear, as it contains lot of missing attributes, noisy data, duplicate or invalid data. This increases the complexity of data classification and prediction. In literature, there are lot of data preprocessing techniques like which has to be done before the application of the algorithm.

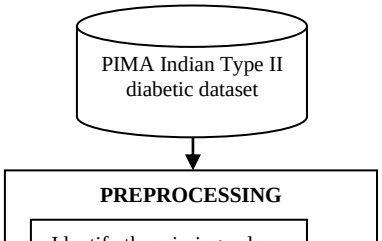

Fig 1.Preprocessing the Input Dataset

The first and foremost step in preprocessing is to identify the missing the values in the dataset. There are two forms of missing values: (1) Missing completely at random: This form does not affect the hypothetical value of the data nor the other variables (2) Missing at random: This form affects the hypothetical value of the data and other variables. Eg: People with high income may not reveal their salaries and some of the women may not be willing to share their ages. After identification of the missing values, the attributes and instances of the dataset are converted into an ASCII file format known as the Attribute Relational File Format (ARFF).The ARFF segregates the dataset into two distinct sections: Header and data section. An example on PIMA dataset:

```
@HEADER
@ RELATION diabetes
@ATTRIBUTE Age NUMERIC
@ATTRIBUTE Number of times pregnant NUMERIC
@DATA
   21,4,diabetic
```

The ARFF helps to partition the dataset into two distinct clusters as diabetic and non diabetic [13]. The main advantage of clustering is, it deeply analyzes the relationship between the attributes of the dataset. The data uncertainty (irrelevant attributes in the cluster) is identified by calculating the value of entropy and relevance measure [1].The attributes are grouped as predecessor (Input: attributes and values) and successor (output: diabetic/non diabetic).

Feature Selection Using Ant Miner:

In the field of data mining, the social and cognitive parameters are supported to discover the exploration and exploitation in a search space [14]. The popular Swarm-based technique is the Ant Miner (AM). In real world, the ants are brilliant in finding the shortest path between nest and food using the substance called performance. This mechanism of real ant was used to resolve the problem of combinatorial optimizations as it involves the principle of cooperation and adaptation.

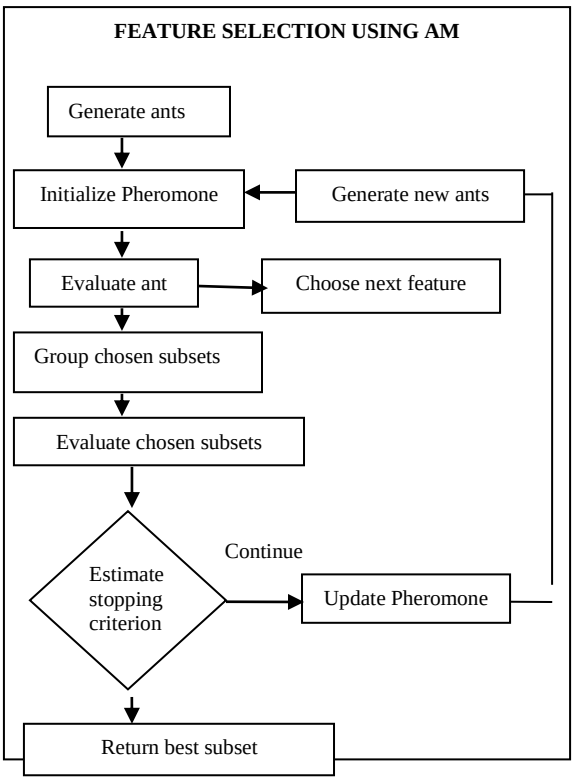

Fig 2: Feature Selection Using AM

In the proposed work, the mechanism of AM is considered as follows,

1) The path traversed (attributes) by the ant is linked with a objective solution (diabetic or not).

2) The quantity of pheromone deposited on the particular path is directly proportional to the objective solution i.e the objective solution highly depends on the probability of the attribute being chosen.

3) When the ant has to select between two or more paths (two or more attributes), the path with the highest precedence is chosen. i.e the path previously traversed by the other ant has the maximum amount of pheromone deposited in the path.

In Ant-Miner algorithm, each ant constructs and modifies a solution for attaining the objective. In

the proposed work, the objective is to obtain classification rules for the prediction of type II diabetes.

If <att 1 AND att 2 AND ….>THEN<diabetic>

Here each "if then" rule corresponds to a triplet of attribute, operator and value. This rule generation of ant miner supports both categorical and continuous values.

Algorithm 1: Feature Selection Using Ant – Miner

TS={all cases}; // Entire Training Set
RL=[];// Rule List Empty
While (TS> max_uncovered_cases)
J=1;//ant index
K=1;// rate of convergence
Initialize with equal amount of pheromone in all trials
Repeat
Ant_i=empty rule
Current rule = current rule +1// constructs rule incrementally
$R_i=R_i+1$
Update pheromone in all trials
$Ant_i=R_i$
Decrease pheromone in other trials
$R_i=R_i-1$//convergence rate updated
Then k=k+1
Else k=1;
End if
J=J+1;
Until (i>number of ants) OR (k>no rules convergence)
Choose R_{best} //Best Rule
Add R_{best} to RL[];
TS = TS – {set of cases correctly traversed by R_{best}};
End.

Classification using GA:

In GA, the chromosome is usually considered as a rigid string of genes (bits) with each gene corresponding to a candidate solution. i.e Each gene can take on a value of either) or 1.The GA generates the best solution using fitness evaluation, selection, crossover and mutation[15]. Every chromosome is compared with target function and the fitness value is estimated.

The fittest chromosomes are selected for the crossover and mutation operations. The genes are swapped (recombined) between two or more individuals who are called the parents. For instance, consider two individuals ('A' and 'B') performing crossover with a string of six genes.

Parent individuals before crossover:

$A_1\ A_2 \mid A_3A_4A_5A_6$

$B_1\ B_2 \mid B_3B_4B_5B_6$

Child individuals after crossover:

$A_1\ A_2 \mid B_3B_4B_5B_6$

$B_1\ B_2 \mid A_3A_4A_5A_6$

In the above string of genes, each character corresponds to a rule set of prediction and each gene symbolizes the rule condition. Thus the crossover operation swaps the rule condition between two prediction rules yielding a new solution. The other type of crossover is known as the uniform crossover in which the position of each gene is fixed to a certain probability 'p'.i.e any number of genes can be swapped between two parents when the probability is high.

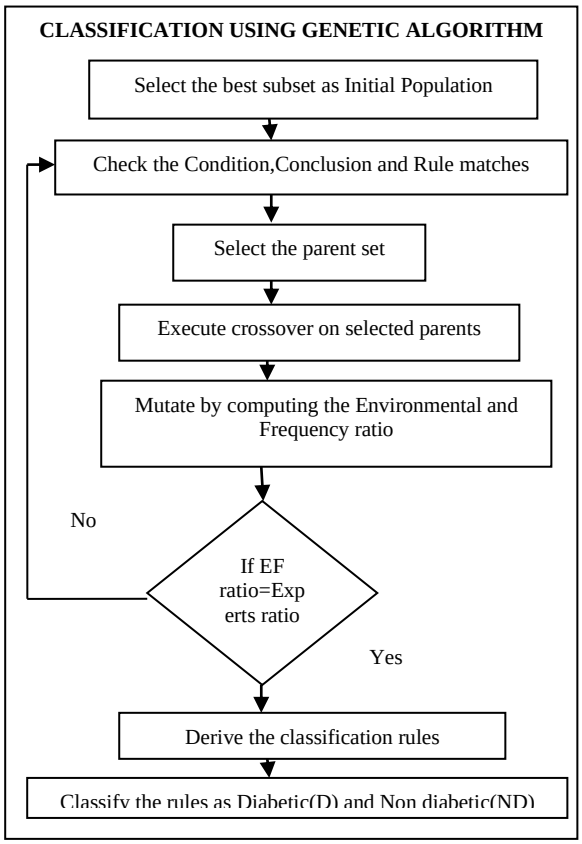

Fig 3: Classification using GA

Uniform crossover:

Before crossover:

$A_1\ \mathbf{A_2}\ A_3\ \mathbf{A_4}\ \mathbf{A_5}\ A_6$

$B_1\ \mathbf{B_2} B_3\ \mathbf{B_4}\ \mathbf{B_5}\ B_6$

After crossover:

$A_1\ \mathbf{B_2}\ A_3\ \mathbf{B_4}\ \mathbf{B_5}\ A_6$

$B_1\ \mathbf{A_2} B_3\ \mathbf{A_4}\ \mathbf{A_5}\ B_6$

The advantage of uniform crossover is that, the swapped genes may not be adjacent to each other. Hence the crossover operator intuitively depends upon the position of the gene. In the rule discovery procedure, each character corresponds

to 'IF-THEN' rule and each gene symbolizes the 'attribute- value' pair of the IF part(rule antecedent).The rule antecedent is a set of unordered conditions i.e the position of the attributes are not fixed. Here the uniform crossover plays a vital role where the attribute-value pairs are swapped independent of their respective position in the gene

The operator mutation works only on a single chromosome in a given period of time by replacing the value of a gene with random value. The best part of mutation is that it can introduce a new individual into the population set even if it doesn't belong to the particular set.

Algorithm 2:
Classification using Genetic Algorithm
TS = {all cases}; // Entire Training Set
RL = [ ]; // Empty Rule List
WHILE (TS > Max_uncovered_cases)
Initialize population as set of n rules
$\{A_i, i=1,2,3\ldots..n\}$
Evaluate fitness function f for each rule
Best fitted rules bestf(A_i) are extracted
From the bestf(A_i)two rules are selected as
Parent 1= bestf(A_i), Parent2= bestf(A_i) ;
Generate Offspring using uniform crossover operator
Add rule Rbest to DiscoveredRuleList;
Repeat the process for the whole training set
TS = TS - {set of cases enclosed by R_{best}};
End

Thus in the proposed approach, the AM employs the entire set of features from the dataset and selects an optimal feature subset. These optimal features are given as an input to the GA. The fitness and objective functions are evaluated and the best predictive results are obtained. The algorithm is repeated until maximum possible accuracy is reached.

RESULTS AND DISCUSSIONS

The benchmark type II diabetic datasets PIMA, LOZHOU, KOGES, Ar and GERIATIC were taken for analysis. Each dataset was divided into the training and test data using 10 fold cross validations. The process of classification was implemented using WEKA Software and the proposed AM-GA algorithm has been implemented using Net Beans IDE.

TABLE I: Performance analysis of data mining methods w.r.t Accuracy

METHODS	DIFFERENT DATASETS USED FOR PREDICTION				
	PIMA	LOZHOU	KOGES	Ar	GERIATIC
PROPOSED AM-GA	98.92	97.34	97.21	98.12	98.32
LINEAR REGRESSION	77.82	78.12	78.87	79.32	81.24
DECISION TREE	89.53	89.12	90.14	90.24	90.32
LOGISTIC REGRESSION	92.12	92.11	92.32	93.43	93.64

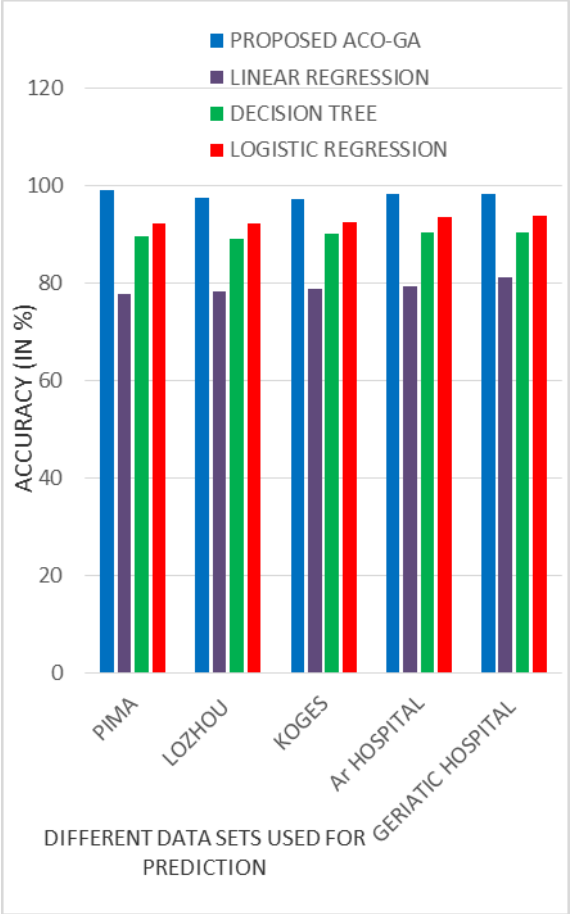

Fig 4: Performance Analysis of Data Mining Methods w.r.t Accuracy

The accuracy rates of various methods respective to their datasets are represented in table. The table clearly highlights that the proposed AM-GA approach has achieved high rate of accuracy compared to the traditional approaches like linear regression, decision tree and logistic regression

irrespective of the datasets being used. It is to be noted that the proposed approach has significantly reached the highest rate of accuracy (98.92%) nearly equal to 100% with PIMA Indian diabetes dataset. It is also to be noted that lowest rate of accuracy was reached when the classifier linear regression was used (77.82%).Figure 2 segregates the performance accuracy of different datasets over different classifiers used for the prediction of type 2 diabetes.

CONCLUSION

The medical services area appeared to be a significant field for creating an immense amount of information in different structures which can be helpful to foresee various infections and furthermore help doctors to take clinical choices. In this paper, the set of features has been optimized by ant colony optimization and genetic algorithm which has resulted in increased classification accuracy and less computational complexity. The result of the proposed model has been compared with other optimization techniques like linear regression, logistic regression and decision tree. The results strongly suggest that AM with GA can aid in the prediction of diabetes. In future, the proposed method might be utilized on continuous applications that need improvement.

REFERENCES

[1] D.Y. Eroglu, K. Kilic., 2017, A novel Hybrid Genetic Local Search Algorithm for feature selection and weighting with an application in strategic decision making in innovation management, Information Sciences, vol.405, (pp.18-32)

[2] Hsiao-Hsien Rau et al., 2016. Development of a web-based liver cancer prediction model for type II diabetes patients by using an artificial neural network, computer methods and programs in biomedicine, vol.125, pp.58–65,

[3] M.F. Ganji, and M. S. Abadeh., 2011. A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis, Expert Systems with Applications, vol. 38, pp. 14650–14659.

[4] E. Dogantekin, A. Dogantekin, D. Avci, and L. Avci., 2010. An intelligent diagnosis system for diabetes on Linear Discriminant Analysis and Adaptive Network Based Fuzzy Inference System: LDA-ANFIS, Digital Signal Processing,vol. 20, pp. 1248–1255.

[5] X.J. Fu and L.P. Wang., 2001 Rule extraction by genetic algorithms based on a simplified RBF neural network, Proceedings of the 2001 Congress on Evolutionary Computation, pp.753-758

[6] Kandhasamy, J. P., & Balamurali, S., 2015. Performance Analysis of Classifier Models to Predict Diabetes Mellitus. Procedia Computer Science, 47(1), 45-51.

[7] Kalaiselvi, C. and Nasira, G.M., 2015. Prediction of heart diseases and cancer in diabetic patients using data mining techniques. Indian Journal of Science and Technology, 8(14).

[8] Karthikeyan, T. and Vembandasamy, K., 2015. A novel algorithm to diagnosis type II diabetes mellitus based on association rule mining using MPSO-LSSVM with outlier detection method. Indian Journal of Science and Technology, 8(S8), pp.310-320.

[9] Karthikeyan, T. and Vembandasamy, K., 2015. A novel algorithm to diagnosis type II diabetes mellitus based on association rule mining using MPSO-LSSVM with outlier detection method. Indian Journal of Science and Technology, 8(S8), pp.310-320.

[10] Aljumah, A.A., Ahamad, M.G. and Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University-Computer and Information Sciences, 25(2), pp.127-136.

[11] Chen, W., Chen, S., Zhang, H. and Wu, T., 2017, November. A hybrid prediction model for type 2 diabetes using K-means and decision tree. In Software Engineering and Service Science (ICSESS), 2017 8th IEEE International Conference on (pp. 386-390). IEEE.

[12] Varma, K.V., Rao, A.A., Lakshmi, T.S.M. and Rao, P.N., 2014. A computational intelligence approach for a better diagnosis of diabetic patients. Computers & Electrical Engineering, 40(5), pp.1758-1765.

[13] Pavate, A. and Ansari, N., 2015, September. Risk Prediction of Disease Complications in Type 2 Diabetes Patients Using Soft Computing Techniques. In Advances in Computing and Communications (ICACC), 2015 Fifth International Conference on (pp. 371-375). IEEE.

[14] Mekruksavanich, S., 2016, August. Medical expert system based ontology for diabetes disease diagnosis. In Software Engineering and Service Science (ICSESS), 2016 7th IEEE International Conference on (pp. 383-389). IEEE.

[15] Guo, Y., Bai, G. and Hu, Y., 2012, December. Using bayes network for prediction of type-2 diabetes. In Internet Technology And Secured Transactions, 2012 International Conference for (pp. 471-472). IEEE.