**Available Online at www.ijcsmc.com**

**International Journal of Computer Science and Mobile Computing**

**A Monthly Journal of Computer Science and Information Technology**

**ISSN 2320–088X**

***IJCSMC, Vol. 4, Issue. 6, June 2015, pg.177 – 186***

**RESEARCH ARTICLE**

# A BFO Based Rule Mining Approach for Diabetic Patient Identification

**Neha Gulia[1], Sugandha Singh[2], Luxmi Sapra[3]**

[1]Computer Science, PDMCE Bhadurgarh, India
[2]Computer Science, PDMCE Bhadurgarh, India
[3]Computer Science, PDMCE Bhadurgarh, India
[1] gulia1neha@gmail.com, [3] luxmi_engg.pdm.ac.in

***ABSTRACT: Medical Disease Prediction is the major and critical data mining application area comes under expert system. This application area requires identifying the critical rules to predict the medical disease. In this paper, A BFO integrated rule mining model is presented to identify the diabetic prediction. The presented model has performed the rule filtration and rule pruning to identify the disease possibility. The experimentation results show that the effective identification of diabetic disease is obtained from the algorithmic approach.***

## 1. INTRODUCTION

Data Mining is sometimes represented as the statistical technique that takes the decision based on historical data. But instead, it is defined as a science to find the input pattern over the large datasets. There are number of application of data mining that separate it from the mathematical or the statistical model and place it under a science or the intelligent system category. The data mining is about to derive the useful information from the available raw set of information. It has its importance in many potential applications in different areas such as financial field, medical care systems etc. In short we can say there is no field that does not require the data mining operations. In business it is being used in many application such as eligible loan candidate detection, fraud detection etc. It is used in network applications, attack detection, load detection, ideal user detection etc

### 1.1 Data

To understand data mining the first concept is to study data. The data is about everything that we see in the form of some information. It can be a digit, text, image, video audio etc. Each system process on dataset. Each system accepts some input and return some result. This input - output is again the data forms.

 

## 1.2 Knowledge Discovery in Databases (KDD) Process:

The knowledge based system includes the intelligent system process that defines the basic data modeling process stages. The stage includes the generation of dataset, selection of relevant attributes from dataset, dataset cleaning etc. All these are taken by most of the data mining applications as the preprocessing stages. As the input is taken from the primary or the secondary sources, it requires lot of modification to convert it to some usable form. This conversion is been done by using the KDD process.

The basic work of KDD process begins as the raw data is accepted from some warehouse. It is the raw dataset which includes the internal and the external dataset [2]. The background information and the meta information is also present with the system. In the next stage, the warehouse is been implement with the relational database concept to provide the optimization to the flexibility of the dataset.

## 1.3 Data Mining

Data mining approaches are the outcome of a huge process model of research and product development. These methodologies define the evolution to the system with process definition along with data specification. In such model, not only the methodology is important but all aspects of data are important. The mining process begins as the dataset search or the generation is defined. Data mining is defined for the specific application respective to the business community.

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

In the field of commercial database system the mining operations is increased very fast. Based on a survey, the data warehouse projects are found to 19% of respondents are beyond the 50 gigabyte level, whereas 59% expect to be there by second quarter of 1996[1].There are number of industries, where these numbers can be much larger. The basic need of improved computational engines shows the importance of data mining todefine in a cost-effective manner along with parallel computer system. It includes the data mining approaches that have existed from more years and also been improved in near future under the different statistical definitions and the featured criteria of reliable, mature, understandable approaches.

## 1.4 The Scope of Data Mining

Data mining is about to identify the similarities between searching the valuable business information from the large database systems such as finding linked products in gigabytes of store scanner data or the mining a mountain for a vein of valuable dataset. Both kind of processes required either shifting through an immense amount of material, or to perform the search intelligently so that exactly match will be performed[4]. Data mining can be done on a database whose size and quality are sufficient. The technology of data mining can generate new business opportunities by providing these capabilities:

- Automated prediction and analysis of various trends and behaviors - Data mining itself automate the process by obtaining the predictive information from large databases. It first setup the questions and then provides the relative solutions. A typical example of such a predictive system is in the marketing field. Data mining uses data on the historical promotional mailings to capture the targets effectively so that the maximum return from

 

market will be achieved. Other predictive problems include the detection of bankruptcy and or the frauds.

- Another application of data mining is Automated discovery of historical patterns dynamically. The presented Data mining system is able to sweep over the databases to identify the hidden patterns. One of such example of pattern discovery is the analysis of retail sales data to identify the seemingly unrelated products so that the effective purchase can be done. Other pattern discovery analysis includes the detection of fraudulent credit card as well as the transactions to identify the anomalous data.

## 1.5 Types of Data Mining

Different types of Data Mining are given as

1) Predictive Data Mining
2) Descriptive Data Mining

Predictive data mining creates a model system described by a given set of data.

Descriptive data mining gives new and unique information inferred from the available set of data.

### 1.6 Applications

There are number of such applications of data mining in the area of Medical care industry. Some of such examples are given as under

- To avoid the member attrition in the medical insurance industry , so that the patterns in the existing dataset to be used and increase the actual member population
- The another problem is such system is the Fraudulent claims of medical Insurance can be determined. In such cases some classification and regression based approaches are used along with improved data sub systems
- There are number of such members who are involved in additional insurance coverage, can be identified from such system.

# 2. Literature Review

Adepele Olukunle[1] has defined a work on association mining under medical imaging. Author defined the work on fast association rule based mining under algorithmic approach. Author defined the medical image dataset processing for disease identification. Author defined the analysis on disease prediction so that analytical results will be derived. Author defined the algorithmic solution to identify the disease.

Carlos Ordonez[2] has defined a work on rule discovery under testing and training set for heart disease prediction. Author defined the algorithmic approach under association rule mining to identify the heart disease. Author defined a constraint specific approach for rule generation so that the reduced rule based association mining will be attained. Author validate the work under different component set so that the effective rule discovery will be obtained from the analytical work..

Gaurav N. Pradhan[4] has presented a work on multi-dimensional time series to identify the

 

disease on medical data. Author defined the analysis on identification of participants obtained from EMG analysis under quantitative behavior and pattern discovery applied on data streaming. Author defined the energy adaptive signal analysis to derive the time series pattern. Author defined a two stage model to discover the frequent patterns over multiple data values so that more effective information will be derived. Author defined the multi dimensional environment so that the association rule generation will be done. Author defined the data series based analysis for effective information analysis.

K.Ravikumar[5] has defined a work on traffic risk analysis using ACO approach. Author defined the accident risk analysis under decision tree adaption so that data description based analysis swill be obtained from the work. Author defined the accidental risk analysis sunder data mining approach. Author defined the pragmatic approach under multi layer propagation so that the relation information processing will be done. The quality analysis and trend pattern discovery so that the efficient information processing based decision tree derivation will be obtained. Author defined the intelligent information processing to derive the heuristic results.

Wei Wang[6] has defined a work on association mining under medial data processing under concept level analysis. Author defined the method driven approach for analysis so that the discritetized results will be processed and information gain will be obtained from the work. Author derived the interesting association rules for the processing.

Mostafa Fathi Ganji[7] has defined a parallel fuzzy rule based learning mechanism under ACO approach so that medical processing will be improved. Author defined the rule based system using ACO and fuzzy approach. Author defined the heuristic method to derive the effective mapped results under probability analysis. Author defined uniformity under rule testing so that the fuzzy cover will be defined. Author defined the meta heuristic algorithm to derive the rules for effective data processing.

Pooia Lalbakhsh[8] has presented a work on rule quality analysis and evaporation so that the ACO improved rule mining will be obtained. Author used the ant effective approach to derive the rules so that the dynamic pheromone evaporation will be obtained. Author defined the accurate rate effective rule discovery.

Ghada Almodaifer[9] has defined a medical association mining based medical data processing so that the association rules will be obtained from medical datasets so that more predictive results will be obtained. Author defined the rule discovery under numerical form and categorical featured data processing.

Qiaoling Duan[10] has defined a work on association mining under multi data base processing. Author defined the indirect rule generation under rule mining approach. Author obtained the correct rules from the work.

P. Kasemthaweesab[11] has presented a work on diabetes to identify the disease under association rule mining. Author defined the work for rule discovery so that more effective results for diabetes detection will be obtained. Author considered muitliple parameter including gender, age and occupation for disease diagnostic.

K.Rameshkumar[13] has defined a work on association rule mining to provide data fileration. The relevancy adaptive rules are generation under association rule mining with specification of transaction set. Author defined the partition adaptive approach for validating the association rules.

Chen[15]presented a analytical behavior analysis approach for predictive system that can be assisted by the medical professionals to identify the heart disease based on the medical information of patients. The presented approach is defined in three main steps given as- In very

 

first stage, the extraction of important features is done. Author taken the decision on disease prediction based on the signal specific constraints with multiple vectors. These vectors includes ECG signal form, slope analysis, heart rate, blood sugar, peak analysis etc. Once the statistical constraints are collected, the neural network model is applied to predict the disease under different feature vectors.

Mrs.G.Subbalakshmi introduced a parameter based probabilistic mechanism to identify the heart disease. Author defined the decision support system based intelligent work model in which naïve bays is used as the classifier. The work is divided in two sub stages, where at first the parameters specific information is collected. These parameters includes blood pressure, sugar, gender, age etc. After generating the statistical dataset, the probabilistic analysis is applied on training and testing sets using naïve bayes approach. The system is implemented by the author as web based questionnaire application. The system is defined as a training tool to predict the possibilities of heart disease.

E. Barati has defined a survey to analyze the medical health care based diagnostic to introduce the data mining to define different data mining approach and provide a brief survey on different classification and predictive approaches. Author analyzed the identification of skin problem under different classifiers. Author obtained the intelligent decision to analyze the various disease aspects so that the disease prediction at the early stage will be obtained.

Milan Kumari presented a study oriented work on different classification algorithm respective to heart disease prediction. The algorithms studied in this paper includes decision tree, artificial neural network, SVM, decision table etc. Author obtained the analytical behavior of all these approaches and provided the functional difference. Author also considered various comparative factors under standard dataset specification. The result analysis is done under accuracy, error rate, and positive rate analysis vectors.

Jyoti Soni defined the data mining approach to perform the analysis on heart disease system. It can intends to provide a survey on different techniques of information retrieval in databases under different approaches that can be used in medical research to perform the analysis on heart disease and to predict the chances of disease occurrence. The author performed the number of such experiments that can be conducted to generate various performance measures so that the analytical results will be derived. Author identified that the decision tree provided the most significant results for this specific database.

Mr. Dhiraj Pandey defined a predictive algorithmic approach for symptom based disease prediction. The defined system and the approach is more extendable and improved for the given dataset generation and developed. There are number mining approaches and the relationships defined with symptoms and to analyze the disorder symptoms in a patient. The author has defined an approach that will produce the hybrid association rules based mining and to define the rules that are displayed in form of tables and graphs.

Jyoti Soni defined another work on Intelligent and Effective Classification system for the prediction of heart disease. The author has defined the weightage to the associated rules. Author has provided a graphical interface to accept the valid symptom dataset and to perform the disease prediction based on the rule level classification. The work is here defined to generate the weighted rules from raw information set and to provide the predictive decision from the dataset. Author improved the decision predictive capabilities so that more accurate decisions from the information set will be taken.

Dr. D. Raghu defined a research on the probabilistic analysis for heart disease so that the identification so that the heart disease analysis is done. Author defined the parameter

 

specification so that the parameter specification based prediction so that the dataset feature analysis under different disease parameters will be done. Author defined the information likelihood analysis for heart disease derivation so that the disease predictive conclusion will be derived.

Shantakumar B.Patil defined a neural network based intelligent mining system to predict the heart disease of the patient. The presented work to the system is defined under an intelligent and effective approach so that the significant disease patterns will be identified and the heart attack chances will be discovered at the earlier stage. The author has defined the multi layer perceptron for the back propagation to improving the prediction of disease. The obtained results actually identify the disease prediction results are more accurate[27].

M.A.Jabbar has defined a work on Knowledge discovery system under the mining association rules for the heart disease prediction. The author also defined the association mining based approach to discover the information from medical data to predict heart disease for Andhra Pradesh. This defined approach is expected to help physicians to make more accurate decisions.

T Srinivasan has defined the neural network based work to perform the decision of clinical disease based on evidence based analysis. Author presented the classification on different neural models. Here three different parametric configurations are applied in experimentation to analyze the learning behavior under neural approach. Author also performed the dumpster shafer based evidence analysis to improve the accuracy of recognition process based on different disease and disorder parameters.

Sellappan Palaniappan defined a data mining approaches based intelligent health disease prediction system. The medical care industry has defined a large patient dataset to analyze the patient information and to take the decision regarding the patient disease. The defined approach includes the specification of heart disease with symptom analysis with classification vector under decision tree and neural network approach.

K. Rajeswari defined a work on Medical Data Mining. The defined approach is about to find the interesting information to take the decision on medical data. The presented approach is about to derive the useful information for heart disease dataset. The author has defined a predictive analytical system to analyze the risk criticality based on the score estimation applied on the dataset of various years.

Smitha.T has defined a research work to focus on a data mining approaches so that the objective of creating a prediction model by using the decision tree for predicting the occurrences of diseases in any area, mainly in slum.

Milan Kumari defined work as a study framework applied on different classification algorithms to predict the cardiovascular disease and relative diagnosis. Author defined the analysis on the classification behavior applied on the decision tree and neural network approach. Author defined the study on the analytical behavior to predict the error chart and to improve the rift rate so that the recognition rate will be improved. Author analyzed various analytical approach in different parameter specific aspects.

K.Srinivas provided the analytical study on different data mining approaches and processes under different datasets. Author analyzed the heart disease and disorder identification under different classification algorithms. These algorithms are rule specific and applied on larger information set.

 

# 3. Proposed Work

## 3.1 Problem Definition

The Knowledge Discovery in Databases (KDD) field of data mining is concerned with the development of methods, techniques and algorithm which can make sense of the available data. KDD is useful in finding trends, patterns, correlations and anomalies in the databases which is helpful to make accurate decisions for the future. Association rule mining finds collections of data attributes that are statistically related to the data available. In this present work, A BFO (Bacteria Foraging Optimization) improved association mining approach is suggested to perform the association mining on medical dataset. The work will be applied on diabetic identification or breast cancer identification under the symptom analysis. In this present work, a layered approach will be applied to predict the disease. At the earlier stage of work, the available dataset will be represented in the form of network. On this network represented dataset, the BFO approach will be applied under the bacterial phenomenon. In the second stage of this work, the rules will be identified based on the predictive analysis by tracking the life of distributed bacterias. This stage will identify the complete possible rule set Once the rules will be generated, at the final stage, the pruning of these associated ruleset will be performed to identify the most effective rules over the dataset

## 3.3 Algorithm

### 3.3.1 Rule Generation Algorithm

The main work associated in this work is to generate the rule set by processing the adaptive diabetic dataset. The work is here defined using BFO based approach for rule identification under association analysis. The rule generation algorithm associated with this work is given here under

RuleAdaptiveAlgorithm(PatientSet)
/* In this work, diabetic patient dataset is considered for disease identification. */
{

1. N=GetInstances(PatientSet)
   [Get the number of disease instances from the dataset]
2. Set class=GetClasAttribute(PatientSet)
   [The ruleset discovery will be here performed respective to class specification under rule formation]
3. Set the parameters for bacterial implementation
   a. NumberOfBacterias
   b. SwimVector
   c. DirectionalAspect
   d. Positional
4. Generate the parametric notification so that the network will be formed from the instance set.
5. Transformed the available dataset in the form of graph network using the parametric specification with rule specification and thresholding.
6. Set specification is here given weight generation so that the role of participating attribute will be identified along with contribution value.
7. Obtain the dataset weight with Ant network specification and attribute set formation with instance set initialization.
8. Obtain the Weighted Ant Network from the Attributeset and InstanceSet
9. For Iter=1 to MAX

 

/* Process the network for maximum number of defined iterations so that the rule generation and filtration will be obtained. */
{
10. Perform the weight adaptive analysis so that the selection or rejection of attribute will be done.
11. The line of sight under bacterial life is analyzed to identify the effective attribute that can participate in the weight adaptive mapping.
12. The cost adaptive analysis is here defined to generate the attribute matrix under bacterial parameter
13. Identify the cost of rule
14. If cost is adaptive select the rule in the dataset
15. RuleSet.Add(AdaptiveRule);
    }
16. Return RuleSet
17. }

}

### 3.3.2 Rule Filteration

Another work associated with this research work is the formation of the rule set based on the analysis applied to perform the selection of adaptive values based on which the rule selection can be performed

RuleSelection(RuleSet)

/*Rule set is the adaptive obtained at the earlier stage based on which the fileration is performed to identify the adaptive rule*/

{

1. N=GetRules(RuleSet)
   [Get the number of available Rules]
2. Define the decision parameters for rule selection called support and confidance threshold
3. For i=1 to N
   /*Process All rules *.
4. {
5. r=GetRule(RuleSet,i);
   [Get the rule from ruleset]
6. Attr=GetAttr(r)
   [Get the participating attributes in the ruleset
7. For J=1 to Attr.length
   [Process all the attributes in the ruleset]
8. {
9. For K=1 to Attrib.length
   [Process all the rule attribute respective to other attribute]
10. {
11. If (ObtainWeight (Attrib(J)), ObtainWeight(Attr(K))
    [Map the weight value respective to which the attribute analysis is performed in the dataset]
    {
12. Count=Count+1
    [Count the available attribute with adaptive weight]

 

```
    }
13. }
14. }
15. If (Count>SThreshold)
    [If the associativity is adaptive under support vector]
16. {
17. If(Count>CThreshold)
    [If associativity is adaptive under confidence vector]
    {
18. FinalRuleSet.Add(r)
    [Include the rule in ruleset]
    }
19. }
20.
21. Return FinalRuleSet;
22. [Obtain the filtered ruleset]

}
```

### 3.4 BFO

Input the bacterial foraging parameters and independent variable, then specify lower and upper limits of the variables and initiate the elimination-dispersal steps, reproduction and chemotactic.

1. Generate the positions of the independent variable randomly for a population of bacteria. Evaluate the objective value of each bacterium.
2. By using the tumbling or swimming process, alters the place of the variables for all the bacteria. Perform reproduction and elimination operation.
3. If the maximum number of chemotactic, reproduction and elimination-dispersal steps is reached, then output the variable corresponding to the overall best bacterium; Otherwise, repeat the process by modifying the position of the variables for all the bacteria using the tumbling /swimming process[13]

1 Firstly Start, there are three principal mechanisms namely, chemo taxis, reproduction and elimination-dispersal.
2. Initialize and then evaluate them.
3 Evaluate moving tuble/swim.
4 Reproduce the best adapted bacteria tend to survive and transmit their genetic characters to succeeding generations
5. After that elimination-dispersal, select parts of the bacteria population to diminish and disperse into random positions in the environment.
6 End

# 4. Conclusion

In this presented work, BFO integrated rule mining model is presented for diabetic disease prediction. The experimentation is performed on real time dataset. The results show that the clear identification of disease prediction is obtained from the work.

 

# References

[1] Adepele Olukunle," A Fast Algorithm for Mining Association Rules in Medical Image Data", Proceedings of the 2002 IEEE Canadian Conference on Electrical & Computer Engineering 0-7803-7514-9/02@2002 IEEE

[2] Carlos Ordonez," Association Rule Discovery With the Train and TestApproach for Heart Disease Prediction", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 10, NO. 2, APRIL 2006, 1089-7771 © 2006 IEEE

[3] Chunxue Shi," Path Planning for Deep Sea Mining Robot Based on ACO-PSO Hybrid Algorithm", 2008 International Conference on Intelligent Computation Technology and Automation

[4] Gaurav N. Pradhan," ASSOCIATION RULE MINING IN MULTIPLE, MULTIDIMENSIONAL TIME SERIES MEDICAL DATA", ICME 2009 978-1-4244-4291-1/09©2009 IEEE

[5] Mr.K.Ravikumar," ACO based spatial Data Mining for Traffic RiskAnalysis".

[6] Wei Wang," Mining Association rules in Medical DataBased on Concept Lattice", Proceedings of the 8th World Congress on Intelligent Control and Automation July 6-9 2010, Jinan, China 978-1-4244-6712-9/10©2010 IEEE

[7] Mostafa Fathi Ganji," Parallel Fuzzy Rule Learning Using an ACO-BasedAlgorithm for Medical Data Mining", 978-1-4244-6439-5/10©2010 IEEE

[8] Pooia Lalbakhsh," Focusing on Rule Quality and PheromoneEvaporation to Improve ACO Rule Mining", 2011 IEEE Symposium on Computers & Informatics 978-1-61284-691-0/11©2011 IEEE

[9] Ghada Almodaifer," Discovering Medical Association Rules from Medical Datasets", 978-1- 61284-704-7/11 ©2011 IEEE

[10] Qiaoling Duan," Mining Indirect Association Rules in Multi-database", 2012 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization 978-1-4673-0915-8/12©2012 IEEE

[11] P. Kasemthaweesab," Association Analysis of Diabetes Mellitus (DM) WithComplication States Based on Association Rules", 978-1-4577-2119-9/12@ 2011 IEEE

 