ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 5, Issue 4, April 2016

A Survey on Hybrid Evolutionary Classification Techniques for Medical Diagnosis

Suman Muwal, Narender Kumar

Abstract*— Knowledge is the important part of human life. There are number of data mining systems that are available today and have many challenges in this field. Today, this is mostly used in the medical diagnosis area to conquer the diagnosis time and to enhance the performance. In medical system, heart disease is the main cause of death for both men and women. That's why early detection of heart disease is necessary. Different types of classification techniques are available in data mining to resolve this problem like Decision tree, C4.5, Bayesian networks, Neural networks, Support vector Machine, Association rule, K-NN, CART etc. Recently, there exist various soft computing techniques for the medical diagnosis like Genetic Algorithm, Ant Colony Optimization, Firefly Algorithm, Cuckoo Search, Artificial Bee Colony, Levy Flight etc. They are also combined with the various other techniques like rough set, fuzzy logic and neural network etc. This paper presents an extensive review of literature on various hybrid technologies used for medical diagnosis.

Index Terms— Decision tree, Metaheutistics, Association Rule, Artificial Neural Network, Rough set, Fuzzy Logic System, SVM

I. INTRODUCTION

Today, the intense lump in data due to the large scale mechanization and computerization of business, easily available and affordable hardware, software and data collection and management tools. The information is "Hidden" in this large data. To handle this huge amount of data and to get the hidden information from the data is a very time consuming process. Also, data is stored in different forms. To get the effectual and fruitful results from these different forms is a very challenging task. Therefore, to extract this hidden information from the data, data mining process is used.

***Suman Muwal**, M.Tech Student, Computer Science and Engineering Department, Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India.*

***Narender Kumar,** Assistant Professor, Computer Science and Engineering Department, Guru Jambheshwar University of Science & Technology, Hisar, Haryana, India.*

II. DATA MINING

Data mining is the process of extraction of useful patterns or knowledge from huge amount of data [2].

According to [1], "Data mining is defined as the process of discovering patterns in data". Various alternative names for data mining is used like Knowledge mining from data, knowledge discovery from data (KDD), knowledge extraction, data/pattern analysis, data archaeology, data dredging, information harvesting etc [2] . Different steps of KDD process are:

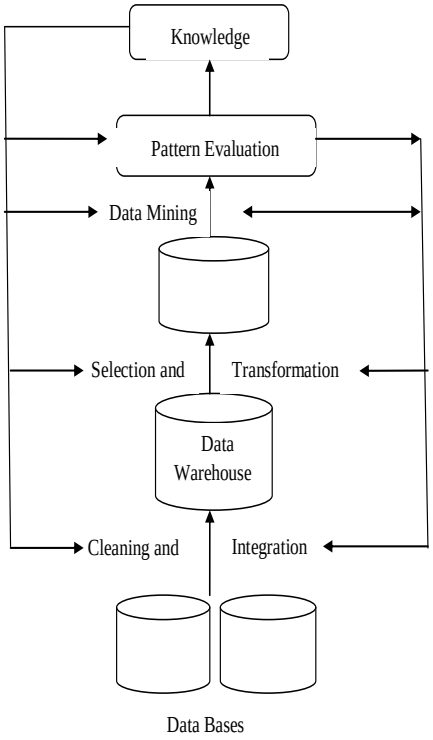

Fig1. Knowledge Discovery Process

1. Data Cleaning: remove noise and correct inconsistencies in data.
2. Data Integration: combination of multiple data sources.
3. Data Selection: selection of data from data base which is suitable to the analysis.
4. Data Transformation: with the help of aggregation and summary operations, data are transformed in the forms which are suitable for mining.
5. Data Mining: extract data patterns by applying intelligent methods.

6. Pattern Evaluation: truly interesting patterns are identified on the basis of some interestingness measures.
7. Knowledge Presentation: For envision of the useful information, the different representation techniques are used.

III. CLASSIFICATION IN DATA MINING

In data mining, According to [2], "Classification is the process of finding a model that describes and distinguishes data classes or concepts for the purpose of being able to use the model to predict the class of objects whose class label is not known". The model is based on the analysis of data objects whose class label is known. The derived model is represented in different forms like if-then rules, decision tree, mathematical formulae and neural networks. Classification process predicts the categorical labels. For example, a medical researcher wants to analyse breast cancer data in order to predict which specific treatment a patient should receive. i.e. whether the breast cancer is benign or malignant. Classification process consists of a two-phase. In first phase, a model is constructed to explain a predetermined set of data classes. This is the training step. In second step, this classifier or model is used for classifying future or unknown objects. The classifier accuracy is the percentage up to which model correctly classified the test samples.

IV. DIFFERENT TYPES OF CLASSIFICATION TECHNIQUES

In data mining, there are different methods which are widely used for classification of data. These are describes as follows:

1. Decision tree induction
2. Artificial Neural Network
3. Association rule analysis
4. Support Vector Machine
5. K-nearest neighbour
6. Bayesian Classification
7. Rule Based Classification

In data mining, there are some advanced methods which are also used for classification. These are:

1. Swarm Intelligence
2. Genetic Classifiers
3. Rough set approach
4. Fuzzy set approach

V. MEDICAL DIAGNOSIS

Very large and complex data are generated by the health care activities. Data mining generate information that can be useful to all in health care for effective treatment and health care management. Today, diagnosis of diseases is important and complex task. In medical system, heart disease is the main cause of death for both men and women. So, exact and efficient execution of this is necessary.

VI. RELATED WORK

The importance of artificial neural network models lies in the truth that they are used to derive a function from observations. When the complexity of the data or task makes the designs of a function by hand impractical, then this is useful. Swarm intelligence systems are well and relatively simple. The use of evolutionary algorithms or we can say that swarm intelligence with ANN is widely used by the various researchers. Arpit Bhardwaj and Aruna Tiwari [3] proposed a model for classification of the breast cancer which combined the Genetic Algorithm and artificial neural network. In this study, due to the destructive nature of simple crossover and mutation operators, authors replaced these operators with the modified crossover and mutation operator. The problem in GP in which the children of good parents generally have low performance than the parents is known as destructive nature. The Proposed algorithm (GONN) used to classify whether the breast cancer caused death or not death. To demonstrate that the GONN was better than other approaches, the authors used the accuracy, sensitivity, specificity, confusion matrix, ROC curve and AUC under ROC curve. In this paper, authors followed some rules in the function set and in the terminal to elaborate ANN with GP. In this, data is divided into training and testing data into four different partitions. The Proposed model compared with the BPNN and Koza and Rice [1991] model. GONN have the highest accuracy of 98.24%, 99.63%, 100% for 50-50, 60-40 and 70-30 training and testing partitions respectively and 100% for 10-fold cross-validation for 50 GP runs for 1-1-1 neural network.

Nguyen Cong Long et al. [4] proposed a hybrid model which integrated the rough set based attribute selection and the interval type 2 fuzzy logic with the nature inspired metaheuristics named as firefly algorithm. For handling the large data set and the uncertainties authors used the interval type 2 fuzzy logic system. A chaos firefly rough set based attribute reduction was used for the attribute reduction and the type 2 interval scaled fuzzy logic was used for the classification. For parameter tuning the interval type 2 fuzzy logic used the chaos firefly algorithm. This proposed model reduced the computational burden and enhanced the performance. The proposed model was applied on the two

medical data sets. The attribute reduction method was compared with the binary particle swarm optimization rough set based attribute reduction and the IT2FLS classifier was compared with the other classifiers like SVM, ANN and naïve bayes. The result showed that the chaos firefly rough set attribute reduction found the minimal attribute reduction from the large data set that helped in the improvement of performance of classification system. On the other hand, the combination of rough set based attribute reduction and the interval type 2 fuzzy logic systems overcomes others techniques in terms of accuracy, convergence speed and processing time.

H. Hannah Inbarani et al [5] proposed a hybrid model which integrated the rough set with particle swarm optimization and used classification technique for medical diagnosis. In this study, to find the minimal set of attributes authors used the PSO based relative reduction and the PSO based quick reduction. Then obtained result was used as inputs to classifier. Different classifiers were used like Naïve Bayer, BayesNet and KStar. The results showed that the attribute reduction using PSO based relative reduct and PSO based quick reduct have both have 15 out of 44 attributes. The best accuracy was 88% with Naïve Bayer classifier for SPECTF dataset.

Sashkala Mishra et al. [6] gave a new metaheuristic Bat-inspired classification approach for microarray data. Many metaheuristics algorithms are present in data mining for classification. Bat algorithm is one of the metaheuristic which solves the multi objective engineering problems. The authors proposed a model which is inspired by bats. Functional link artificial neural network's weights were updated with the help of the proposed algorithm. The proposed model is compared with the PSO-FLANN and FLANN and found that it gave the good accuracy than other two approaches**.**

Decision tree induction is very simple and efficient approach of classification which is easy to understand and also combined with other techniques easily. Association rule learning is a method for discovering interesting relations between variables in large databases. It is used for finding the frequent item sets from the data set. Today, various decision tree induction methods and the association rule mining algorithms are used for the research purposes. Murat Karabatak and M. CevdetInce [7] used the association rules and neural network for the diagnosis of breast cancer. To reduce the dimensions of database, association rule mining was used and for intelligent computation or classification, the neural network was used. Two-step process was used in the proposed work. In first step, the feature extraction and reduction with association rules was implemented and in second step, the reduced input in applied to the neural network structure and classified the breast cancer. For reduction of data set with association rule, AR1 and AR2 techniques were used. The proposed model was compared with the neural network. The AR2+NN gave the best performance with correct classification rate 95.6%.

SVMs can be used to solve various real world problems. Lots of work by using SVM in research has been done by the researchers. Sharifah Hafizah et al. [8] proposed a model which used the Artificial Neural Network and Support Vector Machine for the classification of Liver Cancer. Different Artificial Intelligence techniques are present in the data mining for the classification. Because of the good classification accuracy, the authors used the ANN and SVM for the classification of the cancer. They classified the cancer data as benign tumors or malignant tumors. This study compared the performance of ANN and SVM. The performance measures used in this study were accuracy, sensitivity, specificity and area under curve. For both techniques, four stages were used. These were developing classification model by input variable function, preprocessing and partitioning, setting model parameters and last stage is model implementation. The result showed that SVM classification was better than the ANN classification. In data mining different classification are present with their pros and cons. Different techniques are used by the researcher for the classification. Hybrid modeling is also used to increase the accuracy. Various metaheuristics algorithm or nature inspired algorithms are in data mining for the medical diagnosis or for classification.

Yuehjen E. Shao et al. [9] proposed a novel hybrid model for the classification of heart disease. In heart disease data set there are 75 explanatory variables and one class variable. But almost every study used 13 explanatory variables and one class variable. To address the entire classification problem by using a single technique may not be possible. To overcome this limitation, authors used the LR, MARS, ANN and rough set techniques. The purpose of the study was to calculate the classification performance of hybrid model which integrate the LR, MARS, rough set with ANN. ANN has the good capability to handle the complex nonlinear relationship among variables. In this research the LR, MARS and rough set was used for the reduction of attributes and then the resulting significant attributes were used for the ANN architecture inputs. The results showed that MARS-ANN model was the best model because in this model less number of explanatory variables were used and gave best classification accuracy.

Maya Dimitrova et al. [10] selected the most suitable method or algorithm for a particular medical data set. Different learning techniques or models were compared and found the

best model for the diagnosis of diseases. The result showed that Bayes classifier and SMO model gave the highest accuracy and the best correctly classified cases. Farhana Afroz et al [11] used the various classification techniques like MLP, Bayes net, Naïve Bayesian, J48 graft, Fuzzy lattice reasoning, JRip, fuzzy inference system, adaptive neuro fuzzy inference system on three data mining tools TANAGRA, WEKA and MATLAB. In WEKA, the best algorithm for diabetes diagnosis was J48 graft with accuracy 81.33%. In TANAGRA and MATLAB, naïve Bayesian and ANFIS were the best algorithms with 100% and 75.79% accuracy respectively. Overall comparison on accuracy of these three tools showed that TANAGRA tool was the best tool. Modjtaba Rouhani and Kamran Mansouri [12] used the different ANN architectures for the comparison of the Thyroid diseases diagnosis. Today, in the diagnosis and prognosis of different diseases the ANN and SVM plays a vital role. In this paper, the authors used the several types of ANN architectures and the SVM classification techniques. The different neural network architectures used are radial basis function, general regression neural network, linear vector quantization and probabilistic neural network. In terms of accuracy, the PNN and the RBF gave the best performance. They obtained the overall accuracy of 96%.

Mrs S.M Uma and Dr.E.Kirubakaran [13] used the Intelligent Heart Diseases Prediction System Using A New Hybrid Metaheuristic Algorithm. For solving the health care problems, there is various classification techniques are present in data mining. In this study, the authors proposed a model based on the swarm intelligence which integrated the two swarm intelligence techniques. These were Ant colony optimization and the genetic algorithm. The proposed algorithm was applied on four data sets and compared with the single ACO and the single GA and got that the hybrid ACO/GA algorithm performed well in terms of classification accuracy.

Omar S. Soliman and Eman Abo ElHamd [14] suggested a model for Diagnosing Diabetes Mellitus which used a chaotic Levy Flights Bat Algorithm. Different types of metaheuristics or natured inspired algorithms are present in the swarm intelligence or in data mining. The authors combined the two metaheuristics in their research. They used the bat algorithm with the chaotic levy flights due to the limitation of the bat algorithm. Bat algorithm is one of the meta-heuristic that follows the echolocation behavior of bats. In Bat algorithm, there is problem of trapping in local optima due to the searching behaviour. For enhancing the searching behavior, the chaotic variables have set of properties and these set of properties also helped to prevent the problem of trapping of Bat algorithm into local optima. The efficiency of getting new solutions was increased through randomization and also there was enhancement in the diversity of the solutions. Chances of finding global optimum solution were also increased. Authors applied this algorithm on Pima Indians Diabetes data set. The proposed model was compared with the traditional bat inspired algorithm and found the better result.

M Akhiljabbar et al. [15] proposed a combination of K-Nearest Neighbor and Genetic Algorithm for classification of Heart Disease. For pattern reorganization, the KNN is the simple, easily understood classification method in which the classification is done on the basis of class of their nearest neighbor. But to handle the large amount of data in the medical data set and for irrelevant and redundant data, authors proposed a model which was the combination of KNN and Genetic Algorithm. For global search in complex and large space and to find the optimal solution, genetic algorithm was used. The proposed model gave the better accuracy for heart disease prediction.

H. M. Harb and A. S. Desuky [16] proposed a method of feature Selection based on the Particle Swarm Optimization for the classification of medical data**.** In data mining, there are different steps to get the hidden information from the data. First, the data preprocessing is occurs if necessary. The classification accuracy of the data is affected when the data contains irrelevant and redundant data. Data preprocessing contains feature selection is one of the step. In this study, the authors used the wrapper and filter feature selection approaches with the particle swarm optimization. The features extracted from the above used approaches were applied to the different classifiers like KNN, naïve Bayesian, radial basis function, that improved the classification accuracy as compared to the other techniques like feature selection based on genetic algorithm. They applied these two approaches on three different data sets.

M. Ali Abed and Dr. H. Ali Abed [17] proposed a method which combined the SVM and CSOA for classification of electrocardiograms signals. Electrocardiograms signals (ECG) is the medical test that provide the diagnostic relevant information about the heart functioning. For classification, SVM has been widely used classification technique. In this paper, for parameter optimization of SVM, a novel metaheuristic technique named as cat swarm optimization was used. The proposed algorithm improved the SVM performance in two terms which were feature selection and the parameter optimization.

Sarab Al MuhaideandMohamed El BachirMenai [18] proposed a hybrid metaheuristic which consisted of two

steps. The first step was ant colony optimization and the second step was genetic algorithm. In the ACO step, by using the different subsets of training data decision lists were made-up. Decision lists were constructed by AntMiner+ algorithm. The population of genetic algorithm was initialized to these decision lists. The genetic algorithm worked for the optimization of these decision lists in term of classification accuracy and size. The proposed model gave the good results on the real world medical data sets.

Shilaskar ad Ghatol [19] gave a model which contained the feature selection and the classification techniques. Three data sets from UCI namely Arrhythmia, SPECTF and Heart Disease were used. In this study, The SVM classifier was integrated with forward feature inclusion, back-elimination feature selection and forward feature selection. The results illustrated that there were reduction in the numbers of inputs and also improvement in the accuracy by using the feature selection. For the SPECTF, accuracy of SVM increased 3% and number of features reduced from 44 to 19. For heart disease data set, the accuracy of SVM increased 4% and number of features reduced from 10 to 4.

JyotiSoni, Ujma Ansari et al. [20]. This paper gave the survey of latest techniques that were used in data mining especially in the field of medical research for diagnosis of heart disease. For comparative study, the authors have done the number of experiments and showed that the decision tree method was the best method and sometimes the naïve Bayesian gave the similar accuracy as that of decision tree. They showed that the other techniques like KNN, Neural network and classification based on clustering did not give the good results. They were also concluded that the performance of decision tree and the naïve Bayesian has been improved if they combined with the genetic algorithm. Genetic algorithm gave the reduced set of attribute for the heart disease which was optimal subset of attribute.

VII. CONCLUSION

This paper throws light on various hybrid techniques used for the diagnosis of diseases and illustrated the use of different data mining techniques which helps in the efficient decision making process. Several metaheuristics are available which combines with the various classification techniques. The presented paper showed the wide use of the soft computing techniques or swarm intelligence with the different classification techniques, which gives the better result as compared to the single classification techniques.

VIII. REFERENCES

[1] "Data Mining, 2nd Edition, Ian Witten, Eibe Frank.

[2] "Han and Kamber: Data Mining---Concepts and Techniques, 2nd ed., Morgan Kaufmann, 2006."

[3] A. Bhardwaj and A. Tiwari, "Breast cancer diagnosis using Genetically Optimized Neural Network model," *Expert Systems with Applications* , vol. 42, no. 10, pp. 4611–4620, 2015.

[4] N. C. Long, P. Meesad, and H. Unger, "A highly accurate firefly based algorithm for heart disease prediction," *Expert Systems with Applications*, vol. 42, no. 21, pp. 8221–8231, 2015.

[5] H. Inbarani, A. T. Azar and G. Jothi. "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis," Computer Methods and Programs in Biomedicine, pp. 175-185, 2014.

[6] S. Mishra, D. Mishra and K. Shaw, "A New Meta-heuristic Bat-Inspired Classification Approach for Microarray Data," *Procedia Technology*, vol. 4, pp. 802–806, 2012.

[7] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Systems with Applications*, vol. 36, no. 2, Part 2, pp. 3465–3469, 2009.

[8] R. Sallehuddin, S. H. Ubaidillah and N. H. Mustaffa, "Classification of Liver Cancer Using Artificial Neural Network and Support Vector Machine," *Elsevier Science Proc. Of Int. Conf on Advance in communicatiom Network, and Computing, CNC*, 2014.

[9] Y. E. Shao, C.-D. Hou, and C.-C. Chiu, "Hybrid intelligent modelling schemes for heart disease classification," *Applied Soft Computing*, vol. 14, Part A, pp. 47–52, 2014.

[10] P. Andreeva, M. Dimitrova, P. Radeva, "Data Mining Learning Models and Algorithms for Medical Applications,"

[11] R. M. Rahman and F. Afroz, "Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis," *Journal of Software Engineering and Application*, vol. 06, no. 03, pp. 85–97, 2013.

[12] M. Rouhani and K. Mansouri, "Comparison of Several ANN Architectures on the Thyroid Diseases Grades Diagnosis," in *International Association of Computer Science and Information Technology - Spring Conference, IACSITSC*, pp. 526–528, 2009

[13] D. E. K. and Mrs S. M Uma, "Intelligent Heart Diseases Prediction System Using "A New Hybrid Metaheuristic Algorithm," presented at the International Journal of Engineering Research and Technology, vol. 1 - Issue 8, 2012.

[14] O. S. Soliman and E Abo ElHamd " A Chaotic Levy Flights Bat Algorithm for Diagnosing Diabetes Mellitus," *International Journal of Computer Applications*, vol. 111, no. 1, pp. 36-42, 2015.

[15] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," *Procedia Technology*, vol. 10, pp. 85–94, 2013.

[16] H. M. Harb and A. S. Desuky, " Feature Selection on Classification of Medical Datasets based on Particle Swarm Optimization," *International Journal of Computer Applications*, vol. 104, no. 5, pp. 14–17, 2014.

[17] A. Prof. and A. Prof., "New Hybrid (SVMs-CSOA) Architecture for classifying Electrocardiograms Signals," *International Journal of Advanced Research in Artificial Intelligence*, vol. 4, no. 5, 2015.

[18] S. AlMuhaideb and M. E. B. Menai, "A new hybrid metaheuristic for medical data classification," *Int. J. Metaheuristics*, vol. 3, no. 1, pp. 59–80, 2014.

[19] S. Shilaskar and A. Ghatol " Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," Expert System with Applications, vol. 40, no. 10, pp. 4146–4153, 2013.

[20] J. Soni, U. Ansari, D. Sharma, and S. Soni, " Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, 2011.