ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Disease Prediction with Simulated Annealing

D. Asir Antony Gnana Singh[1], S. Praveena[2], B. Ramya[3], E. Jebamalar Leavline[4]

Teaching Fellow, Dept. of CSE, Anna University, BIT Campus, Tiruchirappalli, India. [1]

UG Student, Dept. of IT, Anna University, BIT Campus, Tiruchirappalli, India. [2,3]

Assistant Professor, Dept. of ECE, Anna University, BIT Campus, Tiruchirappalli, India. [4]

ABSTRACT: The paper proposes a medical diagnosis system using simulated annealing based variable selection. Recent analysis techniques in medical diagnosis use the different classification algorithms to detect the disease. The performance of classification algorithm is predicted by how accurately it predicts the distinct class on specific dataset. Hence it is very important to identify the features that contribute more in identifying the diseases and features with less contribution can be eliminated. The need of feature selection arises when we need to represent the massive medical data with reduced number of features. The objective of this paper is to design an effective algorithm that can remove irrelevant dimensions from large data and to predict more accurately the presence of the disease. The performance of the proposed method is tested an various real world medical datasets and also the performance of the proposed method is compared with existing methods in terms of classification accuracy of multilayer perceptron (MLP), lazy learner (IB1), random forest (RF) classifiers.

KEYWORDS: Disease prediction, Simulated Annealing, Feature selection, Machine learning

I. INTRODUCTION

In recent past, due to the developments in medical sector generate massive data such as radiologic results, medical test results, transcription of the physician's notes, laboratory results, and other sources. Medical equipment are extremely involved in the diagnosis, treatment, and monitoring of patients in healthcare. Classification involves finding rules that partition the data into disjoint groups. Classification analyses the training data set and constructs a model based on the class label and aims to assign a class label to the future unlabelled records. Classifying these massive data is a challenging task in developing medical diagnosis systems. Since the massive data contains irrelevant and redundant data, it reduces the classification accuracy of medical diagnosis system.

The process which involves the selection of subset of relevant features is called as the feature selection. It is also known as variable subset selection or variable selection. It is used for constructing the classification model due to the following reasons:

1. Simplification of models to relatively easier ones to understand by investigator.
2. Smaller teaching times
3. Enhanced generalization

Feature selection methods can be classified into three major categories based on the search technique and selection process namely complete, stochastic and heuristic search. There are three approaches for feature selection: Wrapper, Filter and Embedded. In wrapper approach, the selected subset of features is evaluated by a machine learning algorithm. Filter approach uses some techniques to score the selected subset, ignoring classifier algorithm. In embedded approach, selecting the best subset of features is performed during the process of training.

The drawback of the filter approach is that the process of selecting the best subset of features is independent of the classifier engine. It might cause a bad effect on the output of classification algorithms because the subset is just selected based on correlation between data. The wrapper and embedded approaches do not have such drawback because wrapper uses the same method for evaluating the selected subset of features that is used for classification and embedded approach performs feature selection during the process of training and it is not independent of the classifier algorithm.

The accuracy of the medical diagnosis system plays a vital role in healthcare. The misdiagnosis leads to wrong treatment that may lead to side-effects or death. Therefore, this paper aims to improve the accuracy of the medical diagnosis system by removing the irrelevant and redundant variables using the proposed simulated annealing based variable selection method. This proposed method is tested on the various real world medical datasets. The performance

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

of the proposed method is also tested using different classification algorithms such as function based classifier namely multilayer perceptron (MLP), instance based classifier namely (IB1), and tree based classifier namely Random Forest (RF) in terms of classification accuracy with 10-fold cross validation test mode.

II. RELATED WORK

Several approaches for implementing medical diagnosis system have been addressed in the literature. In Ref. [1], the authors introduced the fuzzy logic approach for medical diagnosis system. The results and outcomes from this study had exposed that fuzzy logic can lead to a reliable result in disease diagnosis [1]. Breast cancer can create major impact among women. At the same time, breast cancer is easily diagnosed early. An automatic diagnosis system using association rules (AR) and neural network (NN) for detecting the breast cancer was proposed in [2]. In this study the dimension of breast cancer data-base is reduced using AR and intelligent classification was achieved using NN. The correct classification rate of the system is 95.6%. This research established that AR and NN model can be used to increase the speed in the automatic diagnostic systems.

Chen et al [3] proposed a support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Anaraki et al performed a detailed review on rough set based feature selection methods [4]. Das et al [5] proposed a system of effective diagnosis of heart disease through neural networks ensembles. Increasing the accuracy of the medical diagnosis attributes through the main problem of the knowledgeable scheme that involves the cooperation of decision making classifications and schemes in the intelligence that predict the behaviors of disease symptoms and the doctors experience are represented by rules whilst the prediction of the possible diseases is identified by the prediction capability of medical expert systems. P. S. Jeetha Lakshmi et al proposed a novel hybrid medical diagnosis system based on genetic data adaptation decision tree and clustering. This paper used Naïve Bayes classifier for detection of heart disease. In this system, medical data is classified into five categories specifically no, low, normal, highland very high [6].

S. Dhanashree et al implemented heart disease prediction system using naive Bayes. The main aim of this paper is to build an Artificial Intelligent System that after analysis of certain parameters and to predict that whether a person is diabetic or not [7]. Hall [8] proposed correlation-based feature subset selection (CRFS) using example of feature subclass based method. Clara Madonna et al [9] proposed a rough-set approach to design a diabetic diagnosis system. In this paper time consumption can be reduced in predicting the disease. Rough set approach is used to design and develop the diagnosis system. Aishwarya et al [10] suggested a genetic algorithm based on medical expert system. To improve the accuracy Extreme Learning Machine based Genetic Algorithm (GA) are used for the selection of the most significant feature set of the dataset.

Peter Szolovits et al [11] proposed an artificial Intelligence in medical diagnosis. In this paper several problems in creation of actual artificial intelligence programs have been resolved. The paper introduced the feature subset selection using wrapper approach in supervised learning. It reduces computation time and set of features can be reduced. Three classifiers namely C4.5, Naïve Bayes, Bayesian networks are used for feature subset selection. Karegowda et al proposed a wrapper approach in feature subset selection [12]. The authors presented various approaches for feature selection and knowledge discovery process in the literature [13-17]. Sribarnasaha proposed a simultaneous feature selection using multi objective framework [18]. This paper focussed on recognizing the accurate partitioning in the relevant subset of features. S. Francisca Rosario et al [19] proposed a simulated annealing for feature selection. This paper used a slow cooling process. The main goal of this algorithm is to reduce the search space.

III. PROPOSED ALGORITHM

This section presents the proposed simulated annealing based variable selection algorithm.

A. *Proposed algorithm :*

Input: Dataset, cooling rate (CR), threshold (T)
Output: selected better attribute.
Step1: Begin
Step2: Read dataset

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Step3: Compute temperature for feature.
Step4: Perform scheduling.
Step5: Select the features which has highest temperature using threshold T, where T=3 Variables.
T_i=n(T_{i+1}= (n + i)), where n=3, i=1.
Step6: Find the solution using classifiers (MLP, RF, and IB1).
Step7: If the solution is optimal the selected features are better features.
Else Go to Step 8.
Step8: Reduced temperature of features by cooling rate (CR) where CR=0.01.
Step9: Perform Step 4 to step 7.
Step10: Stop.

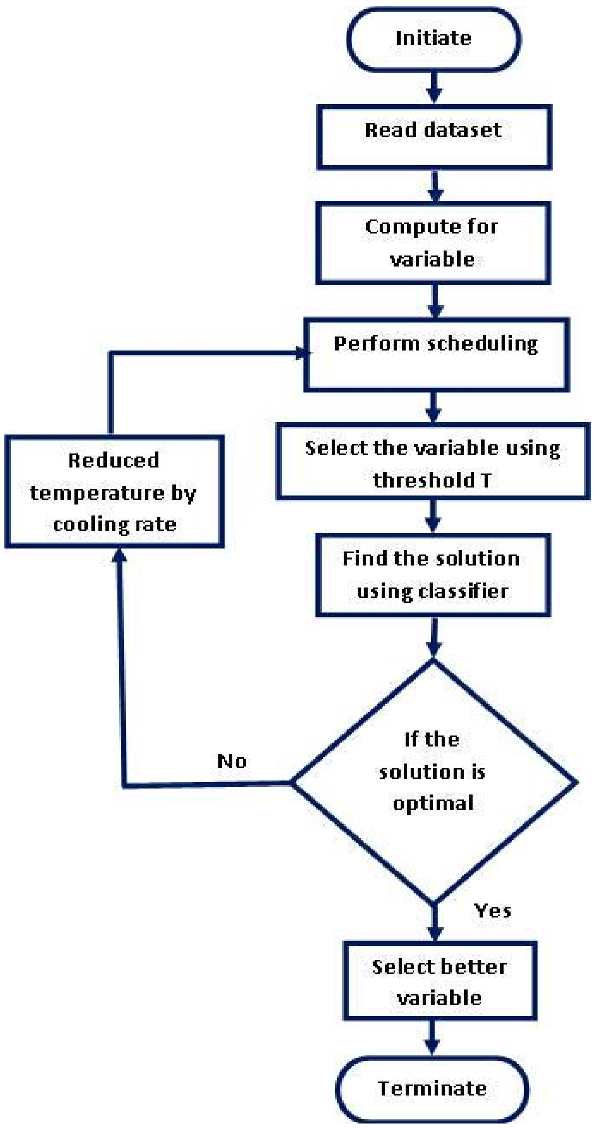

Fig. 1. Flowchart Representation of the proposed algorithm

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

B. *Algorithm Description:*

1. In the proposed work initially that dataset, cooling rate (CR), threshold (T) are given to the algorithm. Then the temperature of the features is calculated using information gain or Gain Ratio. Perform scheduling by ranking the features based on their temperature.
2. Select the features those who have highest temperature using threshold (T)
3. Find the solution for the selected attributes in terms of classification accuracy using MLP, RF, IB1.
4. If the solution is optimal and satisfied then the selected features are considered as selected better variables. Else reduce temperature of the feature using the cooling rate (CR).
5. Then repeat the step4 to step7.Until the optimal solution is reached.

C. *Architecture of the Proposed System:*

The schematic diagram of the proposed diagnostic system is illustrated in Figure 2. The collected real world medical datasets are given to the proposed variable selection algorithm using the simulated annealing (SA) based optimization methodology for selecting the significant variables. Then, classifiers are developed with the selected significant variables of the datasets using different classification algorithms such as function based classifier namely multilayer perceptron (MLP), instance based classifier namely IB1, tree based classifier namely random forest (RF). Then, the developed classifiers are tested on test datasets in terms of classification accuracy with 10-fold cross validation test mode. The performance of the proposed method is also compared with the existing methods. This algorithm is implemented using Java programming language with Net Beans IDE8.0.

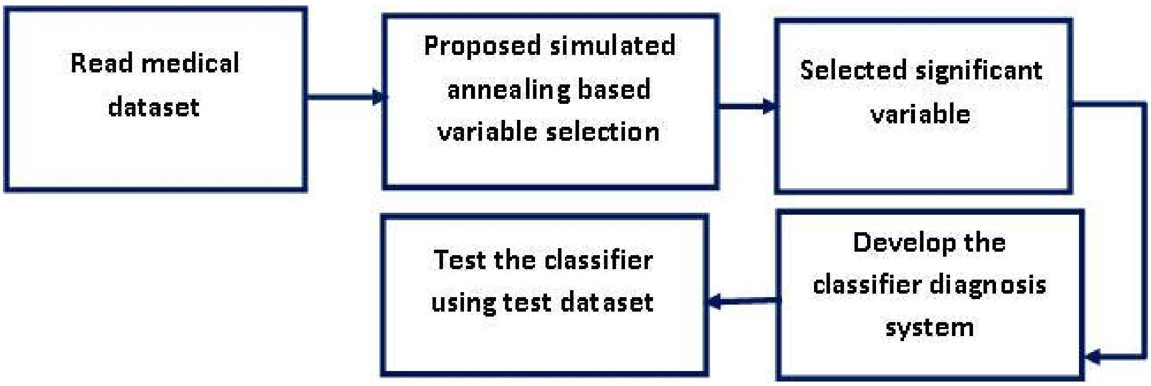

Fig. 2. Schematic diagram of the proposed project

IV. EXPERIMENTAL SETUP

In this experiment, the medical data related to diabetes is considered because the diabetes is one of the leading causes of death in women. The experiments are done using Weka tool. In this study Naïve Bayes algorithm is chosen to analyze the diabetes datasets because it provides better accuracy for medical datasets than the other two frequently used decision tree algorithms J48 and decision stump. Also, this algorithm is implemented using Java programming language with Net Beans IDE8.0. With an intension to find out whether the same feature selection method may lead to best accuracy for various datasets of same domain, various experiments are conducted on three different diabetes datasets.

The data is collected from UCI machine learning repository [www.ics.uci.edu] which is publicly available. The experiments are carried out on the datasets with and without feature selection methods and the results are compared and analyzed. The performance of the classifier is analyzed in terms of classification accuracy in percentage.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

V. RESULTS AND DISCUSSION

In this section, the experimental results are presented. Table 1 to Table 8 show the comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with different number of features selected based on information gain. Table 9 to Table 16 show the comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with different number of features selected based on gain ratio.

From the experimental results presented in Table 1 to Table 8, we observe that the classification accuracy of the disease diagnostic system is improve with the proposed feature selection algorithm compared to the classification accuracy without feature selection.

From the experimental results presented in Table 9 to Table 16, it is evident that the classification accuracy of the disease diagnostic system is improved with the proposed feature selection algorithm compared to the classification accuracy without feature selection.

Table 1: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 3 features selected based on information gain

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy (MLP)	Accuracy (RF)	Accuracy (IB1)	Selected feature	Accuracy (MLP)	Accuracy (RF)	Accuracy (IB1)
Diabetes	9	75.390	74.869	70.182	3	76.432	77.011	70.182
Breast-cancer	10	64.685	69.580	65.734	3	73.195	74.475	73.776
Breast-w	10	95.279	96.566	95.279	3	95.378	95.798	95.991
Heart-c	14	80.858	81.848	76.237	3	82.838	82.838	78.217
Hypothyroid	30	94.167	94.363	91.516	3	96.447	95.572	91.383
Heart-stat log	14	78.148	81.851	75.185	3	85.185	83.333	77.407

Table 2: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 4 features selected based on information gain

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	4	76.692	72.786	68.099
Breast-cancer	10	64.685	69.580	65.734	4	70.629	66.783	70.629
Breast-w	10	95.279	96.566	95.279	4	96.137	95.708	92.990
Heart-c	14	80.858	81.848	76.237	4	79.538	79.538	78.877
Hypothyroid	30	94.167	94.363	91.516	4	96.421	94.618	90.270
Heart-stat log	14	78.148	81.851	75.1852%	4	85.185	82.592	75.555

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Table 3: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 5 features selected based on information gain

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	5	75.260	71.614	69.401
Breast-cancer	10	64.685	69.580	65.734	5	71.678	70.279	65.384
Breast-w	10	95.279	96.566	95.279	5	95.565	95.708	93.133
Heart-c	14	80.858	81.848	76.237	5	80.858	81.848	76.237
Hypothyroid	30	94.167	94.363	91.516	5	96.261	97.454	94.220
Heart-stat log	14	78.148	81.851	75.185	5	82.592	79.259	74.814

Table 4: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 6 features selected based on information gain

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	6	73.958	73.697	68.099
Breast-cancer	10	64.685	69.580	65.734	6	67.832	67.482	62.937
Breast-w	10	95.279	96.566	95.279	6	95.708	95.851	94.134
Heart-c	14	80.858	81.848	76.237	6	79.207	79.868	79.538
Hypothyroid	30	94.167	94.363	91.516	6	79.207	79.868	79.538
Heart-stat log	14	78.148	81.851	75.185	6	78.888	80.000	79.629

Table 5: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 7 features selected based on information gain

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	7	75.781	74.609	67.057
Breast-cancer	10	64.685	69.580	65.734	7	67.482	69.230	66.783
Breast-w	10	95.279	96.566	95.279	7	95.851	96.137	95.422
Heart-c	14	80.858	81.848	76.237	7	81.188	80.858	77.227
Hypothyroid	30	94.167	94.363	91.516	7	96.421	99.257	93.160
Heart-stat log	14	78.148	81.851	75.185	7	81.481	79.629	78.888

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Table 6: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 8 features selected based on information gain

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	8	75.390	74.869	70.182
Breast-cancer	10	64.685	69.580	65.734	8	65.734	69.930	66.783
Breast-w	10	95.279	96.566	95.279	8	94.706	96.423	95.422
Heart-c	14	80.858	81.848	76.237	8	81.518	81.518	77.887
Hypothyroid	30	94.167	94.363	91.516	8	96.739	99.257	92.948
Heart-stat log	14	78.148	81.851	75.185	8	79.259	79.629	77.407

Table 7: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 9 features selected based on information gain

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Breast-cancer	10	64.685	69.580	65.734	9	64.685	69.580	65.734
Breast-w	10	95.279	96.566	95.279	9	95.279	96.566	95.279
Heart-c	14	80.858	81.848	76.237	9	83.498	81.848	77.227
Hypothyroid	30	94.167	94.363	91.516	9	96.315	99.204	92.497
Heart-stat log	14	78.148	81.851	75.185	9	80.740	81.851	81.111

Table 8: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 10 features selected based on information gain

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Heart-c	14	80.858	81.848	76.237	10	81.848	82.508	75.577
Hypothyroid	30	94.167	94.363	91.516	10	96.500	99.284	92.497
Heart-stat log	14	78.148	81.851	75.185	10	81.111	82.592	75.925

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Table 9: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 3 features selected based on gain ratio

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	3	76.432	72.265	70.182
Breast-cancer	10	64.685	69.580	65.734	3	75.174	74.475	67.832
Heart-c	14	80.858	81.848	76.237	3	82.508	82.838	71.947
Hypothyroid	30	94.167	94.363	91.516	3	96.447	95.572	91.383
Heart-stat log	14	78.148	81.851	75.185	3	85.185	83.333	77.407

Table 10: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 4 features selected based on gain ratio

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	4	76.302	72.656	70.182
Breast-cancer	10	64.685	69.580	65.734	4	75.174	72.727	65.035
Breast-w	10	95.279	96.566	95.279	4	96.280	95.994	95.565
Hypothyroid	30	94.167	94.363	91.516	4	96.421	94.618	90.270
Heart-stat log	14	78.148	81.851	75.185	4	83.333	78.888	79.629

Table 11: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 5 features selected based on gain ratio

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Breast-cancer	10	64.685	69.580	65.734	5	71.678	70.279	65.384
Breast-w	10	95.279	96.566	95.279	5	95.852	95.565	94.706
Hypothyroid	30	94.167	94.363	91.516	5	96.261	97.454	94.220
Heart-stat log	14	78.148	81.851	75.185	5	82.963	79.629	79.259

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Table 12: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 6 features selected based on gain ratio

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	6	75.781	74.609	67.057
Breast-cancer	10	64.685	69.580	65.734	6	67.832	67.482	62.937
Breast-w	10	95.279	96.566	95.279	6	95.708	95.851	94.134
Hypothyroid	30	94.167	94.363	91.516	6	97.242	99.098	94.591
Heart-stat log	14	78.148	81.851	75.185	6	78.888	80.000	79.629

Table 13: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 7 features selected based on gain ratio

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	7	75.520	73.697	67.708
Breast-cancer	10	64.685	69.580	65.734	7	67.482	69.230	66.783
Breast-w	10	95.279	96.566	95.279	7	95.422	96.137	94.420
Heart-c	14	80.858	81.848	76.237	7	81.188	80.858	77.227
Hypothyroid	30	94.167	94.363	91.516	7	97.295	99.072	94.565
Heart-stat log	14	78.148	81.851	75.185	7	81.481	79.629	78.888

Table 14: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 8 features selected based on gain ratio

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Diabetes	9	75.390	74.869	70.182	8	75.390	74.869	70.182
Breast-cancer	10	64.685	69.580	65.734	8	65.734	69.930	66.783
Breast-w	10	95.279	96.566	95.279	8	95.135	96.280	94.992
Heart-c	14	80.858	81.848	76.237	8	81.518	80.198	76.897
Hypothyroid	30	94.167	94.363	91.516	8	97.163	99.098	94.538
Heart-stat log	14	78.148	81.851	75.185	8	79.259	79.629	77.407

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Table 15: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 9 features selected based on gain ratio

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Breast-cancer	10	64.685	69.580	65.734	9	64.685	69.580	65.734
Breast-w	10	95.279	96.566	95.279	9	95.279	96.566	95.279
Heart-c	14	80.858	81.848	76.237	9	83.498	81.848	77.227
Hypothyroid	30	94.167	94.363	91.516	9	97.110	99.098	94.326
Heart-stat log	14	78.148	81.851	75.185	9	80.740	81.851	81.111

Table 16: Comparison of classification accuracy (in %) with and without the proposed feature selection algorithm on various datasets with 10 features selected based on gain ratio

Dataset	Without feature selection				With feature selection			
	Full feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3 (IB1)	Selected feature	Accuracy of class1 (MLP)	Accuracy of class2 (RF)	Accuracy of class3
Heart-c	14	80.858	81.848	76.237	10	81.848	82.508	75.577
Hypothyroid	30	94.167	94.363	91.516	10	97.295	99.337	94.485
Heart-stat log	14	78.148	81.851	75.185	10	81.111	82.592	75.925

VI. CONCLUSION

This paper analysed simulated annealing method for feature selection that are suggested by many investigates. Accuracy is most important in the field of medical diagnosis to diagnose the patient’s disease. We conclude that the classifier accuracy has been surely enhanced by the use of feature selection method than the classifier accuracy without feature selection. With an intension to find out whether the same feature selection method may lead to best accuracy for various datasets of same domain, various experiments are conducted on different datasets.

From the results it is clear that, the best feature selection method for a particular dataset depends on the number of attributes, attribute type and instances. Hence, whenever another dataset is considered, one has to experiment on that with various feature selection methods to identify the best one to enhance the classifier accuracy instead of simply considering the previously proved ones related to the same domain. Once the best feature selection method is identified for a particular dataset the same can be used to enhance the classifier accuracy.

REFERENCES

1. Adlassnig, Klaus-Peter. "Fuzzy set theory in medical diagnosis.", IEEE Transactions on Systems, Man and Cybernetics, 16, no. 2 (1986): 260-265.
2. Chou, Shieu-Ming, et al. "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines." Expert Systems with Applications 27.1 (2004): 133-142.
3. Chen, Hui-Ling, Bo Yang, Jie Liu, and Da-You Liu. "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis." Expert Systems with Applications 38, no. 7 (2011): 9014-9022.
4. Anaraki, Javad Rahimipour, and Mahdi Eftekhari. "Rough set based feature selection: a review." In 5th IEEE Conference on Information and Knowledge Technology (IKT), (2013), pp. 301-306.
5. Das, Resul, Ibrahim Turkoglu, and Abdulkadir Sengur. "Effective diagnosis of heart disease through neural networks ensembles." Expert systems with applications 36, no. 4 (2009): 7675-7680.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

6. P. S. Jeetha Lakshmi, S. Saravan Kumar, A. Suresh, “A novel hybrid medical diagnosis system based on genetic data adaptation decision tree and clustering”, ARPN Journal of Engineering and Applied Sciences, vol. 10, no. 16, pp. 7293 – 7299.
7. Medhekar, Dhanashree S., Mayur P. Bote, and Shruti D. Deshmukh. "Heart disease prediction system using naive bayes." Int. J. Enhanced Res. Sci. Technol. Eng 2, no. 3 (2013), pp. 1 – 5.
8. Hall, M.A., 1999. Correlation-based feature selection for machine learning (Doctoral dissertation, The University of Waikato).
9. LJ, Clara Madonna. "Design of a diabetic diagnosis system using rough sets." Cybernetics and Information Technologies 13, no. 3 (2013): 124-139.
10. Aishwarya, Anto, “A medical expert system based on genetic algorithm and extreme learning machine for diabetes disease diagnosis”, International Journal of Science, Engineering and Technology Research (IJSETR), vol 3, no 5, 2014, pp. 1375 – 1380.
11. Szolovits, Peter, Ramesh S. Patil, and William B. Schwartz. "Artificial intelligence in medical diagnosis." Annals of internal medicine 108, no. 1 (1988): 80-87.
12. Karegowda, Asha Gowda, M. A. Jayaram, and A. S. Manjunath. "Feature subset selection problem using wrapper approach in supervised learning." International journal of Computer applications 1, no. 7 (2010): 13-17.
13. D. Asir Antony Gnana Singh ,S.Appavu alias Balamurugan, E.Jebamalar Leavline, “An Unsupervised Feature Selection Algorithm with Feature Ranking for Maximizing Performance of the Classifiers”,International Journal of Automation and Computing , vol. 12, no. 5, pp 511-517, 2014
14. D. Asir Antony Gnana Singh, E. Jebamalar Leavline, E. Priyanka, C.Sumathi , “Feature Selection Using Rough Set For Improving the Performance of the Supervised Learner” International Journal of Advanced Science and Technology, vol. 87. pp.1-8
15. D.AsirAntonyGnanaSingh, S.AppavualiasBalamurugan, E.JebamalarLeavline, "Literature Review on Feature Selection Methods for High-Dimensional Data" International Journal of Computer Applications" Volume 136, Issue 1, February 2016, pp. 9-16 .
16. D.Asir Antony Gnana Singh ,S.Appavu alias Balamurugan, E.Jebamalar Leavline, “Improving the Accuracy of the Supervised Learners using Unsupervised based Variable Selection”, Asian Journal of Information Technology, vol. 30, No.9, pp. 530 – 537, 2014.
17. D.Asir Antony Gnana Singh, E.Jebamalar Leavline, "A Pragmatic Approach on Knowledge Discovery in Databases with WEKA" International Journal of Engineering Technology and Computer Research, Volume 2, Issue 7, 2014, pp. 81-87.
18. Sriparna Sahaa, Rachamadugu Spandanaa, Asif Ekbala, Sanghamitra Bandyopadhyayb, “Simultaneous feature selection and symmetry based clustering using multiobjective framework”, Applied Soft Computing, vol 29, 2015, pp.479–486.
19. S. Francisca Rosario, K.Thangadurai, “Simulated annealing algorithm for feature selection”, International Journal of Computers & Technology, vol. 15, no. 2, pp. 6471 – 6471.