IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-2, Feb 2019 E-ISSN: 2347-2693

Genetic Explorations for Feature Selection

A. Anushya

Department of Computer Science, A.D.M. College for Women (Autonomous), Nagapatnam,

DOI: https://doi.org/10.26438/ijcse/v7i2. 888892 | Available online at: www.ijcseonline.org

Accepted: 17/Feb/2019, Published: 28/Feb/2019

***Abstract*—** In this research, Genetic Algorithm is used for feature selection. Genetic Algorithm has been combined with local search, named as Memetic Algorithm and an algorithm is proposed and named as Compound Featuristic Genetic Algorithm. Next, based on, Class Dependent Feature Subset Selection, an algorithm is proposed namely, Core Featuristic Genetic Algorithm. The performance analyses of existing and proposed feature selection algorithms are functioned on heart dataset to predict the heart disease with minimum number of features. Finally, Fuzzy Decision Tree, Fuzzy Naive Bayes and Fuzzy Neural Networks are applied to the reduced set of the Heart dataset, obtained for classification accuracy.

***Keywords*—**Feature selection, Genetic Algorithm, Compound Featuristic Genetic Algorithm, Core Featuristic Genetic Algorithm, Fuzzy Decision Tree, Fuzzy Naive Bayes and Fuzzy Neural Networks

I. INTRODUCTION

Data mining is a process of extracting valid, previously unknown, comprehensible and actionable information from large databases and using it to make crucial business decisions. Feature selection reduces the number of features before applying data mining algorithm. Genetic Algorithm proposed by John Holland in 1975. It is capable of finding the best subset of features among the other features. It works with a set of candidate solutions called as population and optimal solution after a series of iterative computations. Each solution is called a chromosome. Each iteration is called a "generation". At each generation, two chromosomes are selected as parents for reproduction. The operations in GA are Initialization of parent population, Evaluation, Selection, Crossover and Mutation. The objectives of this paper are to identify key patterns or features from the dataset, To predict heart disease using data mining techniques, To identify and select attributes that are more relevant in relation to heart disease diagnosis and To compare Fuzzy Decision Tree, Fuzzy Naive Bayes and Fuzzy Neural Network Classifiers in predicting heart disease cases.

Rest of the paper is organized as follows, the interconnected former research is tabulated in section II. Section III discusses about feature selection using genetic algorithm, section IV deals about the feature section via class dependent and produces the results about reduced number of features by algorithms. Section V applied the algorithms on heart dataset and displays the results. Section VI expounds the fuzzy classifiers applied with feature selection algorithm on heart data and illustrates the accuracy and section VII concludes research work with future directions.

II. LITERATURE REVIEW

In this section, the previous related research on Feature Selection Using Genetic with Classification are selected and reviewed for this research. In table 1, the most relevant literatures are listed.

Table 1. Literature survey on Feature Selection Using Genetic with Classification

Author(s)	Description(s)
S. Senthamarai Kannan et.al (2010)	Proposed a novel hybrid feature selection algorithm, correlation based memetic algorithm (MA-C)
Malin Bj¨ornsdotter et.al (2010)	Proposed a genetic algorithm with a local search utilizing inherent called Memetic algorithm
K. Rajeswari et.al (2011)	Used Genetic algorithms for feature selection.
J. Anitha et.al (2010)	Applied fuzzy classifier with genetic algorithm
R.Bakyalakshmi et. al (2012)	Offered genetic algorithm with fuzzy data mining algorithm.
Gael de Lannoy et.al (2012)	Intended a method to perform class-specific feature selection in multiclass support vector machines.
Kashif Javed, et.al (2012)	Built a new feature ranking algorithm, termed as class-dependent density-based feature elimination.
Zhou Nina et.al (2012)	Suggested a class-dependent feature selection.

Asha Rajkumar et. al (2010)	Employed tanagra tool to classify the data with naive bayes algorithm, decision list algorithm and k-nn algorithm.
Srinivas, K et.al (2010)	Examined the potential use of classification based on rule based decision tree, naive bayes and artificial neural network to massive volume of healthcare data.
Anbarasi et.al (2010)	Exhibited decision tree were used to predict the diagnosis of heart patients with reduced number of attributes.
K.S.Kavitha et.al (2010)	Applied Hybridization to train the neural network using Genetic algorithm.
Bala Sundar.V et.al (2012)	Obtained high accuracy by the k-means clustering technique.
E.P.Ephzibah et.al (2012)	Proposed genetic algorithm, fuzzy rule based learning and neural networks.
S.Vijiyarani et al., (2013)	Analyzed the classification tree techniques in data mining.

III. FEATURE SELECTION BASED ON GENETIC ALGORITHM

Genetic Algorithm selects that attribute and carry out the mutation operation on it without any surveillance. The pseudo code of Genetic Algorithm is depicted in Figure 1. Also, iterations are limited for repeated calculation, but, when it is implemented, after number of iterations, optimum solution will be achieved. For the above example, Genetic Algorithm might be give attention to the other attributes, which are mostly near to the best solution. Therefore, Genetic Algorithm attempts the premature convergence and consuming much time for computation.

Begin;
Generate random population of P solutions (chromosomes);
For each individual i∈P: calculate fitness (i);
For i=1 to number of generations;
Randomly select an operation (crossover or mutation);
If crossover;
Select two parents at random ia and ib;
Generate on offspring ic = crossover (ia and ib);
Else If mutation;
Select one chromosome i at random;
Generate an offspring ic=mutate (i);
End if;
Calculate the fitness of the offspring ic;
If ic is better than the worst chromosome then replace the worst chromosome by ic;
Next i;
Check if termination=true;
End;

Figure 1. Genetic Algorithm

To overcome these problems, the algorithm is proposed and named as Compound Featuristic Genetic Algorithm. In Compound Featuristic Genetic Algorithm, every operation is same as Genetic Algorithm, but it varies in a petite manner, two parents are drawn from a fixed size population, they breed children, if the child's fitness is better than parent, then the worst parent is replaced by child. The pseudo code of Compound Featuristic Genetic Algorithm is given Figure 2.

Procedure Compound Featuristic Genetic Algorithm
Begin
Initialize population;
for each individual to local-search individual;
repeat
for individual = 1 to #crossovers do
select two parent individual1, individual2 in population randomly;
individual3:=crossover(individual1, individula2);
individual3 := local-search (individual3);
find smallest (child, individual3);
of those find parent with worst fitness;
calculate fitness (child);
if better fitness: exchange (child, individual3);
add individual i3 to population;
end for;
for individual=1 to #mutations do
select an individual of population randomly;
individual{m}:=mutate(individual);
individual {m} := local-search (individual{m});
find smallest F(child, individual{m});
of those find parent with worst fitness;
calculate fitness (child);
if better fitness: exchange (child, individual{m});
add individual {m} to population;
end for;
population :=select (population);
if population converged then
for each individual of best populations do individual := local-search (mutate(individual));
end if
until terminate = true;
End

Figure 2. Compound Featuristic Genetic Algorithm

IV. FEATURE SELECTION USING CLASS DEPENDENT FEATURE SUBSET SELECTION

Based on class dependent feature subset selection, an algorithm is proposed and called as Core Featuristic Gentic Algorithm. The steps involved in Core Featuristic Gentic Algorithm is explained in figure 3.

Step 1: Initially the feature space is clustered based on decision attributes.

Step 2: From each cluster the reduced feature subset is received as Ri, where i=1,2,…,ND.

Step 3: From these subsets the most common attributes are taken out as core subset (Rc) and these attributes are removed from each subset.

Step 4: Then the GA is again applied to select the random number of features from each cluster (Ri) and combine with the core Rc to find the optimum feature subset.

Figure 3. Compound Featuristic Genetic Algorithm

The Genetic Algorithm, Compound Featuristic Gentic Algorithm and Core Featuristic Gentic Algorithm were implemented in MATLAB and conducted experiments on medical data. The comparative analysis is shown in table 2 and performance analysis is illustrated in figure 4.

Table 2. Comparative Analysis

Data sets	**No. of Instances**	**No. of Attributes**				
Dermatology	362	33	29	27	24	
Ecoli	331	7	4	5	3	
Haberman	306	3	2	2	2	
Lung cancer	32	56	47	43	46	
Mammographic masses	830	5	4	4	2	
PIMA Indian diabetics	768	8	6	6	4	
Wine	178	13	9	7	8	
Wisconsin breast cancer	699	9	7	7	6	
Yeast	1484	8	6	6	5	

Compound Featuristic Genetic Algorithm fabricates minimum reduct from the data set holding large number of attributes with large data also. Core Featuristic Genetic Algorithm fabricates minimum reduct from the data set holding large instances.

V. APPLICATIONS OF FEATURE SELECTION TO HEART DATA

The Genetic Algorithm, Compound Featuristic Gentic Algorithm and Core Featuristic Gentic Algorithm were implemented in MATLAB and conducted experiments on heart data. The data are collected from the Cleveland Clinic Foundation, and it is available at the UCI machine learning Repository. A data frame with 297 observations of 13 conditional attributes and 1 decision attribute, which refers to the presence of heart disease in the patient. The decision attribute has the values 0 to 4, 0 denotes healthy and 1,2,3,4 denotes sick. The description of heart dataset is displayed in table 3.

Table 3. Heart Dataset Description

Attributes	**Description**
Age	Age in years
Sex	(1 = male; 0 = female)
Cp	Chest Pain Type (value 1: typical angina, value2: atypical angina, value3: non-angina pain, value 4: Asymptomatic)
Trestbps (mmhg)	Resting blood pressure
Chol(mg/dl)	Serum Cholesterol
Fbs	Fasting Blood Sugar (value 1: >120 mg/dl; value 0:<120 mg/dl)
Restecg	Resting electrographic results (value 0:normal; value 1: having ST-T wave Abnormality; value 2: showing probable or definite left ventricular hypertrophy)
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina (value 1: yes; value 0: no)
Old peak	ST depression induced by exercise relative to rest
Slope	Slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: down sloping)
CA	Number of major vessels colored by floursopy (value 0-3)

From the heart dataset, Genetic Algorithm selected 6 features, such as Cp - Chest Pain Type, Trestbps (mmhg)- Resting blood pressure, Exang - Exercise induced angina, Oldpk - Old peak, CA - No. of vessels colored by floursopy, Thal -Maximum heart rate achieved whereas Compound Featuristic Genetic Algorithm selected 4 features such as Cp - Chest Pain Type, Trestbps (mmhg)- Resting blood pressure, Exang - Exercise induced angina, and CA - Number of vessels colored by floursopy. Also, Core Featuristic Genetic Algorithm selected 5 features:Cp - Chest Pain Type, Trestbps (mmhg) -Resting blood pressure, Fbs - Fasting Blood Sugar, Exang - Exercise induced angina and CA - Number of vessels colored by floursopy. A Comparative Analysis of Genetic Algorithm with Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm on heart data is given in Table 4.

Table 4. Comparative Analysis on heart data

Algorithms	**Number of Reduced attributes**
Genetic Algorithm	6
Compound Featuristic Genetic Algorithm	4
Core Featuristic Genetic Algorithm	5

VI. CLASSIFICATION OF HEART DATA SET WITH FUZZY CLASSIFIERS

Classification is the process of learning a model that describes different classes of data. Learning the model is accomplished by using a training set of data that has already been classified. Then, through the learning, it can be forecast the unknown labeled data. Classification algorithms - Decision Tree, Naive Bayes and Neural Networks are used. A fuzzy classifier is any classifier which uses fuzzy sets either during its training or during its operation. Fuzzy Decision Tree, Fuzzy Naive Bayes and Fuzzy Neural Networks are applied with Genetic Algorithm, Compound Feauteristic Genetic Algorithm and core Featuristic Genetic Algorithm on Heart data. The result analysis is given in table 5.

Table 5. Comparative Analysis of Fuzzy Classifiers on Heart Data

Algorithms	**Accuracy (%)**		
	Fuzzy Decision Tree	**Fuzzy Decision Tree**	**Fuzzy Decision Tree**
Genetic Algorithm	81.14	81.14	81.14
Compound Featuristic Genetic Algorithm	86.52	86.52	86.52
Core Featuristic Genetic Algorithm	87.79	87.79	87.79

VII. CONCLUSION AND FUTURE SCOPE

The main conclusions of the study may be presented in a short Conclusion Section. In this section, the author(s) should also briefly discuss the limitations of the research and Future Scope for improvement.

To improve the efficiency of Genetic Algorithm, two algorithms are proposed and named as Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm. From the experimental analysis, it has been proved that the Compound Featuristic Genetic Algorithm fabricates minimum reduct from the data set holding large number of attributes with large data, whereas Core Featuristic Genetic Algorithm fabricates minimum reduct from the data set holding large instances. In heart data set, among 13 attributes, only 4 and 5 attributes have been chosen by Compound Featuristic Genetic Algorithm and Core Featuristic Genetic Algorithm correspondingly. Fuzzy Decision Tree, Fuzzy Naive Bayes and Fuzzy Neural Network have been utilized to predict the presence of heart disease. Fuzzy Neural Network with Core Featuristic Genetic Algorithm provides the best accuracy. In the future, this work can be expanded by exploring data mining techniques and also employed for other disease condition to predict at early stage with other dataset or images and algorithms to improve the classification accuracy and to build a model. Also, Missing values, noisy data, inconsistencies, and outliers presented a challenge in the data mining process.

REFERENCES

[1] A.Anushya, A. Pethalakshmi, D.Sheela Jeyarani, R.Raja Rajeswari, "A Comparative Study of Decision tree and Naive Bayesian classifiers on medical datasets", Proceedings of the International Conference on Computing and Information Technolgy, 2013.

[2] A.Anushya, A.Pethalakshmi, "A Comparative Study of Fuzzy Classifiers on Heart Data "Proceedings of the 3rd International Conference on Trendz in Information Sciences and Computing (TISC-2011), 978-1-4673-0131-2/11, IEEE Digital Library, 2011.

[3] A.Pethalakshmi, A. Anushya, "A comparative analysis of genetic based feature selection on heart data", International Journal of Computational Intelligence and Informatics, Vol. 2, No. 2, June – September, 2012.

[4] A.Pethalakshmi, A.Anushya, "Dynamic Feature Selection by Genetic on Medical Data", Sub-saharan Journal Computer Science , Vol. 1, No. 1, (ISSN: 2307-9169), 2013.

[5] A.Pethalakshmi, A.Anushya, "Effective feature selection via Featuristic genetic on heart data", International Journal of Computational Intelligence and Informatics, Vol. 2: No. 1, April – June, 2012.

[6] Anbarasi.M, E. Anupriya and N.CH.S.N.Iyengar, " Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm ," International Journal of Engineering Science and Technology Vol. 2(10), 2010, 5370-5376, 2010.

[7] Asha Rajkumar and Mrs. G.Sophia Reena, "Diagnosis Of Heart Disease Using Datamining Algorithm", GJCST, Vol. 10, Issue 10, pp: 38-43, 2010.

[8] Bala Sundar V, T DEVI, N SARAVANAN, " Development of a Data Clustering Algorithm for Predicting Heart", International Journal of Computer Applications (0975 – 888) International Journal of Computer Applications (0975 – 888), Vol 48, No.7, 2012.

[9] Dr. K. Usha Rani, "Analysis of Heart Diseases Dataset using Neural Network Approach", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, 2011.

[10] E.P.Ephzibah, V. Sundarapandian, "A Neuro Fuzzy Expert System for Heart Disease Diagnosis", Computer Science & Engineering: An International Journal (CSEIJ), Vol.2, No.1, 2012.

[11] G. Subbalakshmi, K. Ramesh, M. Chinna Rao, "Decision Support in Heart Disease Prediction System using Naive Bayes," Indian Journal of Computer Science and Engineering (IJCSE), Vol. 2, No. 2, pp: 170-176, 2011.

[12] J. Anitha, C. Kezi Selva Vijila, D. Jude Hemanth, "A hybrid Genetic Algorithm based Fuzzy Approach for Abnormal retinal Image Classification", International Journal of Cognitive Informatics and Natural Intelligence, Vol.4, No.3, pp: 29-43, 2010.

[13] Javed, K. Babri, H.A., Saeed, M, " Feature Selection Based on Class-Dependent Densities for High-Dimensional Binary Data", IEEE Transactions on Knowledge and Data Engineering, Vol: 24 , Issue: 3, pp: 465 -477, 2012.

[14] K. Rajeswari, V. Vaithiyanathan, P. Amirtharaj, " Prediction of Risk Score for Heart Disease in India Using Machine

Intelligence", International Conference on Information and Network Technology(IPCSIT), Vol.4, 2011.

[15] K.S.Kavitha, K.V.Ramakrishnan, Manoj Kumar Singh, "Modeling and design of evolutionary neural network for heart disease detection", International Journal of Computer Science Issues, Vol. 7, Issue 5, September, 2010.

[16] K.Srinivas, G.Raghavendra Rao, A.Govardhan, "Analysis of Attribute Association in Heart Disease Using Data Mining Techniques", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue4, 2012.

[17] K.Srinivas, G.Raghavendra Rao, A.Govardhan, "Analysis of Attribute Association in Heart Disease Using Data Mining Techniques", International Journal of Engineering Research and Applications (IJERA), Vol. 2, Issue4, 2012.

[18] Malin Bj¨ornsdotter & Johan Wessberg, " A Memetic algorithm for selection of 3D clustered features with applications in neuroscience", International Conference on Pattern Recognition, IEEE, 2010.

[19] P. Santhi, V. Murali Bhaskaran, "Improving the Performance of Data Mining Algorithms in Health Care Data", International Journal of Computer Science and Technology, IJCST Vol. 2, Issue 3, ISSN : 2229 – 4333 (Print), ISSN : 0976- 8491 (Online), 2011.

[20] R.Bakyalakshmi, Mr.N.Krishna Kumar , S.Karthika, M.Maheswari, " Minimizing Rules for Medical Dataset using Hybrid Fuzzy Classifier", International Journal of Communications and Engineering, Vol. 02, No.2, Issue: 01, March, 2012.

[21] Raj Kumar et.al, "Classifiction algorithms for Data Mining", A Survey,International Journl of Innovations in Engineering and Technology,Vol.1, Issue 2 ,2012.

[22] S. Senthamarai Kannan, N. Ramaraj, " A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm" Contents lists available at ScienceDirect Knowledge-Based Systems, Vol. 23, pp: 580–585, Elsevier, 2012.

[23] S.Vijiyarani , S.Sudha, "An Efficient Classification Tree Technique for Heart Disease Prediction", Intn mernational Conference on Research Trends in Computer Technologies (ICRTCT - 2013) Proceedings published in International Journal of Computer Applications® (IJCA) (0975 – 8887), 2013.

[24] Zhou Nina, Lipo Wang, "Class-Dependent Feature Selection for Face Recognition, Advances in Neuro-Information Processing", Lecture Notes in Computer Science Volume 5507, 551-558, 2009.

[25] Dipti.N.Punjani et.al, " A comprehensive study of various classification techniques in medical applications using data mining", International journal of Computer science and Engineering, vol.6, issue 6,June,2018.

Authors Profile

Dr. A.Anushya pursed Bachelor of Science from Mother Teresa University, India in 2006, Master of Computer Applications from Bharathidhasan University in year 2009 and Ph.D from Manonmaniam Sundaanar University, 2016. She is currently working as Assistant Professor in Department of Computer Science, A.D.M. College for Women (Autonomous), Nagapatnam. She has published more than 20 research papers in reputed international journals and conferences including IEEE and it's also available online. Her main research work focuses on Data Mining, Image mining, Soft computing and Computational Intelligence based education. She has 5 years of teaching experience and 7 years of Research Experience.

