E-ISSN: 2707-6644
P-ISSN: 2707-6636
IJCPDM 2020; 1(2): 07-12
Received: 05-05-2020
Accepted: 07-06-2020

GVS Sreeram Sarma
Dept. of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India

Corresponding Author:
GVS Sreeram Sarma
Dept. of Computer Science, Sri Venkateswara University, Tirupati, Andhra Pradesh, India

An intelligent hybrid feature selection using correlation coefficient and particle swarm optimization on microarray gene expression data

GVS Sreeram Sarma

DOI: https://doi.org/10.33545/27076636.2020.v1.i2a.10

Abstract
In this paper author is describing concept to apply combination of Particle Swarm Optimization algorithm and Correlation Coefficient algorithm for hybrid features selection to increase classifier accuracy and decrease system execution time. Some datasets such as GENES may contain attributes in thousands and classifying such huge attributes may degrade classifier accuracy and increase system execution time. To overcome from this issue author is using PSO and Correlation algorithm to select important attributes from dataset and ignoring unimportant attributes. These feature selection algorithms will prune unrelated attributes and select few attributes to perform classification. In this paper we are using 'Lymphoma' genes dataset which contains more than 4000 attributes but by applying PSO it will select only 52 important attributes out of 4000. Another dataset called 'SRBCT' contains more than 2000 attributes but PSO will choose few attributes from 2000. The main aim of this project is to select features from dataset by applying PSO features selection algorithm to reduce dataset size.

Keywords: intelligent hybrid feature, correlation coefficient, optimization on microarray

1. Introduction
Hybridization method is used to generate DNA samples in microarray gene expression data. This process can be done in two ways. In the first method, during the hybridization process, messenger RNA (mRNA) is stained using matrices sample taken from tissue or blood stream becomes cDNA. RNA profiling can be noisy and may not be sampled unevenly over time. The second method is the Affymetrix chips are hybridized using oligonucleotides on the surface of the array chip. It is possible to monitor and simultaneous measure thousands of activation levels of gene expression in a single experiment. This is considered as the key advantage of DNA microarray technology. Protein production helps identify the different types of memberships. This is achieved because the gene expression level refers to a specific protein production gene. The clinical medicine progress is only possible because of valuable results produced by microarray experiments performed on a variety of issues of gene expression profile. Microarray data can be applied to the problems of classification of cancer also. This was a recent development in the field of clinical research. Microarray data on cancer DNA is combined with statistical techniques for analysing gene expression profiles to identify potential biomarkers for diagnosis and treatment of various types of cancer.
Results obtained represent the state of the cell. Discriminant analysis of microarray data is an excellent tool for medical diagnostics of diseases, treatment and prevention. The main purpose of the classification is to build an effective model that can identify differentially expressed genes and could also be used to identify classes in the unknown samples. Some of the challenges in the microarray data are the smallest number of training and testing data available, the higher dimensionality of the data and the variations that could sneak in experiments performed to estimate the levels of gene expression. The two main tasks in the analysis of gene expression microarray are feature Selection and classification. To perform the classification process with an acceptable level of accuracy, the process of feature selection becomes crucial. Microarray gene expression data contains hundreds of thousands of genes or feature information. Only a small subset of genes exhibits strong correlation between them. Feature selection is a process that effectively selects differentially expressed genes in the dataset and forms a new subset for efficient classification. There may be situations in which a low-ranked gene could perform well in the rankings and a critical gene

could be left out in the selection of functions. The prediction accuracy would increase only with the best method of feature selection which otherwise would be impossible to understand. Another important measure is to avoid overfitting and build faster and cost-effective models. In this study, we use a hybrid approach that combines the benefits of a filter and a wrapper to perform feature selection. They are easy to use, simple and computationally efficient.

2. Literature Survey

2.1 P. Moradi and M. Gholampour, ‘‘A hybrid particle swarm optimization for feature subset selection by integrating a novel local search strategy,’’ Appl. Soft. Comput., vol. 43, pp. 117–130, Jun. 2016.

The proposed strategy utilizes a neighbourhood search procedure which is inserted in molecule swarm streamlining (PSO) to choose the diminished estimated and striking element subset. The objective of the nearby inquiry method is to control the PSO search procedure to choose particular highlights by utilizing their relationship data. In this manner, the proposed technique chooses the subset of highlights with decreased repetition. A half breed includes determination technique dependent on molecule swarm streamlining is proposed. Our strategy utilizes a novel neighbourhood search to improve the inquiry procedure close to worldwide optima. The technique productively finds the discriminative highlights with decreased correlations. The size of definite list of capabilities is resolved utilizing a subset size recognition scheme. Our technique is contrasted and notable and cutting edge include choice strategies. Highlight determination has been generally utilized in information mining and AI errands to make a model with few highlights which improves the classifier's precision. In this paper, a novel crossover includes choice calculation dependent on molecule swarm enhancement is proposed. The proposed technique called HPSO-LS utilizes a neighbourhood search system which is installed in the molecule swarm improvement to choose the less associated and striking element subset. The objective of the neighbourhood search method is to control the hunt procedure of the molecule swarm enhancement to choose unmistakable highlights by thinking about their relationship data. Additionally, the proposed technique uses a subset size assurance plan to choose a subset of highlights with decreased size. The presentation of the proposed technique has been assessed on 13 benchmark order issues and contrasted and five best in class include choice strategies. Also, HPSO-LS has been contrasted and four notable channel-based strategies including data gain, term fluctuation, fisher score and mRMR and five notable wrapper-based techniques including hereditary calculation, molecule swarm advancement, re-enacted strengthening and insect settlement enhancement. The outcomes exhibited that the proposed technique improves the arrangement exactness contrasted and those of the channel based and wrapper-based element choice strategies. Besides, a few performed measurable tests show that the proposed strategy's prevalence over different strategies is factually noteworthy.

2.2 S. Saremi, S. Mirjalili, and A. Lewis, "Grasshopper advancement calculation: Theory and application," Adv. Eng. Softw., vol. 105, pp. 30–47, Mar. 2017

The way toward finding the best qualities for the factors of a specific issue to limit or amplify a target work is called streamlining. Enhancement issues exist in various fields of studies. To take care of an enhancement issue, various advances should be taken. Right off the bat, the parameters of the issue ought to be recognized. In view of the idea of the parameters, issues might be named consistent or discrete. Also, the imperatives that are applied to the parameters must be perceived [1]. Imperatives separate the enhancement issues into compelled and unconstrained. Thirdly, the targets of the given issue ought to be researched and thought of. For this situation, advancement issues are characterized into single-objective versus multi-target issues [2]. At long last, in light of the recognized kinds of parameters, limitations, and number of targets a reasonable optimiser ought to be picked and utilized to take care of the issue. Numerical improvement mostly depends on inclination-based data of the included capacities so as to locate the ideal arrangement. Albeit such methods are as yet being utilized by various analysts, they have a few weaknesses. Scientific streamlining approaches experience the ill effects of neighbourhood optima entanglement. This alludes to a calculation expecting a nearby arrangement is the worldwide arrangement, subsequently neglecting to acquire the worldwide ideal. They are likewise regularly incapable for issues with obscure or computationally costly determination [3]. Another kind of improvement calculation that mitigates these two downsides is stochastic streamlining [4].

3. Proposed Work

The minimum support and minimum confidence of traditional association rule mining algorithms are set by users themselves. Although some of them refer to the opinions of relevant experts, they are not objective enough and have no theoretical support. If the value is smaller or bigger, it will affect the results. In order to solve the problem of low efficiency of association algorithm caused by frequent dataset replacement, proposed a genetic algorithm combined with immune optimization is proposed to mine association rules. The results show that the optimized algorithm can build redundant rules and improve efficiency. introduce an algorithm based on fuzzy Formal concept analysis and prime number coding. By constructing a frequent fuzzy minimum generator mesh, it can effectively solve the problem that the extraction fuzzy association rules run for a long time.

1. Pre-processing: using this module we will remove out empty values from dataset and convert non digits attribute such as ‘?’ to 0 digits.
2. Generate Model: using this module will convert dataset into train model
3. Features selection: using this module we will apply PSO and Correlation algorithm to select features.
4. Run algorithm module: using this module we will run various algorithms such as J48 tree, random forest and propose EML (Extreme Machine Learning) algorithms
5. Graphs: using this will compare accuracy of all algorithms in graphs

4. Results and Discussions

Fig 1: In above screen click on 'Upload Micro Array Dataset' button to upload genes dataset

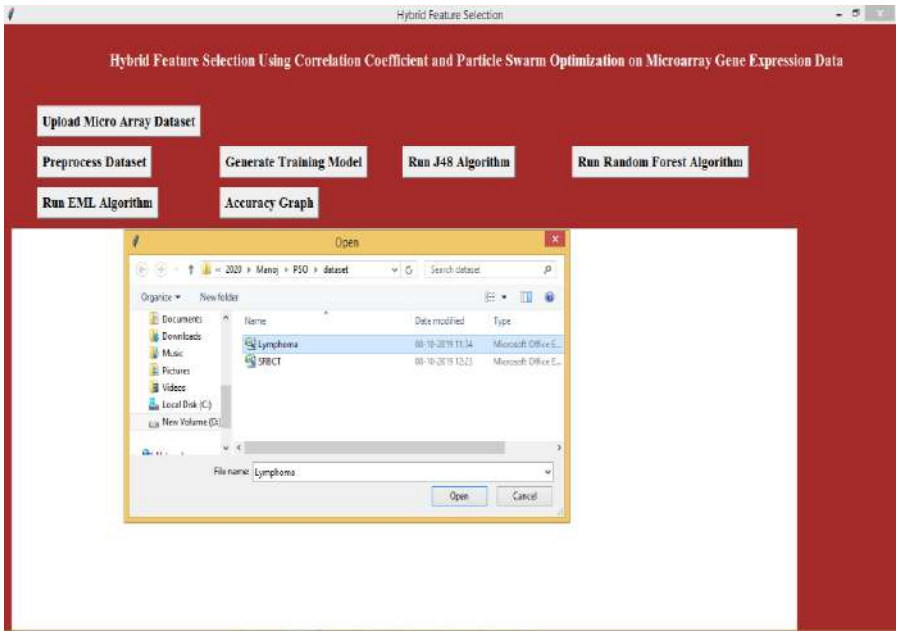

Fig 2: In above screen we are uploading Lymphoma dataset after upload dataset will get below screen

Fig 3: Now click on 'Pre-process Dataset' to remove empty strings

Fig 4: In above screen we are seeing come values from the dataset now click on 'Generate Training Model' to create train and test dataset.

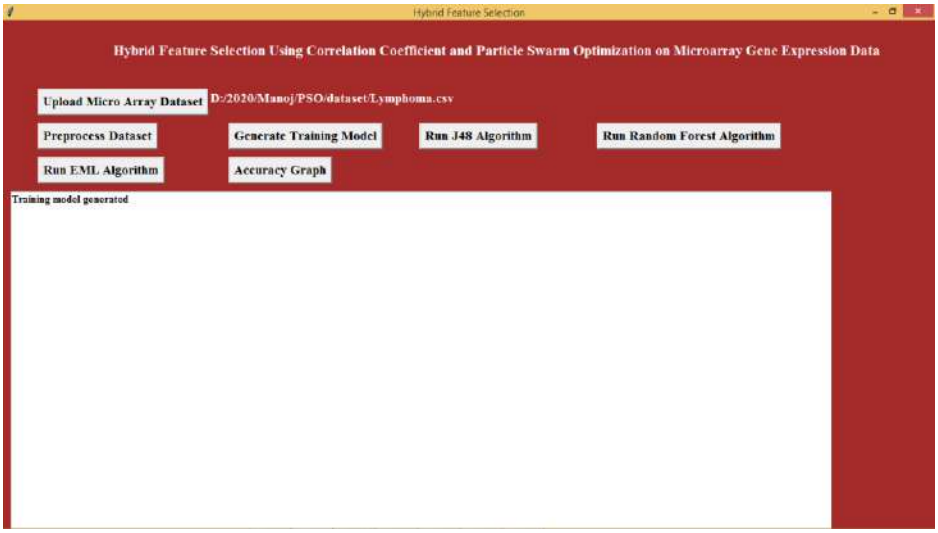

Fig 5: After training model click on 'Run J48 Algorithm' to run j48 decision tree classifier

Fig 6: In above screen for J48 we got 42% accuracy now run random forest algorithm

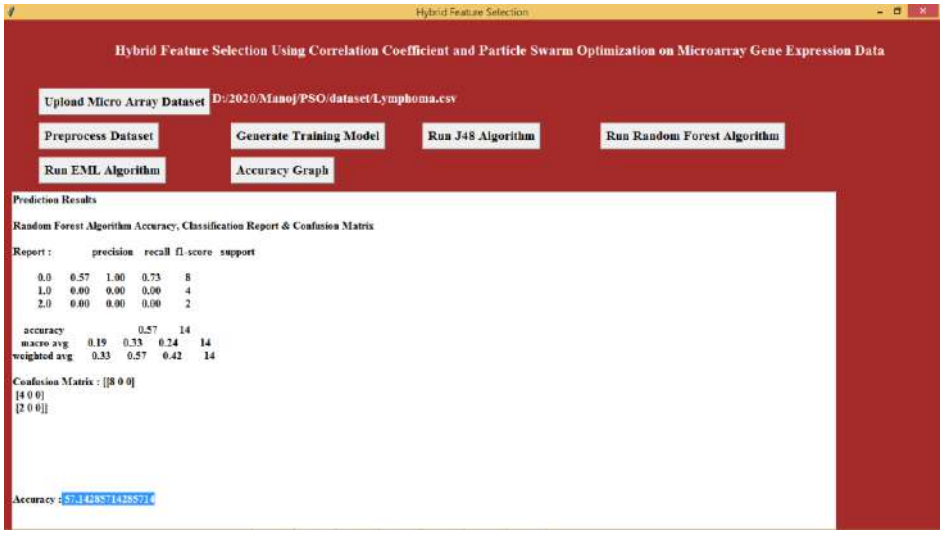

Fig 7: In above screen for random forest accuracy is 57% now click on Run EML Algorithm

Fig 8: In above screen selected text we can see total 4027 attributes are there but after applying PSO we got 52 and in below lines we can see accuracy is 64% for ELM which is higher than other algorithms

Fig 9: In above graph we can see accuracy for J48, random forest and EML algorithms. In above graph x-axis represents algorithm name and y-axis represents accuracy

5. Conclusion

In this paper, we have discussed the classifier accuracy of the proposed hybrid approach that combines the correlation coefficient with particle swarm optimization. This is compared with the traditional tree-based classifiers like J48, Random Forest, Random Tree, Decision Stump and Genetic Algorithm as well. It is evident that the extreme learning machines classifier produces more or comparatively better accuracy than the other tree-based classifiers available in literature. The proposed hybrid method that has higher potential in aiding further research in the area of feature selection simplified the process of gene selection which is evident from the experimental results. The proposed method significantly reduces the number of genes needed for

classification and has also contributed to the improvement in classifier accuracy. The proposed method has greater scope of application to problems in other domains in future.

6. References

1. Sushmita Mitra, Ranajit Das, Yoichi Hayashi. "Genetic Networks and Soft Computing", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011, 8(1).
2. Cheng-San Yang, Li-Yeh Chuang, Chao-Hsuan Ke, Cheng-Hong Yang. "A Hybrid Feature Selection Method for Microarray Classification", IAENG International Journal of Computer Science, 21 August, 2008
3. Cheng-Huei Yang, Li-Yeh Chuang, Cheng-Hong Yang. "IG-GA: A Hybrid Filter/Wrapper Method for Feature Selection of Microarray Data", Journal of Medical and Biological Engineering, 30(1), 23-28
4. Pradipta Maji, Chandra Das. "Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification", IEEE Transactions on Nano Bioscience, 2012, 11(2).
5. Li-Yeh Chuang, Hsueh-Wei Chang, Chung-Jui Tu, Cheng-Hong Yang. "Improved binary PSO for feature selection using gene expression data", Computational Biology and Chemistry. 2008; 32(1):29-38.
6. Alok Sharma, Seiya Imoto, Satoru Miyano. "A Top-r Feature Selection Algorithm for Microarray Gene Expression Data", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2012, 9(3).
7. Argiris Sakellariou, Despina Sanoudou, George Spyrou. "Investigating the Minimum Required Number of Genes for the Classification of Neuromuscular Disease Microarray Data", IEEE Transactions on Information Technology in Biomedicine. 2011, 15(3).
8. Jagath Rajapakse C, Piyushkumar Mundra A. "Multiclass Gene Selection Using Pareto-Fronts", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013, 10(1).
9. Jialei Wang, Peilin Zhao, Steven Hoi CH, Rong Jin. "Online Feature Selection and Its Applications", IEEE Transactions on Knowledge and Data Engineering, 2014, 26(3)
10. Qinbao Song, Jingjie Ni, Guangtao Wang. "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, 2013, 25(1).
11. Sheng Liu, Ronak Patel Y, Pankaj Daga R, Haining Liu, Gang Fu, Robert J *et al*. "Combined Rule Extraction and Feature Elimination in Supervised Classification", IEEE Transactions on Nano Bioscience, 2012, 11(3)
12. Yukyee Leung, Yeungsam Hung. "A Multiple-Filter-Multiple-Wrapper Approach to Gene Selection and Microarray Data Classification", IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2010, 7(1)
13. Guoli Ji, Zijiang Yang, Wenjie You. "PLS-Based Gene Selection and Identification of Tumor-Specific Genes", IEEE Transactions on Systems, Man, and Cybernetics – Part C: Applications and Reviews, 2011, 41(6).