International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 05, May-2015

# A Hybrid Genetic Algorithm to Improve Feature Selection

Suruchi Koul
Student, Dept. CSE/IT, ITM University
Gurgaon, India

Rita Chhikara
Faculty, Dept. CSE/IT, ITM University
Gurgaon, India

***Abstract*— Feature selection techniques have turned into a clear need in numerous applications. Feature Selection removes redundant or irrelevant features and can improve the accuracy of classifiers. Notwithstanding the substantial pool of methods that have as of now been created in the machine learning and data mining fields, no single criterion isbest for all applications.This paper proposes a framework to obtain better results using Relief and Correlation Feature selection which are filter methodsfor generation of feature pool for hybrid GA which is a wrapper method using Naïve Bayesian classifier. Thus, we obtain combined advantages of both filter and wrappermethods to get better subset of features and not compromising with the classification accuracy, which is the current need of the hour.**

***Keywords—Filter,Wrapper, Genetic algorithm, fitness function, feature pool,Relief, Correlation Feature Selection ,Naïve Bayes.***

## I. INTRODUCTION

Amid the most recent period, the rise of huge measures of multivariate information in different applications has contributed to the need of Feature Selection.Feature selection strategies consider over how to distinguish and select useful features for building models which can translate information better. Feature selection can decrease the computational cost by lessening dimensionality of information; enhance the projection execution and the intelligibility of the models by classification of repetitive and insignificantfeatures.

Feature Selection has its applications in various fields such as in Geneticsit is used to find out the solution to some vastly occurring diseases such as diabetes, cancer, arthritis and many more by rightly classifying the features. The bioinformatics field is just not only concerned with the human, but it is also a subject for animal's features and various characteristics of flora and fauna. But due to higher dimensionality of data, the task gets difficult and the need for Feature Selection technique comes in picture.

The aim of feature selection is typical, the most vital ones presenting: (a) to abstain from over fitting and enhance model execution, i.e. forecast execution on account of managed classification and improved cluster location on account of clustering, (b) to give quicker and new savvy models and (c) to pick up a more profound understanding into the hidden techniques that created the information [1]. Notwithstanding, the focal points of Feature selection strategies come at a positive cost, as the search for a subset of significant features presents an extra layer of unpredictability in the displaying assignment [2]. Rather than simply improving the constraints of the model for the complete Feature subset, we now requisite to discover the ideal model constraints for the ideal Feature subset, as there is no ensure that the ideal constraints for the complete feature set are just as ideal for the ideal feature subset [4]. Accordingly, the pursuit in the model speculation space is expanded by another measurement: the one of discovering the ideal subset of important features. Various methodologieshavebeen studied for feature selection.

However, feature selection can further be categorized into 3 categories.

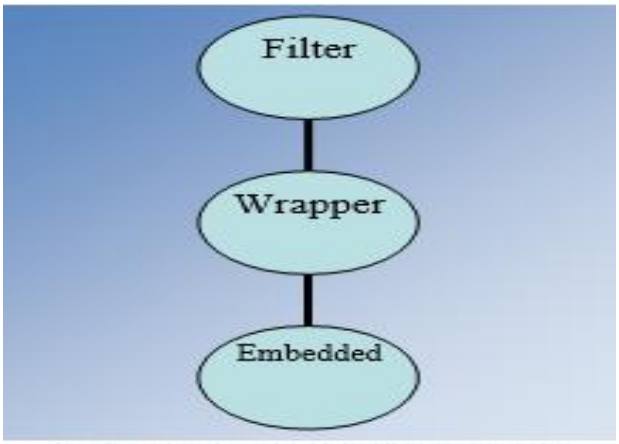


**Fig. 1 Classification of feature selection**

This paper proposes to improve Feature selection by using combined advantages of Relief, correlation Feature selection which are filter methods and GA, which is a wrapper approach. The proposed work is performed in two steps. Firstly, we construct a feature pool for our GA using these filter methods. Then this resultant data subset is given to GA which works with the Naïve Bayes classifier, a variant of supervised learning. Therefore, we are able to work with hybrid GA having combined advantages of both filter and wrapper approach and thus selecting the optimal features with better classification accuracy and time efficiency.

### A. *The Naive Bayes*

Naive Bayes classifier is taking into account Bayes hypothesis. This classifier calculation utilizes restrictive autonomy, implies it expect that a quality esteem on a given class is free of the estimations of other traits[4].

The Bayes hypothesis is as per the following: Let X={x1, x2... xn} be an arrangement of n qualities. We need to focus P (H|X), the likelihood that the theory H embraces given confirmation i.e. information test X.

As per Bayes hypothesis the P (H|X) is communicated as:

$$P(H|X) = P(X|H)\,P(H)/\,P(X) \quad (1)$$

### B. *Gentic Algorithm*

Genetic Algorithms (GAs) are versatile heuristic hunt calculation in view of the transformative thoughts of characteristic determination and hereditary qualities. All things considered they articulate to a keen exploitation of an arbitrary hunt used to tackle streamlining issues. Albeit randomized, GAs are in no way, shape or form irregular, rather they misuse authentic data to through the pursuit into the area of improved execution inside the hunt space. The fundamental procedures of the GAs are intended to recreate forms in common frameworks vital for advancement, exceptionally those take after the standards first set around Charles Darwin of "survival of the fittest". Since in nature, competitionamid people for sparse assets brings about the fittest people commanding over the weaker ones [5] [6].

The GA works with three operators Fig.2.

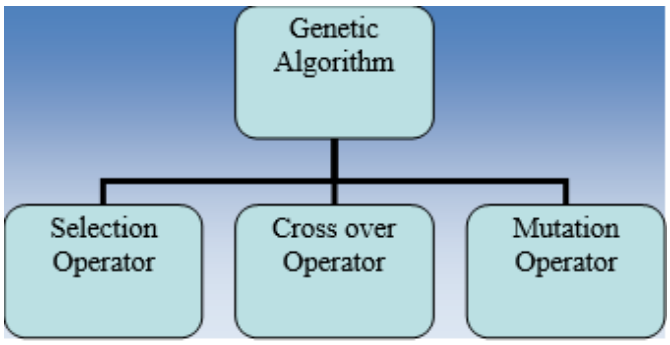


**Fig. 2 Genetic Algorithm operators**

The capacity of the algorithm to investigate and analyze at the same time, a developing measure of hypothetical avocation, and effective application to day to day issues reinforces the decision that GAs are a capable, vigorous enhancement system.

### C. *RELIEF Algorithm*

This algorithm "estimates the relevance of features according to how well their values distinguish among the instances of the same and different classes that are near each other"[7]. These types of filters are suited to high-dimensional datasets, in terms of accuracy, time, and memory efficiency.

### D. *Correlation Feature Selection*

It is a correlation based filter method. CFS evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them[8]. The subsets of features that are having high correlation with the class while having low inter-correlation are preferred.

## II. METHODOLOGY

### A. *Feature Group*

The feature group[5] or pool is a gathering of applicant features to be chosen by the genetic calculation to build a feature subset. As opposed to utilizing all features, we take arrays of features chosen by a few feature choice calculations to structure the group. In the proposed methodology, dataset is given to Feature selection algorithms namely Relief and correlation Feature selection which are filter methods and the resultant data subset consisting of top 20 features selected from these algorithms act as feature pool for our GA.

### B. *Fitness Function*

The key objective of this research work is to get maximumcataloging accuracy of the feature subset and reduce the size of the feature subset. To do so, the resulting fitness function[5] is called upon:

$$F = w*c(\,)\,x + (1-w)*(1/s(x)) \quad (2)$$

Where, x is a feature vector demonstrating a feature subset certain and w is a constraint between 0 and 1. The function is collected of 2 parts. The $I^{st}$ part is a subjective classification accuracy c(x) from a classifier and the $II^{nd}$ part is subjective size s(x) of the feature subset signified by x. For a given w, the fitness of adiscrete x is amplified.

### C. *Induction Algorithm*

The genetic algorithm is autonomous of the genetic learning algorithm used by a Naive Bayes classifier. Multiple algorithms, such as Naïve Bayes, decision trees and artificial neural network, can be adaptablycombined into the methodused by [5]. Here the implementation is done over the Naïve Bayes classifier inNETBEANS software.

*D. Genetic Operator*

- Selection

Roulette wheel selection is used to probabilistically select entities from a population for future breeding. The probability of selecting specificentity $h_i$ is resolute by[5]:

$$P(h_i) = \frac{F(h_i)}{\sum_{i=1}^{p} F(h_i)}$$

The probability that an specificentity will be nominated is proportionate to its own fitness and is inversely proportionate to the fitness of the other challengingtheories in the current population.

- Crossover

Theiterative crossover operator[5] is used. The crossover facti is selected at arbitrary so that the first i bits are donated by one parent and the continuing bits by the second parent.

- *Mutation*

Each entity has a possibility pm to mutate. A randomly chosen number of n bitsare to be overturned in every mutation stage[5].

## III. WORKFLOW

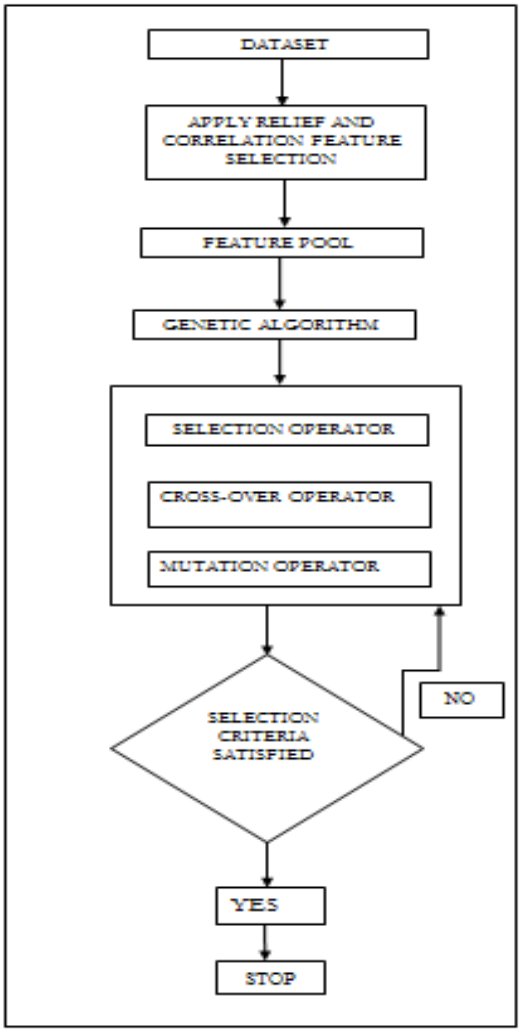


Fig. 3. showing the proposed workflow with the following steps:

- Step1: Four datasets are used as input to Feature Selection Algorithm.
- Step2:Relief and Correlation Feature selection both filter methods are applied to the respective datasets to extract the features for creating feature pool for GA.
- Step3: Once the features are extracted, then genetic algorithm is applied that works according to the natural selection of population, with the three operators and their work is mentioned in the above section.
- Step4: Now the final decision depends on the value of optimal solution of the testing as per the equation --1. If the output is optimal and meets the criteria then the process will stop and if not then again GA will implement it.
- This way we were able to combine the benefits of filter and wrapper methods into one making it hybridGA which is providing us optimal results with higher classification accuracy.

## IV. IMPLEMENTATION

This proposal has been implemented in NetBeans Java. During implementation, the graphical user interface (GUI) has been created with 4 tabs, for the GA setting, processing, logs and performance plot.

The number of class is entered in the processing tab, this will divide the subsets and the iterations i.e. number of times the execution will occur is entered by the GA tab Fig.4. for the subsets.

The matching classes or extracted features are presented in the logs.

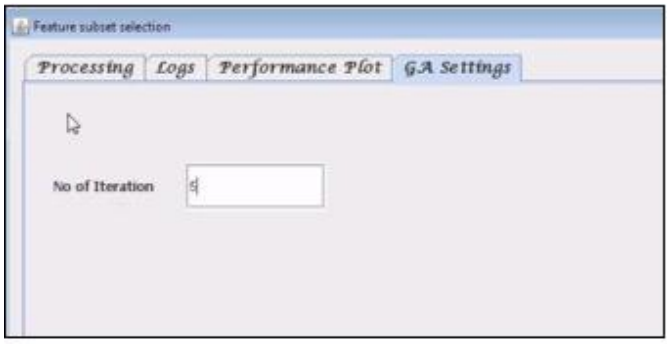


Fig. 4 The feature subset selection

After iteration the best results are sent to the pool for global extraction. The final best result will be considered whose value is highest in the pool.

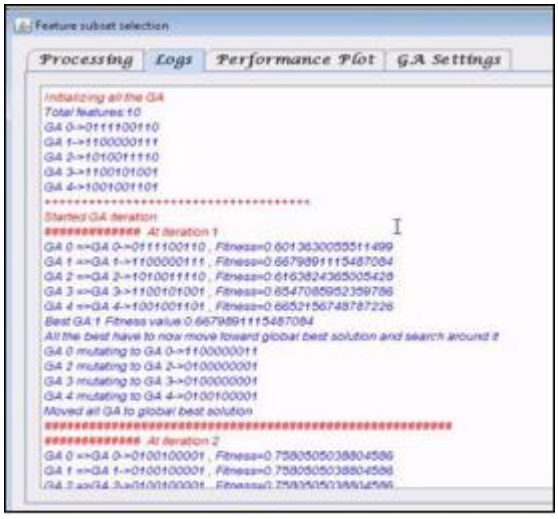


Fig. 5 GA output with the optimal solution

## V. EXPERIMENTAL RESULTS AND ANALYSIS

TABLE I Number of Features and instances in datasets used

| DATASET | No. of Features | No. of Instances |
|---|---|---|
| SPECTF-Heart | 23 | 187 |
| Soybean-small | 35 | 47 |
| Steganalysis | 275 | 4001 |
| Internet-ad | 1559 | 2168 |

Dataset SPECTF-Heart consists of 23 features and 187 instances. Dataset soybean-small contains 35 features and 47 instances. Dataset Steganalysis contains 275 features with 4001 instances. High dimensional dataset Internet-ad consists of 1559 features and 2168 instances. All datasets have been taken from UCI Machine Learning repository.

TABLE II Number of features reduced after Feature Selection

| DATASET | Without Feature Selection | With Feature Selection | |
|---|---|---|---|
| | | *Relief* | *Correlation Feature Selection* |
| SPECTF-Heart | 23 | 4 | 4 |
| Soybean-small | 35 | 9 | 10 |
| Steganalysis | 275 | 131 | 135 |
| Internet-ad | 1559 | 26 | 45 |

It can be observed from table II that Dataset SPECTF-Heart containing **23** features was reduced to **4** features after using Relief and Correlation Feature Selection, similarly, for Dataset Soybean-small which contains **35** were reduced to **9**, for Dataset Steganalysis containing **275** features which were reduced to **131**and Dataset Internet-Ad which had **1559** features were reduced to **26**.This shows how the irrelevant or redundant features are eliminated by using Feature Selection Methods. At this point we need to debate that the reduction in number of features should not be at the cost of accuracy.

TABLE III Comparison of Accuracy and Time using Relief and Correlation Feature selection methods

| | Without Feature Selection | | With Feature Selection | | | |
|---|---|---|---|---|---|---|
| | | | *Correlation Feature Selection* | | *Relief* | |
| **DATASET** | ***Accuracy*** | ***Time*** | ***Accuracy*** | ***Time*** | ***Accuracy*** | ***Time*** |
| SPECTF-Heart | 68.40% | 0.1s | 72.72% | 0.1s | 72.72% | 0.1s |
| Soybean-small | 97.87% | 0.01 s | 99% | 0.1s | 99% | 0.1s |
| Steganalysis | 99.13% | 0.2s | 99.40% | 0.12 s | 99.47% | 0.12s |
| Internet-ad | 95.10% | 0.71 s | 95.01% | 0.03 s | 95.01% | 0.01s |

It can be observed from table III that accuracy has been increased even with reduced feature set for all the datasets. Consider dataset SPECTF-Heart where accuracy was 68.40% without feature selection, after performing feature selection it is increased to 72.72%.

TABLE IV Top 20 features selected by Relief Algorithm

| **DATASET** | **Top 20 Features selected by Relief** |
|---|---|
| SPECTF-Heart | a1,a14,a17,a22 |
| Soybean-small | a4,a12,a21,a22,a23,a24,a26,a28 |
| Steganalysis | a113,a39,a115,a50,a117,a112,a226,a21,a250,a13, a185,a179,a266,a10,a187,a251,a114,a198,a233,a191 |
| Internet-ad | a3,a1400,a352,a1244,a969,a533,a1230,a1023,a277,a399, a1484,a62,a1279,a376,a1088,a91,a875,a171,a5,a82 |

The top 20 features selected by Relief Algorithm for all the datasets are as shown in table IV. We will select top 20 features ranked by Relief for the construction of feature pool for our GA.

TABLE V Top 20 features selected by Correlation Feature Selection

| **DATASET** | **Top 20 Features selected by Correlation Feature selection** |
|---|---|
| SPECTF-Heart | a1,a14,a17,a22 |
| Soybean-small | a4,a12,a21,a22,a23,a24,a26,a28,a35 |
| Steganalysis | a2,a10,a11,a13,a15,a21,a23,a30,a31,a35, a39,a43,a46,a50,a58,a59,a67,a68,a70,a72 |
| Internet-ad | a3,a5,a20,a62,a82,a91,a171,a184,a277,a352, a376,a399,a533,a706,a875,a969,a1023,a1088,a1230,a1244 |

Table V shows top 20 features that are selected by Correlation Feature selection.These features are selected to construct the feature pool for our GA.

TABLE VI Optimized Features selected by GA from the feature pool

| **DATASET** | **Features Selected by GA** |
|---|---|
| SPECTF-Heart | a28,a35,a2,a4,a5,a6,a8 |
| Soybean-small | a22,a14,a4,a9,a16 |
| Steganalysis | a2,a11,a13,a15,a21,a58,a68,a113,a114,a179,a185,a187,a226,a233 |
| Internet-ad | a5,a20,a82,a533,a1400,a1484 |

Table VI shows the optimized features selected by our hybrid GA approach. The combination of first 20 features selected through filter approaches; Relief and Correlation Feature Selection are given as input to proposed GA wrapper approach to further find the relevant features at the same time maintaining the classification accuracy also. The features selected through GA are as depicted in table 4.8. Thus, combining the advantages of filter and wrapper methods we are getting benefits of the hybrid GA approach that provides us optimized results.

TABLE VII Comparison of % reduction of features between Relief and proposed hybrid GA

| | **% Reduction in Features** | |
|---|---|---|
| **DATASET** | ***RELIEF*** | ***Hybrid GA*** |
| SPECTF-Heart | 83% | 87% |
| Soybean-small | 75% | 86% |
| Steganalysis | 53% | 96.30% |
| Internet-ad | 98.40% | 99.70% |

It can be observed from Table VII that proposed hybrid GA performs better as we are achieving higher classification accuracy and % reduction in features is more in case of hybrid GA, thus obtaining better results.

## VI. CONCLUSION

Feature Selection for classification has a great significance in the current and upcoming times not only in medical sciences but also in other fields.This paper presents a hybrid genetic algorithm approach for improving feature selection.Because of employing different feature choice criteria, diverse feature subset systems frequently give altogether superior results. Our hybrid genetic algorithm coerces numerous features selection criteria to discover subsets of enlightening qualities that can improve classification accuracy.

On average, the strategy based on the hybrid GA approach leads to better prediction rates.

## VII. ACKNOWLEDGEMENT

REFERENCES

[1] Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. J. Mach Learn Res., 3, 1157–1182.

[2] Liu,H. and Motoda,H. (1998) Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Norwell, MA.

[3] Daelemans,W., et al. (2003) Combined optimization of feature selection and algorithm parameter interaction in machine learning of language. A review of feature selection techniques, In Proceedings of the 14th European Conference on Machine Learning (ECML-2003), pp. 84–95.

[4] Rish, Irina, An empirical study of the naive Bayes classifier in IJCAI Workshop on Empirical Methods in AI, 2001.

[5] Feng Tan, Xuezheng Fu, Yanqing Zhang, Anu G. Bourgeois, Improving Feature Subset Selection Using a Genetic Algorithm for Microarray Gene Expression Data, 2006 IEEE Congress on Evolutionary Computation Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006.

[6] Yvan S., In aki I. Pedro L. , A Review of Feature Selection Techniques In Bioinformatics, Bioinformatics,Vol. 23 no. 19 2007, pages 2507–2517.

[7] K. Kira, L. A. Rendell, The feature selection problem: Traditional methods and a new algorithm in Proc. of the Tenth National Conference on Artificial Intelligence, pp. 129-134, 1992.

[8] M. Hall, Correlation-based Feature Selection for Machine Learning, 1999.