Indian Journal of Science and Technology, Vol 9(4), DOI: 10.17485/ijst/2016/v9i4/87039, January 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Metaheuristic Approach for Efficient Feature Selection: A Data Classification Perspective

B. Amarnath[1*] and S. Appavu alias Balamurugan[2]

[1]M. S. University, Tirunelveli - 627012, Tamil Nadu, India; amars_88@yahoo.co.in
[2]Department of I. T., KLNCIT, Sivagangai - 630612, Tamil Nadu, India; datasciencebala@yahoo.com

Abstract

Background/Objectives: Feature selection has been an active research area in pattern recognition, statistics, and data mining communities. The main objective of the proposed feature selection method is to choose a subset of input variables by eliminating features with little or no predictive information using Meta -heuristic approach. **Methods/Statistical Analysis:** The presented method uses genetic algorithm for selecting the optimal feature subset from the datasets. **Findings:** The purpose of this method is to reduce the dimension of the original thereby improves classification accuracy of the selected feature subsets. The experiment performed with various standard dataset revealed that the proposed method is superior to most of the existing feature selection methods in terms of feature subset selection, classification accuracy and running time.

Keywords: Classification and Genetic Algorithms, Data Mining, Feature Selection

1. Introduction

Feature selection is one of the longest existing methods that deal with these problems. Its objective is to select a minimal subset of features according to some reasonable criteria so that the original task can be achieved equally well, if not better. Feature selection is a process of selecting a minimum subset of n features from the original set of N features (where n≤ N) based on certain evaluation criterion. It is a best subset contains the least number of relevant features by discarding the unimportant features from original set of N features.

In general, Feature selection method consists of four basic steps: Subset evaluation, stopping criterion, result validation and three main categories of feature selection methods: Wrapper, Filter and Embedded methods. In filter method, the subset selection is performed without involving any mining algorithms, only be relaying on general characteristics of the data. In this approach, the subset selection is independent of any particular classification algorithm and the selected subsets are supposed to be useful for several different classification algorithms. This method is quite popular mainly due to their computational efficiency even for large datasets but the less computational effort and quality of selected features are the main disadvantages of this method. In wrapper approach, feature selection is performed by taking into account the classification algorithm that will be applied to the selected features. It selects a subset of features that is "optimized" for a given classification algorithm to improve the mining performance. This method selects the best subset with higher predictive accuracy with the expensive computation effort, but the selected feature cannot be used for several different classification algorithms. The embedded method attempts to take advantage of both filter and wrapper methods by exploiting their different evaluation criteria in different search stages.

In feature selection, genetic algorithm is used as a random selection algorithm, capable of effectively exploring large search spaces, which is usually required in case of attribute selection. If the original feature set contains an N number of features, the total number of competing candidate subsets to be generated is 2N, which is a huge

*Author for correspondence

number even for medium sized N. further unlike many search algorithms, which perform a local, greedy search, GA perform a global search. Genetic algorithms have demonstrated substantial improvement over a variety of random and local search methods .this is accomplished by their ability to exploit accumulating information about an initially unknown search space in order to bias subsequent search in to promising subspaces. A Study on Bio-Inspired Metaheuristics for Solving Vehicle Routing Problem was proposed by Yesodha and Amudha[1]

A number of approaches to feature subset selection have been proposed in the literature, a few of them only are referred here. These approaches involve searching for an optimal subset of feature based on some criteria of interest. Michale, Raymer et al.[2] presented a new approach to feature extraction in which feature selection, feature extraction, and classifier training are performed simultaneously using a genetic algorithm. Shah S. C. A. and Kusiak[3] have applied genetic algorithm for feature selection for mining SNP's in association studies. The analysis of SNP's determines relationships between genotypic and phenotypic information and helps in the identification of SNP's related to a disease. Bhargavi and Jothi[4] analyzed the effectiveness of classification algorithm using a genetic algorithm, fuzzy classification and fuzzy clustering on the collected supervised and unsupervised soil data. Laetita Jourdan, Clarisse Dhaennens et al.[5] were interested in discovering genetic feature and environment that are involved in multifactorial disease such as obesity and diabetes. Approach being explored to improve the usefulness of machine learning techniques for generating classification rules for compile, real world data using genetic algorithm. Yang and Jihoon Honavar [6] presented an approach to the multicriteria optimization problem of feature subset selection in GA a framework for efficient feature selection simultaneously by evolving a binary code alongside the chromosome structure used for evolving the rules. HalehVafaie and Ibrahim F. Imam[7] presented a comparison between two feature selection methods, the important score, which is based on a greedy like search and a genetic algorithm based method, in order to better understand their strengths and limitations and their area of applications. A review of current evolutionary algorithms for feature ranking, in data mining tasks involving automated learning was done by Ruxandrastonean and Florin[8]. In[9,10] genetic algorithm is used to determine the attributes which contribute more towards the diagnosis of heart ailments which indirectly reduces the number of tests which are needed to be taken by a patient proposed a feature extraction method using GA and they also used GA for classification and pattern discovery from data. Elakadi, Amine et al.[11] proposed a two stage feature selection algorithm by combining MRMR and GA. In the first stage MRMR is used as a filter to remove the redundant feature and classification performances. Jayaram, Karegowda et al.[12] proposed wrapper method used genetic algorithm as random search technique wrapped with different induction algorithm, namely decision tree C4.5, Bayes networks and radial basis function as subset evaluating mechanism.

2. Proposed Feature Selection Method

In this paper, we introduce a genetic algorithm based for feature selection method. Since the genetic algorithms are basically a domain independent search technique, it is ideal for application where domain knowledge and theory is difficult or impossible to provide. In the proposed method, feature selection the minimum subset with optimal running time. At first, the original dataset is used as input to the genetic algorithm to select a reduced feature subset by removing the irrelevant feature. The feature are chosen at the end of this space are considered as a reliable feature and are sent to the next phase for final evaluation. In the next phase, the test set is used to evaluate the quality of the generated hypotheses with estimated classification accuracy as the performance measure. The performance of the proposed algorithm is evaluated using several UCI datasets. It can see that our method has a strong search capability in the problem space and can discover optimal feature than other existing feature selection methods. The main idea of our feature selection method is enumerated below.

Input full database

- Initialize the population with random individuals of certain size i.e. group individuals with different chromosomes. The size of the initial population should be determined properly by user to include as many possible solutions as possible.
- Evaluate the fitness of individuals in the initial population and rank them according to their fitness.
- Select a certain number of individuals with high fitness of the population and retain them in the next generation.

- Check whether the termination conditions are satisfied, if so the evolution syops and the optimal result representation by the best individual is returned is returned, the evolution continues and the next generation is produced.
- Evaluate the fitness value of all offspring.
- Repeat steps 3–5 until some convergence criteria are met
- Output :selected feature subset

3. System Implementation

The proposed feature selection algorithm using GA carried out in the Matlab 6.5 development environment. Matlab is an engineering and scientific data analysis tool development by the math works, which allows matrix manipulations, implementation of algorithms, and interfacing with programs written in other datasets from the performance of different classification methods in the literature this dataset consists of numeric and nominal attributes. In our experiment, these data sets were the lower scoring features were preserved by the GA while the lowest scoring feature were discarded and the system selects the best feature subsets with reduced number of features. The outputted reduced data set is covered into a CSV file and fed in to the WEKA tool. This selected reduced subset is tested by eight different classification algorithms. Which are available in the tool for our study. All these experiments are performed using 10 fold cross validation method, i.e predictive classification values are averaged over ten disjoint training and testing samples out of the main dataset. Thus the effectiveness of our method is demonstrated through empirical study on UCI data sets and a group of classification algorithms. This empirical evaluation was performed in Intel core i3 CPU running at 3.4 GHz and 4 GB RAM and the results are reported in the following sections.

4. Results and Discussion

In this section, we report our experiments on ten data sets, namely, diabetes, segment-challenge, soybean, vote, ionosphere, dermatology, lung cancer, wine, hepatitis and vehicle. Among the data set the segment challenge is the larger one, which has 20 attribute with 1500 examples while the ionosphere data sets has 35 features with 351 examples. The diabetes datasets has a9 features and 768 examples while the soybean has 36 features with 351 examples. The diabetes datasets has 9 features and 768 examples while the soybean has 36 feature and 683 examples. The summary of this data set is presented in Table 1.

In order to empirically test many of the hypotheses about the benefits and disadvantages of our method we performed several experiments on these datasets by applying eight important classification algorithms. We wanted to compare the performance of the GA based method on different data sets and by evaluating the benefits of the method in terms of selected features, classification accuracy and processing time. This experiment is implemented using MATLAB and a machine learning open source software WEKA. The genetic algorithm is implemented in Matlab to select the best subset from the original feature set. Weka is used to choose the classification algorithm from it its range of classification algorithms to measure the classification accuracy of our proposed method. The eight different classification algorithms are chosen and applied to all the five final selected subsets for our study. Finally, classification accuracy of our proposed method is compared with eight classification algorithms and the results are presented in Table 3.

The objective of this work is to evaluate the performance of the proposed Genetic based feature selection method. To achieve this objective, three evaluation metrics number of features selected, classification accuracy and processing time have been used to test our method with the existing methods. The experimental results of ten datasets are listed in Table 3. As stated earlier, feature subset selection method improves the performance of the

Table 1. Description of Dataset used for the experimentation

Name of the dataset	Number of the attributes	No. of instances
Diabetes	9	768
Segment-challenge	20	1500
Soybean	36	683
Vote	17	435
Ionosphere	35	351
Dermatology	35	366
Lung cancer	57	32
Wine	14	178
Hepatitis	20	155
Vehicle	19	846

classifier, since feature selection is not only concerned with reducing the number of features, but also eliminating the variables that produce nose or, are correlated with other already selected variations. To demonstrate the method, first, the entire sets of features were used in predicting the output. Then the reduced subsets using a genetic feature selector were used to predict the output.

4.1 Feature selection

In this experiment, we compared the proposed method in terms of the number of features selected from the original dataset. By selecting the reduced feature subset, we applied our proposed feature selection algorithm in each dataset and obtained the smallest number of features across all datasets. We implemented the proposed feature selection in Matlab 6.5. The Genetic feature selector is managed to select only very few attributes from these data sets. The results of ten datasets in terms of the number of features selected from the original dataset are listed in Table 2.

By selecting the feature subset, the proposed GFS algorithm obtained the smallest number of features across all datasets except for wine. More significantly, GFS is able to reduce the number of features up to a four feature in vote dataset. Among the ten data sets, genetic feature selector is able to select only 21 features from 57 features for the lung cancer dataset. It also removes adequate features from Ionosphere, vote, Segment challenge, and Dermatology and soybean datasets. From the values shown in the Table, out of ten data sets for vote dataset it selects 25 percentages of features and for the wine it selects the 86 percentages of features with an average dimensionality reduction of 59 percentages of the original data set.

Table 2. Number of features selected from the original dataset with Genetic Feature Selector

Datsets	All features	Features selected	Feature removed	% reduction
Ionosphere	34	14	20	59
Soybean	35	22	13	37
Diabetes	8	4	4	50
Segment	19	8	11	58
Vote	16	4	12	75
Dermatology	35	22	13	37
Lung cancer	57	21	36	63
Wine	14	12	2	14
Hepatitis	20	11	9	45
Vehicle	19	11	9	47

Table 3. Accuracy of each dataset of different classification algorithms

Dataset	Classification Algorithms																	
	NB		J48		SMO		JRIP		DT		Rand. Frt		MI.Prtn		Kstar		Avg. Accuracy	
	All	GA	All	GA	All	GA	All	GA	All	GA	All	GA	All	GA	All	GA	All	GA
Ionosphere	82.6	90.3	91.4	92.5	88.6	88.0	89.7	89.1	89.4	89.4	92.8	92.3	91.1	92.5	84.6	86.6	88.8	90.1
Soybean	92.9	92.0	91.5	90.7	93.7	93.5	91.9	91.3	84.3	80.8	92.0	91.6	93.4	93.9	87.9	89.1	90.9	90.4
Diabetes	76.3	77.4	73.8	74.8	77.3	76.8	76.0	75.5	71.2	73.5	73.8	73.9	75.3	75.5	69.1	69.9	74.1	74.7
Segment	81.0	82.8	95.7	95.4	91.9	89.4	93.7	94	87.4	88.2	96.9	97.8	96.7	94.4	96.6	97.4	92.4	92.4
Vote	90.1	96.0	96.3	96.0	96.0	95.63	95.4	95.4	94.94	95.63	95.63	95.17	94.7	95.86	93.33	96.0	94.6	95.7
Dermatology	97.3	98.1	94	94.3	95.4	97.3	86.9	91	86.7	87.4	94.8	94.2	96.2	96.4	94.3	96.7	93.2	94.4
Lung	50	75	50	62.5	40.6	65.6	43.8	50	92.5	92.5	56.2	65.6	37.5	65.6	40.6	92.5	47.7	63.7
Wine	97.2	97.2	93.8	93.8	98.3	98.3	91.6	89.9	88.8	88.8	97.2	96.6	97.2	99.4	98.9	97.8	95.3	95.2
Hepatitis	84.5	84.5	83.9	83.2	85.2	83.2	78.0	76.8	76.1	77.4	80	83.9	80	83.9	81.9	83.9	81.2	82.1
Vehicle	44.8	48.5	72.5	68.3	74.3	58	68.6	62.6	65.7	66	77	72.8	81.7	71.1	71.4	66.2	69.5	64.2
Avg. Accuracy	79.7	84.2	84.3	85.2	84.1	84.6	81.6	81.6	80.7	80.9	85.6	86.4	84.4	86.9	81.7	84.6	82.8	80.7

Table 4. Running time of GA and full Features on various datasets for each classification algorithm

Dataset	NB		J48		SMO		JRIP		DT		Rnd.Frt		MI.Prtn		Kstar		Avg. Running Time	
	All	GA	All	GA	All	GA	All	GA	All	GA	All	GA	All	GA	All	GA	All	GA
Ionosphere	0.01	0	0.08	0.01	0	0.08	0.04	0.06	0.05	0.13	0.04	0.1	2.27	0.54	0	0	0.31	0.12
Soybean	0	0.01	0.03	0.01	1.49	1.12	0.03	0.09	0.24	0.23	0.19	0.29	31.78	15.94	0	0	4.22	2.21
Diabetes	0.01	0	0.01	0.01	0.09	0.06	0.09	0.09	0.17	0.04	0.16	0.13	0.66	0.46	0	0	0.15	0.10
Segment	0.02	0	0.04	0.02	0.4	0.28	0.19	0.14	0.16	0.07	0.13	0.07	5.31	2.53	0	0	0.78	0.39
Vote	0	0	0	0	0.09	0.01	0.03	0.02	0.15	0	0.15	0.01	0.84	0.2	0	0	0.16	0.03
Dermatology	0.01	0.01	0.03	0.03	0.28	0.21	0.09	0.07	0.19	0.17	0.14	0.1	23.66	10.68	0	0	3.05	1.41
Lung	0	0	0	0	0.02	0.02	0.01	0	0.02	0	0.01	0.01	5.75	0.78	0	0	0.73	0.10
Wine	0	0.01	0.01	0.01	0.02	0.02	0.01	0.1	0.01	0.01	0.01	0.01	0.29	0.23	0	0	0.04	0.05
Hepatitis	0	0	0.05	0	0.05	0.01	0.03	0.01	0.04	0.01	0.09	0.01	0.36	0.15	0	0	0.08	0.02
Vehicle	0.01	0	0.04	0.02	0.07	0.05	0.17	0.06	0.07	0.03	0.07	0.06	2.09	1.13	0	0	0.32	0.17
	0.006	0.003	0.029	0.011	0.251	0.186	0.069	0.064	0.11	0.069	0.099	0.079	7.301	3.264	0	0	0.98	0.46

4.2 Classification Accuracy

To demonstrate the method, first the entire sets of features were used in predicting the output by running all the eight classification algorithms and the values are calculated for the original data sets. Then the reduced subsets were used to predict the output using the same classification algorithms. The results obtained in both cases are shown in Table 3. This table gives the performance of the entire set of features with respect to the reduced dataset generated by our proposed method. From respect to reduced dataset generated by GFS. It is seen from the work that our proposed algorithm reduces the number of features while at the same time increased the classification accuracy. From Table 3, we observe that feature reduction using GFS improves the accuracy of most of the classifiers and some of the values are relatively constant. The classification accuracy of all datasets for each classification algorithm is shown in Figures 4–13.

It is seen from the table that our proposed method is efficient in comparison with other methods for most of the datasets. The proposed algorithm reduces the number of features while at the same time increased the classification accuracy. From Table 3, we observe that feature reduction using GA improves the accuracy of most of the classifiers and some of the values are relatively constant. The value in the table shows the average learning accuracy of the different learning algorithms on different feature sets. From the average accuracy over all datsets, we observe our method improves the accuracy of NB, SMO, Decision table, Multilayer perceptron and K-star classifiers. The overall performance of K-star algorithm is better and a significant improvement was shown by NB classifier for all datasets excluding soybean dataset than all other classifiers. The performance of JRIP is not relatively better than other classifiers, however it is relatively constant. From this experiment, we conclude that after attribute reduction by GFS, the classification accuracy either increases or remains relatively same. For all the classifiers, the average performance of a selected feature subset using our method was better than the performance of the original data set. From the experimental results, we observe that, the classification accuracy of a selected feature subset of our method shows superior results than without feature selection.

4.3 Processing Time

After determining the performance of our method with various classification algorithms, the running time for the selected feature subset was evaluated. The Processing time of the classifiers is tabulated in the Table 4 and it gives the related results of running time in comparison with the full dataset without feature selection takes much time to process the datasets, but the running time is fast and acceptable for genetic feature selector. This approach

consistently reduces the computational time across all the data sets by about 50 percentage compared to the running time of the full data sets. By and K-star take very less time to generate the models and multilayer perception takes much time than other classifiers. From our work the average running time of the classifiers of our method is better than the average running time of the original data. Hence, from the statistics obtained from our work, we observe that at an average our proposed method achieved very encouraging results in terms of processing time also. Hence, the reported results indicate that our feature selection strategy using the genetic algorithm can yield a significant reduction in the number of features and simultaneously produce improvements in classification accuracy and processing time. Thus we conclude that attribute reduction, classification accuracy and processing time by our method using genetic algorithms show superior results than other feature selection methods.

5. Conclusion

In this paper we have analyzed the performance of our proposed genetic feature selector on different data sets. The experimental results have shown that proposed method is more effective and gives better results than other feature selection methods. Therefore, our algorithm is a useful tool for selecting the best feature subset with reduced number of features and for better classification accuracy. The results shown previously itself advocate about the capabilities of our method. Specialty of the proposed method can more clearly be stated as follows:

- The proposed genetic feature selector is very simple and light because GA is used to search the optimal subset of features.
- The results indicate that our feature selection strategy yield a significant reduction in number of features from the original datasets.
- Classification accuracy of our method is either equally good or better than many of the existing feature selection methods.
- Our method significantly improves time than any other feature selection methods with fewer features.

6. References

1. Yesodha R, Amudha T. A study on bio-inspired metaheuristics for solving vehicle routing problem. Indian Journal of Science and Technology. 2015; 8(25):1–9.
2. Raymer ML, Punch WF, Goodman ED, Kuhn LA, Jain AK. Dimensionality reduction using genetic algorithms. IEEE Transactions on Evolutionary Computation. 2000; 4(2):164–71.
3. Shah SC, Kusiak A. Data mining and genetic algorithm based gene/SNP selection. Artificial Intelligence in Medicine. 2004; 31(3):183–96.
4. Bhargavi P, Jothi S. Soil classification using data mining Techniques. A comparative study. International Journal of Engineering Trends and Technology. 2011 Jul-Aug:55–9.
5. Jourdan L, Dhaenens CE, Talbi G. A genetic algorithm for feature selection in data micheal mining genetics. 4th Metaheuristics International Conference (MIC'2001); Porto, Portugal. 2001.
6. Yang J, Honavar V. Feature subset selection using a genetic algorithm. The Spring International Series in Engineering Computer Science. 1997. p. 117–36.
7. Vafaie H, Imam IF. Feature selection methods: Genetic algorithms vs. greedy–like search. Proceedings of the 3rd International Fuzzy Systems and Intelligent Control Conference; Louisville, KY. 1994 Mar. p. 1–11.
8. Stoean R, Gorunescu F. A survey on feature ranking by means of evolutionary computation. Annals of the University of Craiova, Mathematics and Computer Science Series. 2013; 40(1):100–5.
9. Anbarasi M, Anupriya E, Iyengar NCSN. Enhanced prediction of heart disease with feature subset selection using genetic algorithm. International Journal of Engineering Science and Technology. 2010; 2(10):5370–6.
10. Pie M, Goodman ED, Punch WF. Feature extraction using genetic algorithms. Michigan State University: GARAGe; 1997.
11. Elakadi A, Amine A, El Ouardighi A, Aboutajdine D. A new gene selection approach based on Minimum Redundancy-Maxmimum Relevance (MRMR) and Genetic Algorithm (GA). IEEE/ACS International Conference on Computer Systems and Applications, AICCSA'09; Rabat. 2009. p. 69–75.
12. Jayaram MA, Karegowda AG, Manjunath AS. Feature subset selection problem using wrapper approach in supervised learning. International Journal of Computer Application. 2010; 1(7):13–7.