

An Effective Algorithm for Correlation Attribute Subset Selection by Using Genetic Algorithm Based On Naive Bays Classifier

Mr. Shiv Kumar Sharma, Mr. Ashwani Kumar, Mr. Ram Kumar Sharma

Abstract— In recent years, application of feature selection methods in various datasets has greatly increased. Feature selection is an important topic in data mining, especially for high dimensional datasets. Feature selection (also known as subset selection) is a process commonly used in machine learning, wherein subsets of the features available from the data are selected for application of a learning algorithm. The main idea of feature selection is to choose a subset of input variables by eliminating features with little or no predictive information. The challenging task in feature selection is how to obtain an optimal subset of relevant and non redundant features which will give an optimal solution without increasing the complexity of the modeling task. Feature selection that selects a subset of most salient features and removes irrelevant, redundant and noisy features is a process commonly employed in machine learning to solve the high dimensionality problem. It focuses learning algorithms on most useful aspects of data, thereby making learning task faster and more accurate. A data warehouse is designed to consolidate and maintain all features that are relevant for the analysis processes.

Index Terms— co-relation based GA, Feature Selection, Feature Selection Methods, Feature Selection Algorithms, GABAS.

I. INTRODUCTION

In recent years, the need to apply feature selection methods in medical datasets has greatly increased. This is because most medical datasets have large number of samples of high dimensional features.

A "feature" or "attribute" or "variable" refers to an aspect of the data. Usually before collecting data, features are specified or chosen. Features can be discrete, continuous, or nominal. Generally, features are characterized as:

- i. Relevant:** These are features which have an influence on the output and their role cannot be assumed by the rest.
- ii. Irrelevant:** Irrelevant features are defined as those features not having any influence on the output, and whose values are generated at random for each example.
- iii. Redundant:** A redundancy exists whenever a feature can take the role of another (perhaps the simplest way to model redundancy).

Shiv Kumar Sharma, M.TECH (CSE) Research Scholar, IEC-CET, Greater Noida

Ashwani Kumar, Assistant Professor, Department of Information Technology IEC-CET, Greater Noida

R. K. Sharma, Assistant Professor, Department of Information Technology, NIET Greater Noida

II. PROBLEM IDENTIFICATION

Feature subset selection can be used as the technique to identifying and removing as many redundant and irrelevant features as possible. This is because:

(i) redundant features do not help in getting a better predictor for that they provide mostly information which is already present in other feature(s), and

(ii) Irrelevant features do not contribute to the predictive accuracy. A number of approaches to feature subset selection have been proposed in the literature, a few of them only are referred here.

(iii) Relevant: These are features which have an influence on the output and their role cannot be assumed by the rest.

But in my research paper we find the many assumption of Co-related feature selection problem in data mining.

Feature Selection is the essential step in data mining. Individual Evaluation and Subset Evaluation are two major techniques in feature selection. Individual Evaluation means assigning weight to an individual feature. Subset Evaluation is construction of feature subset. The general criteria for feature selection methods are the classification accuracy and the class distribution. The classification accuracy does not significantly decrease and the resulting class distribution, given only the values for selected features. Feature Selection can support many applications, it include the problems involving high dimensional data.

Figure 1 describes feature selection steps. The four key steps in feature selection are-

- a. Subset generation
- b. Subset Evaluation
- c. Stopping criteria
- d. Result validation

The feature selection is used to select relevant features by removing irrelevant and redundant features to improve the performance and to speed up the learning process.

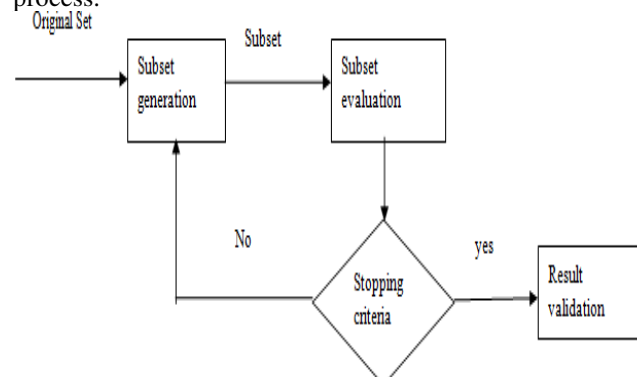


Figure 1. The general procedure for feature selection

An Effective Algorithm for Correlation Attribute Subset Selection by Using Genetic Algorithm Based On Naive Bays Classifier

III. PROPOSED WORK AND METHODOLOGY

Works reported so far in the area of feature subset selection for dimensionality reduction could not claimed that the solution provided by them in the most optimal solution it's because correlation attribute feature subset selection is an optimization problem so the scope of work remains open further and algorithm likes ACO, GA, co-relation based GA, Meta heuristic Search and PSO have been applied to subset selection in the past. In my research paper we are working on, correlation attribute subset selection done by using genetic algorithm that based on naive bays classifier. Its aim is to improve the performance results of classifiers but using a significantly reduced set of features. Genetic Algorithms as an optimization tool is proposed to be applied where Naïve Bayes Classifier will be used to compute the Classification accuracy that will be taken as the fitness value of the individual subset.

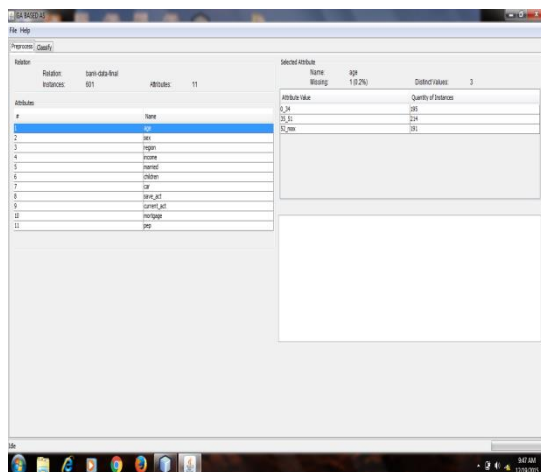
IV. RESULT ANALYSIS

In this proposed method a source bank dataset will be taken as input in arff file; arff is Attribute relation file format. After that all the attributes of datasets are encoded. A number of attributes are select randomly. The classification accuracy is compute with selected attributes. a complete detail of implemented tool is discussed along with the description of results obtained.

A tool is designed in Java to select the subset of features automatically based on GABASS. Tool has a GUI as shown in figure 6.1 where three command buttons are provided named;

- a. File,
- b. Preprocess
- c. Classify

By clicking on the file button an ARFF format dataset file is browsed and taken as input. An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. One such a list of features can be seen in figure 2. An attribute value and its quantities of instances on clicking a particular attribute.



Attribute	Name	Range	Missing	Quantity of Instances
1	sex	0, 1	0	285
2	age	18, 80	0	285
3	marital	0, 1, 2, 3, 4, 5	0	285
4	divorce	0, 1	0	285
5	children	0, 1, 2, 3, 4, 5	0	285
6	nativity	0, 1, 2, 3, 4, 5	0	285
7	income	0, 1, 2, 3, 4, 5	0	285
8	nativity	0, 1, 2, 3, 4, 5	0	285
9	nativity	0, 1, 2, 3, 4, 5	0	285
10	nativity	0, 1, 2, 3, 4, 5	0	285
11	nativity	0, 1, 2, 3, 4, 5	0	285

Figure 2. Attribute list

By clicking on the Classify button a number of input boxes, some check boxes and two buttons are shown in figure 3. An input boxes are used for take user define values in respect to different parameters. All check boxes are optional and used to depend on the user.

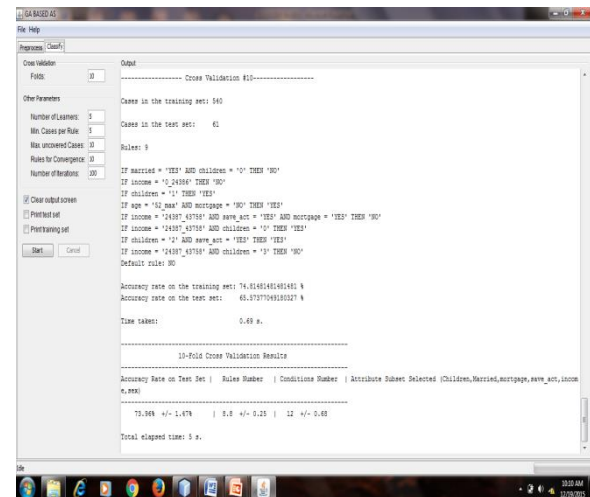
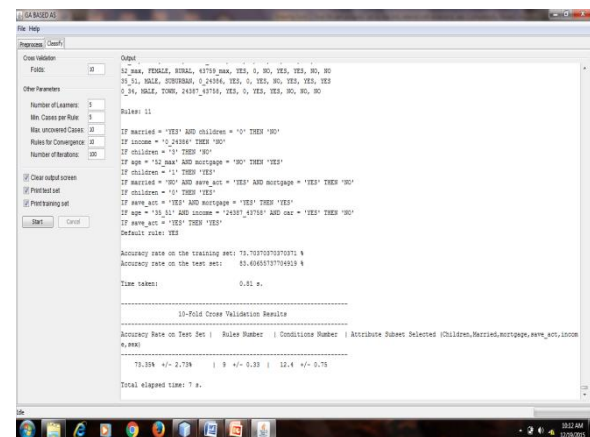


Figure 3 List of parameters

After the k fold validation results, a subset of selected attribute is shown in last of output screen. Here k is taken by the user in term of number. Classification ratio is show the performance of the program.



Attribute	Name	Range	Missing	Quantity of Instances
1	sex	0, 1	0	285
2	age	18, 80	0	285
3	marital	0, 1, 2, 3, 4, 5	0	285
4	divorce	0, 1	0	285
5	children	0, 1, 2, 3, 4, 5	0	285
6	nativity	0, 1, 2, 3, 4, 5	0	285
7	income	0, 1, 2, 3, 4, 5	0	285
8	nativity	0, 1, 2, 3, 4, 5	0	285
9	nativity	0, 1, 2, 3, 4, 5	0	285
10	nativity	0, 1, 2, 3, 4, 5	0	285
11	nativity	0, 1, 2, 3, 4, 5	0	285

Figure 4. Selected subset of Attribute

In this work, five different methods are used for feature selection. Forward Selection Multi cross Validation, Bootstrap backward elimination, Relief, MIFS and proposed GABASS method are implemented and five different feature subsets were obtained. Forward Selection Multi cross Validation and Bootstrap backward elimination are wrapper based method; Relief and MIFS are filter based method. To calculate the classification accuracy for above described methods; SIPINA tool of TANAGRA software is used. The selected feature subsets by these five methods are detailed in following table. The k-fold cross validation method was used to measure the performances.

V. CONCLUSION

In the Feature selection methodology is the first task of any learning approach to define a relevant set of features. Several methods are proposed to deal with the problem of feature selection including filter, wrapper and embedded methods. In this work, I focus on feature subset selection to select a minimally sized subset of optimal features.

Feature Selection is optimization problem; genetic algorithm based attribute subset selection using naïve bayes classifier is used for this purpose. GABASS are found to be the best technique for selection purpose when there is very

large population. The GABASS provides good results and their power lies in the good adaptation to the various and fast changing environments.

VI. FUTURE WORK

Future work will involve experiments on the datasets from different domains. The GABASS algorithm tested on different domains previously. The difference in performance and accuracy of different ensemble approaches will be evaluated.

GABASS can give more efficient results and the optimization process can become much easier and faster. The one more important aspect of future work is of finding more factors that can compare two test suits for their goodness, so that efficiency of selection process can be improved.

REFERENCES

- [1] I. Inza, P. Larranaga, R. Etxeberria, B. Sierra, "*Feature Subset Selection by Bayesian network-based optimization*" Artificial Intelligence 123, pp. 157–184, 2000
- [2] Isabelle Guyon, Andre Elisseeff, "*An Introduction to Variable and Feature Selection*", Journal of Machine Learning Research 3, pp. 1157-1182, 2003.
- [3] Lei Yu, Huan Liu, "*Efficient Feature Selection via Analysis of Relevance and Redundancy*", Journal of Machine Learning Research 5, pp. 1205–1224, 2004
- [4] Felix Garcia Lopez, Miguel Garcia Torres, Belen Melian Batista, Jose A. Moreno Perez, J. Marcos Moreno-Vega, "*Solving feature subset selection problem by a Parallel Scatter Search*" European Journal of Operational Research 169, pp. 477–489, 2006.
- [5] Dunja Mladeni, "*Feature Selection for Dimensionality Reduction*" LNCS 3940, pp. 84–102, 2006.
- [6] C.-R. Jiang, C.-C. Liu, X. J. Zhou, and H. Huang, "Optimal ranking in multi-label classification using local precision rates," Statistica Sinica, vol. 24, no. 4, pp. 1547–1570, 2014.
- [7] M. S. Mohamad, S. Deris, S. M. Yatim, and M. R. Othman, "Feature selection method using genetic algorithm for the classification of small and high dimension data," in Proceedings of the 1st International Symposium on Information and Communication Technology, pp. 1–4, 2004.