In the Proceedings of the Florida AI Research Symposium-FLAIRS-94

AN EMPIRICAL COMPARISON BETWEEN GLOBAL AND GREEDY-LIKE SEARCH FOR FEATURE SELECTION

Ibrahim F. Imam and Haleh Vafaie
Center for Artificial Intelligence
George Mason University
Fairfax, VA, 22030
{iimam@aic.gmu.edu & vafaie@aic.gmu.edu}

ABSTRACT

The paper presents a comparison between two feature selection methods; the Importance Score (IS) and a genetic algorithm-based (GA) method. The goal of both is to achieve better performing rules produced by the AQ15 learning system. The IS method performs a greedy-like search based on an attributional score that represents the importance of each attribute in classifying the decision classes. IS uses the rule testing system Atest to evaluate the performance of the selected feature sets. The genetic algorithm method explores, in an efficient way, the space of all possible subsets to obtain the set of features that maximizes the predictive accuracy of the learned rules. The GA method uses the GENESIS system to globally search the space. It uses an Evaluation Function for providing a feedback about the fitness of each feature subset. The comparison is done on three real world problems, wind bracing design, accident data, and Soybean data.

Key words: feature selection, machine learning, genetic algorithms

1. INTRODUCTION

Whenever the learning process took place, each feature can increase the cost and the time of learning. On the other hand, ignoring one feature may cause problems in achieving high recognition rate. Most of the attempt to solve such a problem is based on selecting features based only on system performance rather than domain knowledge. This approach requires an effectiveness criteria for measuring the performance of the system. One can view systems within this approach as feature selection methods based on the dependencies between the given features. Most of such systems assume that some prior information about the data is available. Moreover, the search procedure plays a very significant role in the success of systems of this approach. Different methods which use this approach for feature selection include genetic algorithms, statistical methods, basic search algorithms, and other heuristic methods.

This paper presents a comparison between two feature selection methods, the Importance Score (IS) and a genetic algorithm-based (GA) method. Both systems perform a search for a set of features from which the AQ learning system (Mich 86) will generate the most predictive accurate rules. The IS (Imam 93) method performs a greedy-like search to obtain the minimum set of features that maximizes the recognition of AQ learned rules. IS uses the rule testing system Atest (Rein 84) for evaluating the performance of the selected feature sets. The genetic algorithm method explores, in an efficient way, the space of all possible subsets to obtain the set of features that maximizes the predictive accuracy of the learned rules. This method uses the GENESIS system (Gref 91) with standard parameter setting to globally search the space. It uses an evaluation function (Vafa 92) for providing a feedback about the fitness of each feature subset. The comparison is done on three real world problems: "Wind bracing" —a data set of wind-bracing design examples classified by the structural worth of buildings (Arci 92), "Accident factors"—a data set of construction accident features in which accident examples are classified by the age of the workers (Arci 92), and “Soybean”—a data set used for learning rules to diagnose several soybean diseases, (Mich 79).

2. FEATURE SELECTION METHODS

2.1. The Importance Score Method

The Importance Score (IS) is a feature selection method that searches for the minimum set of features, the most important ones, which produce better predictive accuracy. The method assigns a score for each feature based on its correlation with the description of the decision classes (which are learned by any of the AQ systems). The features are listed in descending order by their scores and a greedy-like search is performed to select the best feature set. A dynamic threshold is used to start the search with some set of features, then to add a new feature to the selected set. The method uses a user-define tolerance to determine when to end the search. Figure 1 shows a description of the IS method.

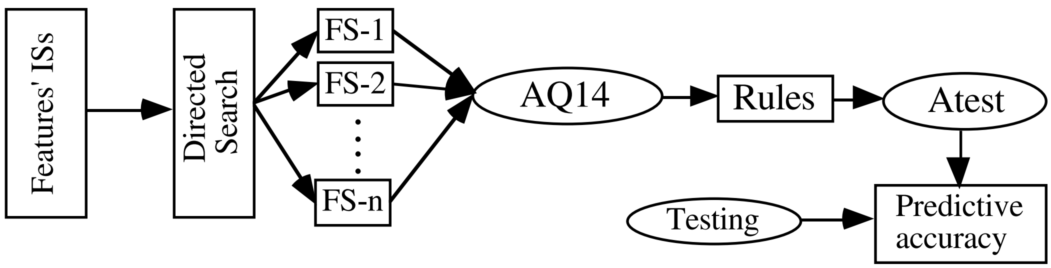

Figure 1: The Importance Score method.

Given n decision classes $C_1,..,C_n$, and m features $A_1,..,A_m$, and assuming that for each feature A_j there are Ec_{ij} examples that match rules containing A_j, and belong to class C_i (i=1,..,n; j=1,..,m), the importance score for A_j is calculated as follows:

$$IS(A_j) = \left(\sum_{i=1}^{n} Ec_{ij}\right) / \left(\max_{j} \sum_{i=1}^{n} Ec_{ij}\right) \qquad (1)$$

The search for the optimal feature set is directed by a dynamic threshold which takes initial value between zero and one (Figure 2). The initial threshold is chosen such that the number of features with higher importance scores is at least equal to the square root of the total number of feature (e.g. for 7 features, the first threshold should be less than the highest three importance scores). Any feature has importance score greater than that threshold is considered relevant. The data is modified to include only the relevant features.

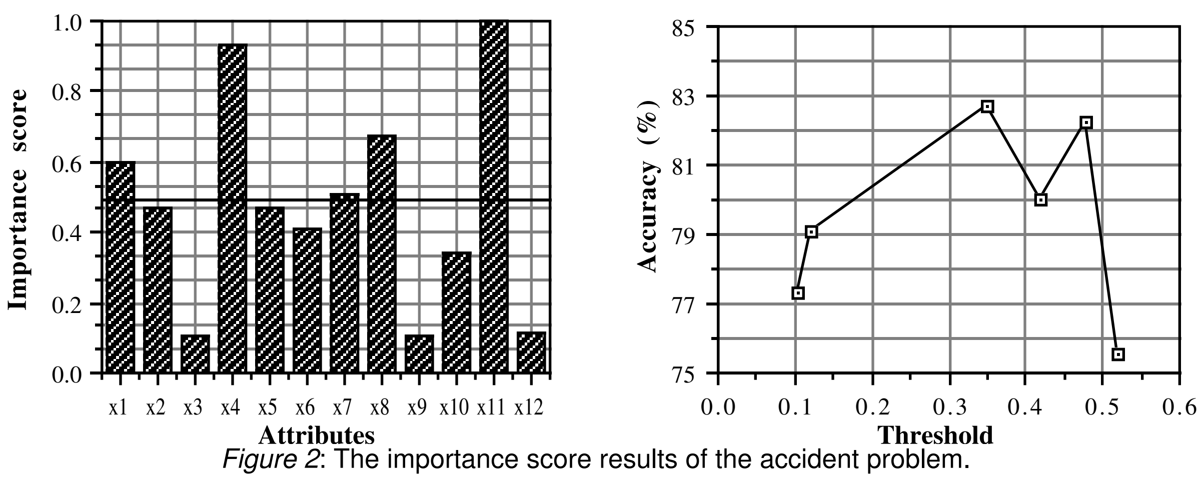

Figure 2: The importance score results of the accident problem.

AQ14 is invoked using the modified data and the output rules are tested against the testing examples. Then, the method reduces the threshold to include another feature. In case of multiple features with the same score, the feature to be selected is the one which exists in rules with the largest number of complexes. A new feature set is selected by reducing the threshold to include only one extra feature, and the method is repeated. The change in the predictive accuracy of the learned rules is reported after each iteration. The search ends whenever the drop in the accuracy is greater than a given tolerance.

In the experiments reported here, the tolerance is set to 5%, and the number of iterations (number of times the threshold is modified and new feature set is selected) did not exceed one more than half the number of original features. The optimal threshold is the value which determines set of features that produces the maximum accuracy. Figures 2 shows the importance scores of the given features and the relation between the threshold and the predictive accuracy for the accident problem.

2.2. Genetic Algorithms

Genetic algorithms (GAs) are adaptive search techniques (Holl 75) that can be good tools for searching for the best set of features. In this research, the architecture of a feature selection system using a genetic algorithm (GA) is given in Figure 3. In this approach an initial set of features as well as a training set of examples in the form of feature vectors extracted from data are provided as input. GAs are used to explore the space of all subsets of the given feature set. Each of the selected feature subsets is evaluated (its fitness measured) by invoking the classification process with the correspondingly reduced feature space and training set, and measuring the accuracy of the rules produced. The best feature subset found is then output as the recommended set of features.

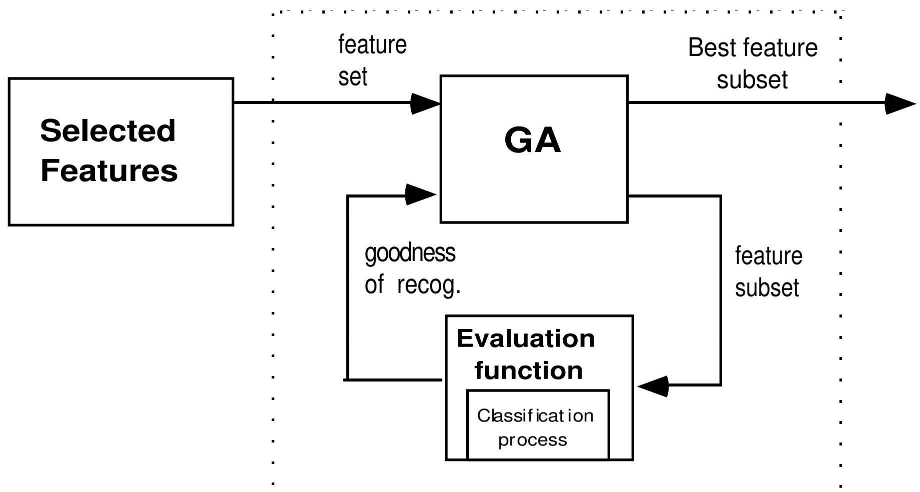

Figure 3: Block diagram of the adaptive feature selection process.

Selecting an appropriate evaluation function and an adequate representation is an essential step for successful application of GAs to any problem domain. This GA method represents each individual as a binary string of length n, where n is the number of features and each bit represents presence or absence of the associated feature. For more details see (Vafa 91). The process of evaluation is shown in Figure 4.

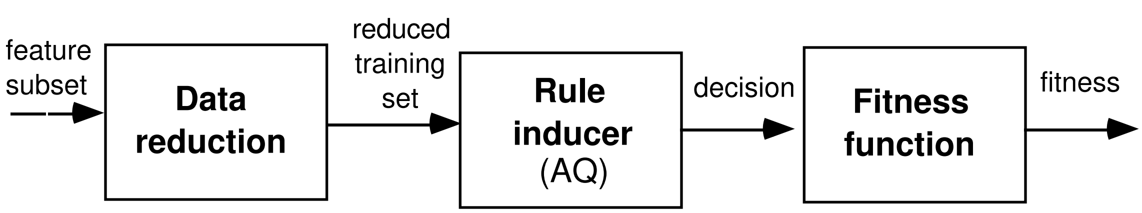

Figure 4: Feature set evaluation procedure.

The evaluation procedure as shown is divided into three steps. After a subset of features is selected, the initial training data is reduced. This is done by removing the values for features that are not in the selected subset of feature vectors. The second step is to apply the AQ system to the new reduced training data to generate the decision rules for each of the given classes in the training data. The last step is to evaluate the rules produced by the classification procedure with respect to their accuracies on the test data.

3. RESULTS AND CONCLUSION

In performing the experiments reported here, AQ14 was used for learning decision rules. In order to keep as many characteristics as possible constant, the parameters of the AQ14 system were held constant for both methods. For the GA approach, GENESIS (Gref 84), a general purpose genetic algorithm program, was used with the standard parameter settings recommended in (DeJo 75): a population size=50, a mutation rate= 0.001, and a crossover rate=0.6.

The comparison is done on three real world problems with different feature sizes. The wind bracing, with small number of features, which describes the structural worth of buildings in terms of different construction features. The Accident factors, which describes the relation between the age of the construction workers and other personal and construction accidental features. The Soybean problem, with large number of features, which describes the relation between a group of fifty symptoms and their relation to seventeen of the soybean diseases.

Figure 5 shows a comparison between the IS and the GA methods. The comparison included the number of iterations, number of selected features, and the predictive accuracy from Atest. Their evaluation function accuracy is indicated between brackets.

Problem		AQ only	IS	GA
Wind bracing	No. of iterations	1	4	85
	No. of features	7	5	5
	Testing accuracy	86.74%	90.43% (92.17%)	90.43% (92.17%)
Accident	No. of iterations	1	7	454
	No. of features	12	6	7
	Testing accuracy	79.22%	80.0% (83.11%)	80.89% (88.88%)
Soybean	No. of iterations	1	10	967
	No. of features	50	14	30
	Testing accuracy	83.66%	91.5% (87.58%)	99.35% (99.35%)

Figure 5: Comparison between IS and GA.

It was very interesting to observe which feature set was selected by both methods for each problem. In the wind bracing problem, the two methods obtained the same set of features {x1, x2, x4, x5, x7}. In the accident data, the importance score method achieved its best performance with a set of six features {x1, x2, x3, x4, x8, x11}, while the genetic algorithm reached its best accuracy with a set of seven features {x1, x2, x4, x5, x6, x7, x10}.

In the soybean data, the importance score achieved its best performance with a set of (14) features {x1, x3, x10, x13, x16, x19, x21, x22, x23, x28, x34, x36, x40, x50}. The genetic algorithm reached its best accuracy with a set of (30) features {x5, x6, x7, x8, x10, x11, x13, x17, x19, x20, x21, x23, x24, x25, x26, x27, x28, x29, x31, x34, x35, x36, x37, x39, x41, x46, x47, x48, x49, x50}.

The results of these experiments show a very strong relation between the size of the original set of features and the behavior of both systems. In simple databases (data that is described by a few number of features), both methods can achieve the best set of features which perform the maximum accuracy, but the GA method does take longer time to achieve such results. Whenever the size of the original set of features increases, the directed search of the Importance Score provides a small feature set with fairly high predictive accuracy. Also, the number of iterations that it takes to learn and test rules is a fixed function of the total number of features. On the other hand, the genetic algorithm search is pointed towards achieving the highest possible predictive accuracy. The Importance Score searches a space of n possible sets, where n is the number of original features, while the genetic algorithm efficiently searches the space of 2^n possible sets.

The results of this experiment show that although the IS method achieved better results than running the AQ algorithm alone, it did not perform as well as the GA-based method. One might suspect that there are a large number of interactions between the features used to describe this data. This then causes the greedy-like search algorithm of IS method to get trapped at a local optima. However, the GA-based method can reach a global optima at the expense of increasing the computational effort. An advantage of the importance score method is that it provide smaller set of features with improved better accuracy in very few search iterations.

ACKNOWLEDGMENT

The authors thank Eric Bloedorn, Ken De Jong, and Ryszard Michalski for their valuable comments.

This research was conducted at George Mason University. The research is supported in part by the National Science Foundation under grant No. IRI-9020266, in part by the Advanced Research Projects Agency under the grant No. N00014-91-J-1854, administered by the Office of Naval Research, and the grant No. F49620-92-J-0549, administered by the Air Force Office of Scientific Research, and in part by the Office of Naval Research under grant No. N00014-91-J-1351.

REFERENCES

[Arci 92] Arciszewski, T., Bloedorn, E., Michalski, R., Mustafa, M., and Wnek, J., "Constructive Induction in Structural Design", Report of Machine Learning and Inference Labratory, MLI-92-7, Center for AI, George Mason Un., 1992.

[DeJo 75] De Jong, K., "*Analysis of the behavior of a class of genetic adaptive systems*," Ph.D. Thesis, Department of Computer and Communications Sciences, University of Michigan, Ann Arbor, MI., 1975.

[Gref 91] Grefenstette, John J., Davis, L., and Cerys, D., "Genesis and OOGA: Two Genetic Algorithms Systems," TSP: Melrose, MA. 1991.

[Holl 93] Holland, J. H.,. "*Adaptation in Natural and Artificial Systems*," University of Michigan Press, Ann Arbor, MI., 1975.

[Imam 93] Imam, I.F., Michalski, R.S., and Kerschberg, L., “Discovering Attribute Dependence in Databases by Integrating Symbolic Learning and Statistical Analysis Techniques”, Proceeding of AAAI-93 Workshop on Knowledge Discovery in Databases, Washington D.C., July 11-12, 1993.

[Mich 79] Michalski, R.S. and Chilausky, R.L., “Knowledge Acquisition by Encoding Expert Rules vs. Computer Induction from Examples: a Case Study Involving Soybean Pathology”, in the International Journal for Man-Machine Studies, May, 1979.

[Mich 86] Michalski, R.S., Mozetic, I., Hong, J., and Lavrac, N., “The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains,” *Proceedings of AAAI-86*, pp. 1041-1045, Philadelphia, PA: , 1986.

[Rein 84] Reinke, R.E., “Knowledge Acquisition and Refinement Tools for the ADVISE Meta-Expert System”, Master thesis, ISG 84-4, Urbana, Illinois, July, 1984.

[Vafa 91] Vafaie, H., and De Jong, K., “Imporving the Performance of a Rule Induction System Using Genetic Algorithms”, Proceeding of the First International Workshop on Multistrategy Learning, Harpers Ferry, WV., 1991.

[Vafa 92] Vafaie, H., and De Jong, K., “Genetic Algorithms as a Tool for Feature Selection in Machine Learning”, Proceeding of the 4th International Conference on Tools with Artificial Intelligence, Arlington, VA, November, 1992.