Feature Selection Methods: Genetic Algorithms vs. Greedy-like Search

Haleh Vafaie and Ibrahim F. Imam
George Mason University, Fairfax, VA, 22030

Abstract. This paper presents a comparison between two feature selection methods, the Importance Score (IS) which is based on a greedy-like search and a genetic algorithm-based (GA) method, in order to better understand their strengths and limitations and their area of application. The results of our experiments show a very strong relation between the nature of the data and the behavior of both systems. The Importance Score method is more efficient when dealing with little noise and small number of interacting features, while the genetic algorithms can provide a more robust solution at the expense of increased computational effort.
Keywords. feature selection, machine learning, genetic algorithms, search.

1. INTRODUCTION

Feature selection is a problem that has to be addressed in many areas, especially in artificial intelligence. The main issues in developing feature selection techniques are choosing a small feature set in order to reduce the cost and running time of a given system, as well as achieving an acceptably high recognition rate. This has led to the development of a variety of techniques for selecting an optimal subset of features from a larger set of possible features. These feature selection techniques fall into two main categories. In the first approach problem specific strategies are developed based on the domain knowledge in order to reduce the number of features used to a manageable size (Dom 89). The second approach is used when the domain knowledge is unavailable or expensive to exploit. In this case, generic heuristics, essentially greedy algorithms, are applied to select a subset "d" of the available "m" features (Kittler 78).

The experiments reported in this paper compare two techniques which belong to the second category, in order to better understand their strengths and limitations and their area of application. The selected feature selection algorithms (Important Score method and GA-based technique) contain the basic components as shown in Figure 1 (Vafaie 93). The search procedures used by the Importance Score (IS) technique and the genetic algorithm-based (GA) method require no domain knowledge to assist the search process. Both IS and GA systems perform a search for achieving the highest predictive accuracy of rules produced by the AQ learning system as evaluated by their criterion function. The IS method performs a greedy-like search to obtain the minimum set of features that maximizes the recognition of AQ learned rules, while 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.

where the importance scores are calculated. These rules are then examined by using a fitness function which will be described later.

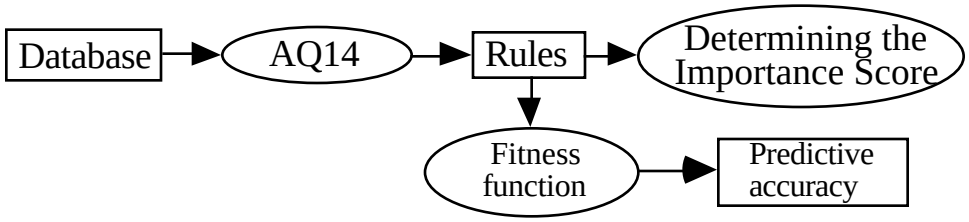

Figure 2: The importance score method.

The process of finding a feature subset using the Importance Score method is shown in Figure 3. To determine importance scores, the first step is to run AQ14 (for more information see Reinke, 1984) against the training data to learn initial rules. Each decision class is described by a set of rules, where a rule consists of a set of conditions or complexes. There are, two weights, t-weight and u-weight, that are assigned to each rule in order to measure its strength. The *t*-weight (*total-weight*) of a rule belonging to a class is the number of examples of that class which are covered by that rule. The u-weight (*unique-weight*) of a rule describing a class is the number of examples of that class which are only covered by this rule. The importance score of a feature A is calculated by first adding the t-weights of all the rules that contain A in one of their complexes, and dividing this sum by the total number of t-weights in all of the rules. The importance score ranges from zero, whenever a feature does not appear in any of the rules, to one when the feature is very valuable for describing the learned rules (i.e. the feature exists in every learned rule). More formally, the Importance score for a feature Aj is given by:

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

where: n is the number of decision classes,

m is the number of features, A1, ..., Am, and

Ec_{ij} is the number of examples matching the rules that contain Aj, and belong to class Ci (i=1,..,n; j=1,.., m)

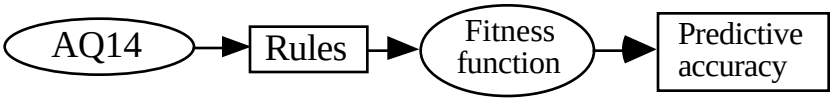

Figure 3: The criterion function for IS method

The IS method then ranks the features in descending order measured by their scores, and search for the minimum set of features with highest scores which performs the best predictive accuracy. Unlike the normal greedy search, the importance score starts the search with a set of features. Its search is directed by a dynamic threshold which takes initial value between zero and one. 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 features (e.g. for 7 features, the first threshold should be less than the highest three importance scores). Then, the features which have importance score greater than that threshold are considered relevant, and the data is modified to include only the relevant features. The rule inducer system AQ14 is ran against the modified data and the output rules are tested using the testing examples. The process continues by reducing the threshold to include another feature. In the case of having more than one feature with the same important score, the feature to be selected is the one which belongs to a rule with large number of complexes. The method stops whenever the decrease in the predictive accuracy is greater than a given tolerance. In the experiments done so far, the tolerance is set to 5%. The optimal threshold is the value that when applied in conjunction with the Chi-square relevant method produces the maximum accuracy.

The Chi-square test is applied for determining the correlation between the decision classes and the features. The Chi-square test is only used when it is significant with the data. Then, the original data is modified to include the union of features that are statistically relevant and all of those with higher importance score than the optimal threshold.

As mentioned in earlier work, while Chi-square is used to statistically ascertain the correlation between the decision classes and the other features, the importance score is used to heuristically verify the relation between them. In some cases, the Chi-square test was not of help in discovering such a good correlation. More details on the Chi-square method can be found in (Chan, 1991; Imam, 1993).

The performance of a feature subset and the predictive accuracy of the AQ produced classification rules are measured by applying a fitness function which will be described in later sections.

The AQ algorithm

The AQ algorithm is a rule induction technique used to produce a complete and consistent description of classes of examples [Michalski 86, 83]. A class description is formed by a collection of disjuncts of decision rules describing all the training examples given for that particular class. A decision rule is simply a set of conjuncts of allowable tests of feature values. For more detailed description see [Vafaie 91].

Fitness Function

In order to efficiently use the criterion function, it is necessary to define a fitness function which properly assesses the decision rules generated by the AQ algorithm. The fitness function must be able to discriminate between correct and incorrect classification of examples, given the AQ

created rules. Finding an appropriate function is not a trivial task, due to the noisy nature of most real world data.

The fitness function takes as an input a set of feature or attribute definitions, a set of decision rules created by the AQ algorithm, and a collection of testing examples defining the feature values for each example. The fitness function then evaluates the AQ generated rules on the testing examples as follows.

For every testing example a match score (Vafaie 91) is evaluated for each of the classification rules generated by the AQ algorithm, in order to find the rule(s) with the highest or best match. At the end of this process, if there is more than one rule having the highest match, one rule will be selected based on the chosen conflict resolution process. This rule then represents the classification for the given testing example. After all the testing example have been classified using AQ generated rules, the overall fitness function will be evaluated by adding the weighted sum of the match score of all the correct recognitions and subtracting the weighted sum of the match score of all of the incorrect recognitions. For a detailed description of the fitness function see (Vafaie, 1993).

2.2 *Genetic algorithm-Based Method*

The presented method uses a genetic algorithm for feature selection. Genetic algorithms (GAs), a form of inductive learning strategy, are adaptive search techniques initially introduced by Holland (Holland, 1975). Genetic algorithms derive their name from the fact that their operations are similar to the mechanics of genetic models of natural systems.

Genetic algorithms typically maintain a constant-sized population of individuals which represent samples of the space to be searched. Each individual is evaluated on the basis of its overall fitness with respect to the given application domain. New individuals (samples of the search space) are produced by selecting high performing individuals to produce "offspring" which retain many of the features of their "parents". This eventually leads to a population that has improved fitness with respect to the given goal.

New individuals (offspring) for the next generation are formed by using two main genetic operators, crossover and mutation. Crossover operates by randomly selecting a point in the two selected parents gene structures and exchanging the remaining segments of the parents to create new offspring. Therefore, crossover combines the features of two individuals to create two similar offspring. Mutation operates by randomly changing one or more components of a selected individual. It acts as a population perturbation operator and is a means for inserting new information into the population. This operator prevents any stagnation that might occur during the search process.

Genetic algorithms have demonstrated substantial improvement over a variety of random and local search methods (De Jong, 1975). This is accomplished by their ability to exploit accumulating information about an initially unknown search space in order to bias subsequent search into promising subspaces. Since GAs are basically a domain independent search technique, they are ideal for applications where domain knowledge and theory is difficult or impossible to provide (De Jong, 1988).

The main issues in applying GAs to any problem are selecting an appropriate representation and an adequate evaluation function. For detailed description of both of these issues for the problem of feature selection see (Vafaie, 1991).

Representation

The natural representation for the feature selection problem is precisely the one described earlier, namely a binary string of length N representing the presence or absence of each of the N possible features. The advantage of this representation is that the classical GA's operators as described before (binary mutation and crossover) can easily be applied to this representation without any modification. This eliminates the need for designing new genetic operators, or making any other changes to the standard form of genetic algorithms.

Evaluation function

Selecting an appropriate evaluation function is an essential step for successful application of GAs to any problem domain. Evaluation functions provide GAs with the feed-back about the fitness of each individual in the population. GAs then use this feed-back to bias the search process so as to provide an improvement in the population's average fitness.

The process of evaluation involves the steps presented in Figure 4 (Vafaie & De Jong, 1991). The evaluation function is solely based on the performance of the classification process used, in order to select the appropriate feature set, without attempting to bias the search toward small feature subsets.

3. EXPERIMENTAL RESULTS

In performing the experiments reported here, AQ14 was used for learning decision rules. In order to keep as many things as possible constant, the parameters of the AQ14 system were held constant for both methods. For the GA approach, GENESIS (Grefenstette, 1984), a general purpose genetic algorithm program, was used with the standard parameter settings recommended in (De Jong, 1975): a population size=50, a mutation rate= 0.001, and a crossover rate=0.6. Table 1 summarizes the results of our experiments comparing the best performance of the IS method to that of the GA-based strategy.

In the first experiment, a natural database designed for diagnosing breast cancer (Michalski, 1986) was used. This data was originally extracted from the MLI database of George Mason University and each case was described using nine features. There is a very small amount of noise in the data. In this experiment both methods find an optimal subset, but the IS method requires fewer iterations.

The second experiment is based on texture images which were randomly selected from Brodatz (Brodatz, 1966) album of textures as depicted in Figure 5. Two hundred feature vectors, each containing eight features were randomly extracted from an arbitrary selected area from each of the chosen textures.

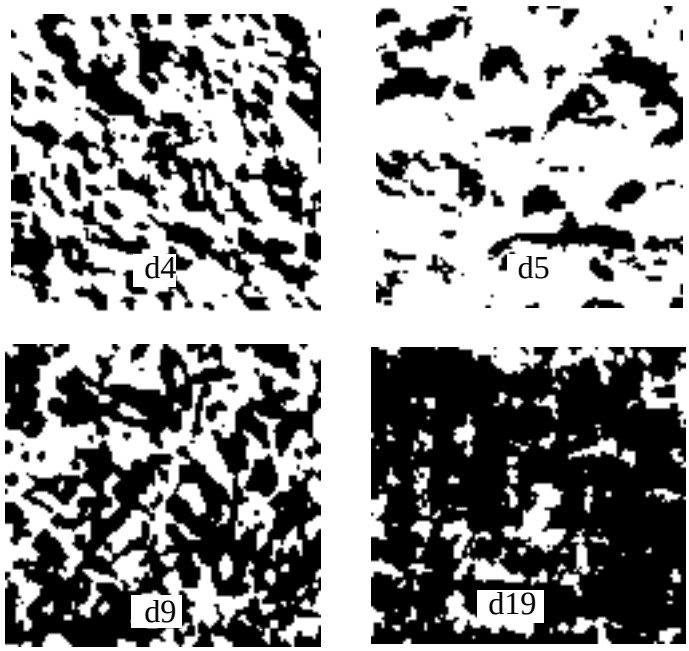

Figure 5: The texture images used in experiment 2.

As it is shown in Table 1, the IS method was unable to improve the predictive accuracy with respect to running the AQ algorithm alone. This is the effect of the fact that image data, specially textures, are very noisy. However, the GA-based method improved the results by producing a higher accuracy and reducing the number of features required for learning the

classification rules. This is a result of GA's insensitivity to noise. The robustness of GA-based method is quite evident here.

In the last experiment, the coal data was used in order to predict the amount of electricity generated from it given examples of the state by state demand where, each example is the demand of a state for a year. These examples are described using 19 features.

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.

Problem		AQ	IS	GA
Breast Cancer	No. of iterations	1	6	93
	No. of features	9	3	3
	Testing accuracy	61.63	75.58	75.58
Texture	No. of iterations	1	7	285
	No. of features	8	8	3
	Testing accuracy	70.35	70.35	78.90
Coal	No. of iterations	1	10	504
	No. of features	19	8	10
	Testing accuracy	71.86	82.63	93.41

Table 1: The result of the performed experiments
No. of iterations is the number of times AQ is used to learn classification rules

4. SUMMARY AND CONCLUSIONS

The goal of the research reported here was to understand better their strengths and limitations and the area of applications of a greedy-like and a GA-based feature selection algorithms and to use that knowledge to develop more robust approaches. The results suggest that the IS method and in general greedy-like searches have a tendency to get trapped on local peaks caused by noise or interdependencies among features. It should also be noted that the IS method runs fast and is more efficient when there is little noise or the number of interacting features is small. The GA-based method proved quite effective in improving the robustness of feature selection over a range of problems at the expense of increased computational complexity.

An interesting open question is whether a multistrategy approach could be developed which could combine the two methods since, in general, information about the degree of interactions or the amount of noise is not available a priori.

ACKNOWLEDGMENT

The authors wish to thank Mr. Ron Capone for providing the authors with the coal data and Bill Spears for furnishing the breast cancer data. We also like to thank Dr. Ken De Jong, Mitch Potter, and Eric Bloedorn 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 Defense 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

Brodatz, P., "*A Photographic Album for Arts and Design*," Dover Publishing Co., Toronto, Canada, 1966.

Chan, K. C., and Wong, A. K., "A Statistical Technique for Extracting Classificatory Knowledge from Databases", *Knowledge Discovery In Databases*, Piatetsky-Shapiro, G., Frawley, W., (Eds.), AAAI Press, 1991.

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.

De Jong, K., "Learning with Genetic Algorithms : An overview," *Machine Learning* Vol. 3, Kluwer Academic publishers, 1988.

Dom, B., Niblack, W., and Sheinvald, J., "Feature selection with stochastic complexity," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Rosemont, IL., 1989.

Grefenstette, John J., Technical Report CS-83-11, Computer Science Dept., Vanderbilt Univ., 1984.

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

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

Kittler, J., " Feature set search algorithms," in *Pattern Recognition and Signal Processing*, C.H. Chen, Ed., Sijthoff and Noordhoff, The Netherlands, 1978.

Michalski, R.S., "A Theory and Methodology of Inductive Learning", *Artificial Intelligence*, Vol. 20, pp. 111-116, 1983.

Michalski, R.S., "AQVAL/1-Computer Implementation of a Variable-Valued Logic System VL1 and Examples of its Application to Pattern Recognition" in *Proceeding of the First International Joint Conference on Pattern Recognition*, Washington, DC, pp. 3-17, October 30-November, 1973.

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.

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

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.

Vafaie, H. and De Jong, K., "Imporving the Performance of a Rule Induction System Using Genetic Algorithms", in Machine Learning: A Multistrategy Approach, Vol. IV, R.S. Michalski and G.Tecuci (Eds.), Morgan Kaufmann, San Mateo, CA, 1993.

Vafaie, H., and De Jong, K.A., "Improving the performance of a Rule Induction System Using Genetic Algorithms," *Proceedings of the First International Workshop on MULTISTRATEGY LEARNING*, Harpers Ferry, W. Virginia, USA, 1991.