Genetic Algorithms as a Tool for Feature Selection in Machine Learning

Haleh Vafaie and Kenneth De Jong

Center for Artificial Intelligence, George Mason University

Abstract

This paper describes an approach being explored to improve the usefulness of machine learning techniques for generating classification rules for complex, real world data. The approach involves the use of genetic algorithms as a "front end" to traditional rule induction systems in order to identify and select the best subset of features to be used by the rule induction system. This approach has been implemented and tested on difficult texture classification problems. The results are encouraging and indicate significant advantages to the presented approach in this domain.

1.0 Introduction

In recent years there has been a significant increase in research on automatic image recognition in more realistic contexts involving noise, changing lighting conditions, and shifting viewpoints. The corresponding increase in difficulty in designing effective classification procedures for the important components of these more complex recognition problems has led to an interest in machine techniques as a possible strategy for automatically producing classification rules. This paper describes part of a larger effort to apply machine learning techniques to such problems in an attempt to generate and improve the classification rules required for various recognition tasks. The immediate problem attacked is that of texture recognition in the context of noise and changing lighting conditions. In this context standard rule induction systems like AQ15 produce sets of classification rules which are sub-optimal in two respects. First, there is a need to minimize the number of features actually used for classification, since each feature used adds to the design and manufacturing costs as well as the running time of a recognition system. At the same time there is a need to achieve high recognition rates in the presence of noise and changing environmental conditions.

This paper describes an approach being explored to improve the usefulness of machine learning techniques for such problems. The approach described here involves the use of genetic algorithms as a "front end" to traditional rule induction systems in order to identify and select the best subset of features to be used by the rule induction system. The results presented suggest that genetic algorithms are a useful tool for solving difficult feature selection problems in which both the size of the feature set and the performance of the underlying system are important design considerations.

2.0 Feature Selection

Since each feature used as part of a classification procedure can increase the cost and running time of a recognition system, there is strong motivation within the image processing community to design and implement systems with small feature sets. At the same time there is a potentially opposing need to include a sufficient set of features to achieve high recognition rates under difficult conditions. This has led to the development of a variety of techniques within the image processing community for finding an "optimal" subset of features from a larger set of possible features. These feature selection strategies fall into two main categories.

The first approach selects features independent of their effect on classification performance. The difficulty here is in identifying an appropriate set of transformations so that the smaller set of features preserve most of the information provided by the original data and are more reliable because of the removal of redundant and noisy features.

The second approach directly selects a subset "d" of the available "m" features in such a way as to not significantly degrading the performance of the classifier system [5]. The main issue for this approach is how to account for dependencies between features when ordering them initially and selecting an effective subset in a later step.

The machine learning community has only attacked the problem of "optimal" feature selection indirectly in that the traditional biases for simple classification rules (trees) leads to efficient induction procedures for producing individual rules (trees) containing only a few features to be evaluated. However, each rule (tree) can and frequently does use a different set of features, resulting in much larger cumulative features sets than those typically acceptable for image classification problems. This problem is magnified by the tendency of traditional machine learning algorithms to overfit the training data, particularly in the context of noisy data, resulting in the need for a variety of ad hoc truncating (pruning) procedures for simplifying the induced rules (trees).

The conclusion of these observations is that there is a significant opportunity for improving the usefulness of traditional machine learning techniques for automatically generating useful classification procedures if there were an effective means for finding feature subsets which are "optimal" from the point of view of size and performance. In the following sections an approach using genetic algorithms is described in some detail and its effectiveness illustrated on a class of difficult texture recognition problems.

3.0 Feature Selection Architecture

The overall architecture of the proposed system is given in Figure 1. It is assumed that an initial set of features will be provided as input as well as a training set representing positive and negative examples of the various classes for which classification is to be performed. A search procedure is used to explore the space of all subsets of the given feature set. The performance of each of the selected feature subsets is measured by invoking an evaluation function with the correspondingly reduced feature space and training set, and measuring the specified classification result. The best feature subset found is then output as the recommended set of features to be used in the actual design of the recognition system.

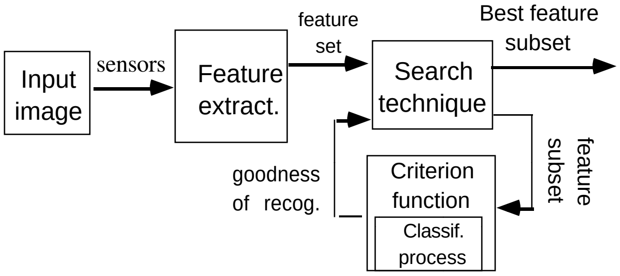

Figure 1: Block diagram of the adaptive feature selection process

The performance of a feature subset is measured by applying the evaluation procedure presented in Figure 2. The evaluation procedure as shown is divided into three main steps. After a feature subset is selected, the initial training data, consisting of the entire set of feature vectors and class assignments corresponding to examples from each of the given classes, is reduced. This is done by removing the values for features that are not in the selected subset of feature vectors.

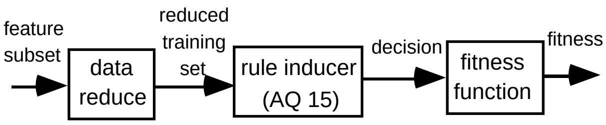

Figure 2: Feature set evaluation procedure

The second step is to perform rule induction on the new reduced training data in order to generate classification rules for use in a recognition system. In our case we use AQ15, a rule induction technique used to produce a complete and consistent description of classes of examples [6]. A class description is formed by a set 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 a more detailed description see [7].

The last step is to evaluate the classification performance of the induced rules on the unseen test data. How this is done varies from one feature selection method to another and will be described more precisely in the following sections.

4.0 Feature Selection Through Traditional Statistical Methods

Our first attempt at improving the performance of the texture classifier involved the use of a traditional statistical feature selection method. Such methods involve defining both a *search procedure* and an *evaluation procedure.*

A standard approach to feature selection involves the use of sequential backward selection (SBS), a top down search procedure that starts with the complete set of features and discards one feature at a time until the desired number of features have been deleted. For detailed description see [3].

As noted earlier, the performance of a feature subset is measured by applying the evaluation process presented in Figure 2. After a feature subset is selected, AQ15 is applied to the new reduced training data to generate the decision rules for each of the given classes in the training data. The final step is to evaluate the rules produced by AQ15 with respect to their classification performance on unseen test data. This last step varies from method to method. We have adopted a statistical measure of fitness based on Euclidean distance measures of class separability which is frequently used in traditional feature selection techniques.

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 (for more detailed description see [7]) 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, a statistical separability measure is computed as the estimate of fitness of the given feature set. The basic idea is to find feature subsets which increase (maximize) the distance between the classes to be recognized [3]. More formally, one would like to maximize

$$J(e) = 1/2 \sum_{i=1}^{c} P_i \sum_{j=1}^{c} P_j \, 1/n_i \, n_j \sum_{k=1}^{n_i} \sum_{l=1}^{n_j} \partial(e_{ik}, e_{jl})$$

where, $\partial(e_{ik}, e_{jl})$ represents the distance between two elements. Typically, this distance measure is Euclidean distance since it allows for both analytical and computational simplifications of the interclass distance criterion [3]. Then, $\partial(e_{ik}, e_{jl}) = (e_{ik} - e_{jl})^t (e_{ik} - e_{jl})$. For detailed explanation refer to [3].

4.1 Initial Experimental Results

The AQ15 system used for rule induction has a number of parameters which affects its own performance on a given problem class. An attempt was made to identify reasonable values for these parameters for the texture classification problems used. (for more details, see [7]).

In these experiments four texture images were randomly selected from Brodatz [1] album of textures. These images are water, beach pebbles, hand made paper, and cotton canvas as depicted in [1] and [7]. Two hundred feature vectors, each containing 18 features were then randomly extracted from an arbitrary selected area of 30 by 30 pixels from each of the chosen textures. These feature vectors were divided equally between training examples used for the generation of decision rules, and testing examples used to measure the performance of the produced rules.

The initial experimental results using the traditional SBS feature selection technique described above are summarized in Figures 3-5. Figure 3 shows that some improvement in Euclidean separability measure was achieved by using the SBS search technique to produce trial feature sets for testing and evaluation. Figure 4 indicates a corresponding decrease in the size of the feature set. However, in Figure 5, we see that the recognition rate (measured in terms of the % of correct classifications) has clearly decreased. This is due in part to the fact that statistical separability measures (based on Euclidean distance) do not necessarily correlate directly to classification performance. In our case, this effect is compounded by the inherent noise in the image data. Both the AQ15 program and the SBS search procedure, by trying to produce optimal results for the training data, can easily overfit the noisy data resulting in actual decreases in performance on unseen test data.

Our hypothesis, based on these initial results, was that a more robust feature selection strategy was required in order to simultaneously improve the feature selection and the classification performance in these kinds of noisy domains.

5.0 Feature Selection Using GAs

Genetic algorithms (GAs) are best known for their ability to efficiently search large spaces about which little is known *a priori*. Since genetic algorithms are relatively insensitive to noise, they seem to be an excellent choice for the basis of a more robust feature selection strategy for improving the performance of our texture classification system. In this section we describe this approach in more detail.

5.1 Genetic Algorithms

Genetic algorithms (GAs), a form of inductive learning strategy, are adaptive search techniques which have demonstrated substantial improvement over a variety of random and local search methods [2]. 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 [2].

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 [7].

In the feature selection problem the main interest is in representing the space of all possible subsets of the given feature set. Then, the simplest form of representation is binary representation where, each feature in the candidate feature set is considered as a binary gene and each individual consists of fixed-length binary string representing some subset of the given feature set. An individual of length l corresponds to a l-dimensional binary feature vector X, where each bit represents the elimination or inclusion of the associated feature. Then, $x_i = 0$ represents elimination and $x_i = 1$ indicates inclusion of the ith feature.

5.2 Evaluation function

Choosing an appropriate evaluation function is an essential step for successful application of GAs to any problem domain. As before, the process of evaluation involved the steps presented in Figure 2. The only variation was to implement a more performance-oriented fitness function that is better suited for genetic algorithms. In order to use genetic algorithms as the search procedure, it is necessary to define a fitness function which properly assesses the decision rules generated by the AQ algorithm. Each testing example is classified using the AQ generated rules as described before. If this is the appropriate classification, then the testing example has been recognized correctly. After all the testing examples have been classified, the overall fitness function will be evaluated by adding the weighted sum of the match score of all of the

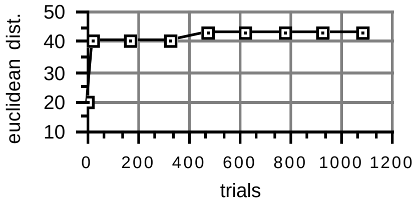

Figure 3: The improvement of Euclidean distance measure over time

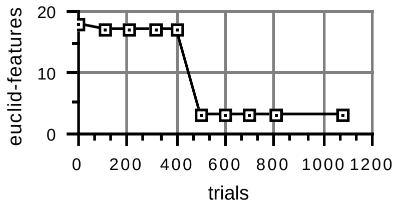

Figure 4: The number of features used by the best individual

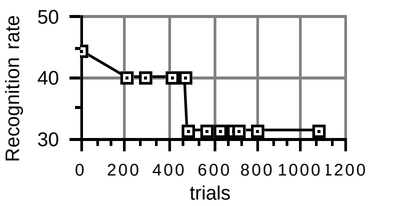

Figure 5: The improvement in feature set fitness over time

correct recognitions and subtracting the weighted sum of the match score of all of the incorrect recognitions (for a detailed explanation see [7]), i.e.

$$\mathbf{F} = \sum_{i=1}^{n} S_i * W_i - \sum_{j=n+1}^{m} S_j * W_j$$

The range of the value of **F** is dependent on the number of testing events and their weights. In order to normalize and scale the fitness function **F** to a value acceptable for GAs, the following operations were performed:

$$\text{Fitness} = 100 - [(\mathbf{F} / \text{TW}) *100]$$

where:

$$\text{TW} = \text{total weighted testing examples} = \sum_{i=1}^{m} W_i$$

As indicated in the above equations, after the value of **F** was normalized to the range [-100, 100], the subtraction ensures that the final evaluation is always positive (the most convenient form of fitness for GAs), with lower values representing better classification performance.

5.3 Experimental Results

In performing the experiments reported here, the same AQ15 system was used with the same parameter settings as described earlier. In addition, GENESIS [4], a general purpose genetic algorithm program, was used as the search procedure (replacing SBS). We used the standard parameter settings for GENESIS.

In the experiments reported for the GA-based approach, equal recognition weights (i.e., W=1) were assigned to all the classes in order to perform a fair comparison between the two presented approaches. The experiments were performed on the texture images described before. The results are summarized in Figures 6 and 7 and provide encouraging support for the presented GA approach. Figure 6 shows the steady improvement in the fitness of the feature subsets being evaluated as a function of the number of trails of the genetic algorithm. This indicates very clearly that the performance of rule induction systems (as measured by recognition rates) can be improved in these domains by appropriate feature subset selection.

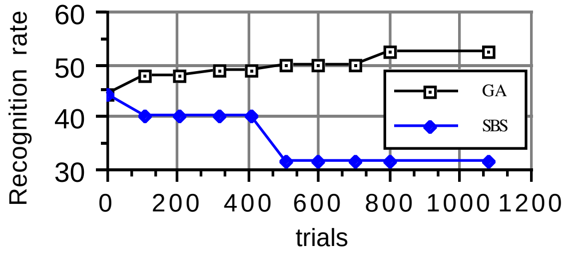

Figure 6: The improvement in feature set fitness over time

Figure 7 shows that the number of features in the best feature set decreased for both approaches. However, the feature subset found by statistical measures was substantially smaller than that found by the GA-based system. Figure 6 indicates that this was achieved at the cost of poorer performance. The advantage of the GA approach is to simultaneously improve both figures of merit.

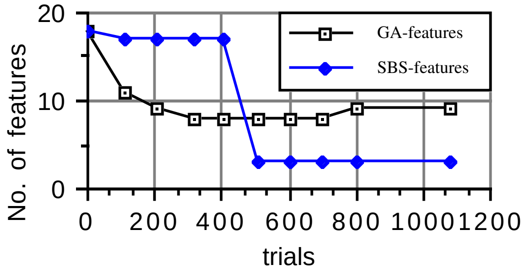

Figure 7: The number of features used by the best individual

6.0 Summary and Conclusions

The experimental results obtained indicate the potential advantages of using feature selection techniques to improve rule induction techniques. The reported results indicate that an adaptive feature selection strategy using genetic algorithms can yield a significant reduction in the number of features required for texture classification and simultaneously produce improvements in recognition rates of the rules produced by AQ15. This is a step towards the application of machine learning techniques for automating the of constructing classification systems for difficult image processing problems.

Acknowledgments

This research was done in the Artificial Intelligence Center of George Mason University. The activities of the Center are supported in part by the Defense Advanced Research Projects Agency under grants administrated by the office of Naval Research, No. N00014-87-K-0874, and No. N00014-91-J-1854, in part by the Office of Naval Research under grant s No. N00014-88-K-0226, No. N00014-88-K-0397, No. N00014-90-J-4059, and No. N00014-91-J-1351, and in part by the National Science Foundation under grant No. IRI-9020266.

References

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

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

[3] Devijver, P., and Kittler, J. "*PATTERN RECOGNITION: A STATISTICAL APPROACH,*" Prentice Hall, 1982.

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

[5] Ichino, M., and Sklansky, J.. "Optimum Feature selection by zero-one Integer Programming," *IEEE Transactions on Systems, Man, and Cybernetics*, Vol. 14, No. 5, 1984a.

[6] Michalski, R.S., Mozetic, I., Hong, J.R., and Lavrac, N.. "The Multi-purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains, *AAAI*, 1986.

[7] 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.