International Joint Conference on Artificial Intelligence (IJCAI), August 19-25, 1995, Montreal, Canada.

Hybrid Learning Using Genetic Algorithms and Decision Trees for Pattern Classification

J. Bala, J. Huang and H. Vafaie
School of Information Technology and Engineering
George Mason University
Fairfax, VA 22030

K. DeJong and H. Wechsler
Department of Computer Science
George Mason University
Fairfax, VA 22030

Abstract

This paper introduces a hybrid learning methodology that integrates genetic algorithms (GAs) and decision tree learning (ID3) in order to evolve optimal subsets of discriminatory features for robust pattern classification. A GA is used to search the space of all possible subsets of a large set of candidate discrimination features. For a given feature subset, ID3 is invoked to produce a decision tree. The classification performance of the decision tree on unseen data is used as a measure of fitness for the given feature set, which, in turn, is used by the GA to evolve better feature sets. This GA-ID3 process iterates until a feature subset is found with satisfactory classification performance. Experimental results are presented which illustrate the feasibility of our approach on difficult problems involving recognizing visual concepts in satellite and facial image data. The results also show improved classification performance and reduced description complexity when compared against standard methods for feature selection.

1 Introduction

Pattern classification, a difficult but fundamental task in AI, depends heavily on the particular choice of features used by the classifier. One usually starts with a given set of features and then attempts to derive an optimal subset of features leading to high classification performance. A standard approach involves ranking the features of a candidate feature set according to some criteria involving 2nd order statistics (ANOVA) and/or information theory based measures such as "infomax", and then deleting lower ranked features. Ranking by itself is usually not enough because the criteria used do not measure the effectiveness of the features selected on the actual classification task itself, nor do they capture possible non-linear interactions among the features.

This paper provides specific answers to the problems raised above and describes a hybrid learning approach for optimal feature selection and the derivation of robust pattern classifiers. Our novel approach, which includes a genetic algorithm (GA) and a tree induction system (ID3), minimizes the number of features used for classification while simultaneously achieving improved classifications rates. A GA is used to search the space of all possible subsets of a large set of candidate discrimination features. For a given feature subset, ID3 is invoked to produce a decision tree. The classification performance of the decision tree on unseen data is used as a measure of fitness for the given feature set, which, in turn, is used by the GA to evolve better feature sets. This GA-ID3 process iterates until a feature subset is found with satisfactory classification performance. Experimental results are presented which illustrate the feasibility of our approach on difficult problems involving recognizing visual concepts in satellite and facial image data. The results also show improved classification performance and reduced description complexity when compared against standard methods for feature selection.

2 Background

Any object or pattern that has to be recognized and/or classified must possess a number of discriminatory properties or features. The first step in any recognition process, performed either by a machine or by a human being, is to choose candidate discriminatory features and evaluate them for their usefulness. Feature selection in pattern recognition involves the derivation of salient features from the raw input data in order to reduce the amount of data used for classification and simultaneously provide enhanced discriminatory power. The number of features needed to successfully perform a given classification task depends on the discriminatory qualities of the selected features.

The selection of an appropriate set of features is one of the most difficult tasks in the design of pattern classification system. At the lowest level, the raw feature data is not nice clean symbolic data like "green", but rather noisy sensor data (e.g., spectral properties) the characteristics of which are complex and irregular. In addition, there is considerable interaction among low level features which must be identified and exploited. However, the typical number of possible features is so large as to prohibit any systematic exploration of all but a few possible interaction types (e.g., pairwise interactions). Large feature sets with noisy numerical data also provide considerable difficulty for traditional symbolic inductive learning systems. The running time of the learning system and the accuracy and complexity of the output rapidly fall below an acceptable level.

The rationale behind our approach is the belief [Michalski, 1994] that further advances in pattern analysis and classification require the integration of various learning processes in a modular fashion. Learning systems that employ several strategies can potentially offer significant advantages over single-strategy systems. Since the type of input and acquired knowledge are more flexible, such hybrid systems can be applied to a wider range of problems. Examples of such integration include combinations of genetic algorithms and neural networks [Gruau and Whitley, 1993] and genetic algorithms and rule-based systems [Bala *et al*, 1994] [Vafaie and De Jong, 1994].

The integration of genetic algorithms and inductive decision tree learning for optimal feature selection and pattern classification is a novel application of such an approach and is the topic of this paper. We have selected ID3-like induction algorithms, which use entropy as an information measure during tree derivation. This same entropy underlies also the infomax principle - maximum information preservation between successive processing layers. Self-organization in perceptual networks and the development of receptive fields has been shown to be driven by such a principle. Specifically, Linsker (1988) has reported that a perceptual system develops to recognize relevant features of its environment using the infomax principle.

The integration of genetic algorithms and decision tree learning advocated in this paper is also part of a broader issue being actively explored, namely, that evolution and learning can work synergistically [Hinton and Nowlan, 1987]. The ability to learn can be shown to ease the burden on evolution. Evolution (genotype learning) only has to get close to the goal; (phenotype) learning can then fine tune the behavior [Muhlenbein and Kinderman, 1989]. Although Darwinian theory does not allow for the inheritance of acquired characteristics (Lamarckian evolution), learning (acquired behaviors) can still influence the course of evolution. The Baldwin effect where local search is employed to change the fitness of strings, but the acquired improvements do not change the genetic encoding of the individual is under active study [Whitley *et al*, 1994]. One can gain a further perspective on the Lamarckian hypothesis by moving up from the individual chromosome (agent) to ecosystems (species) and addressing cultural evolution as well [Wechsler, 1993].

3 GA-ID3 Hybrid Learning

The basic idea of our hybrid system is to use GAs to efficiently explore the space of all possible subsets of a given feature set in order to find feature subsets which are of low order and high discriminatory power. In order to achieve this goal, we felt that fitness evaluation had to involve direct measures of size and classification performance, rather than measures such as the ranking methods discussed in the previous section. The speed of ID3 suggested the feasibility of the approach shown in Figure 1.

An initial set of features is provided together with a training set of the measured feature vectors extracted from raw data corresponding to examples of concepts for which the decision tree is to be induced. The genetic algorithm (GA) is used to explore the space of all subsets of the given feature set where preference is given to those features sets which achieve better classification performance using smaller dimensionality feature sets. Each of the selected feature subsets is evaluated (its fitness measured) by testing the decision tree produced by ID3 [Quinlan, 86]. The above process is iterated along evolutionary lines and the best feature subset found is then recommended to be used in the actual design of the pattern classification system.

In order for a GA to efficiently search such large spaces, one must give careful thought to both the representation chosen and the evaluation function. In this case, there is a very natural representation of the space of all possible subsets of a feature set, namely, a fixed-length binary string representation in which the value of the ith gene {0,1} indicates whether or not the ith feature from the overall feature set is included in the specified feature subset. Thus, each individual in a GA population consists of fixed-length binary string representing some subset of the given feature set. The advantage of this representation is that a standard and well understood GA can be used without any modification.

Each member of the current GA population represents a competing feature subset that must be evaluated to provide fitness feedback to the evolutionary process. This is achieved by invoking ID3 with the specified feature subset and a set of

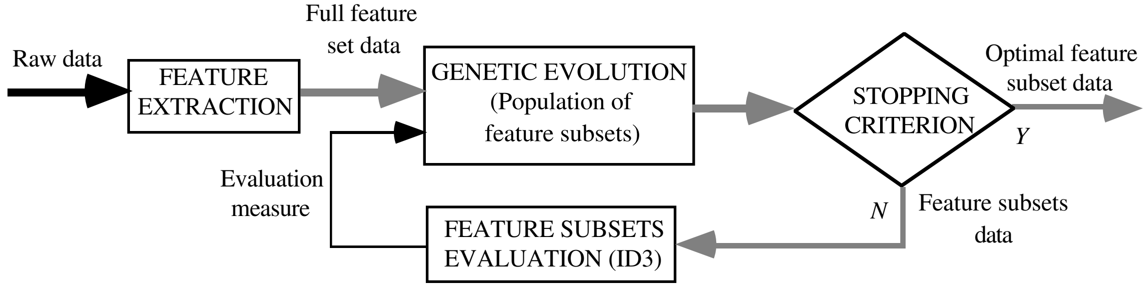


Figure 1: Hybrid Learning System Using Genetic Algorithms and Decision Trees

training data (reduced to include only the feature values of the specified features). The decision tree produced by ID3 is then tested for classification accuracy on a set of unseen evaluation data. Its accuracy together with the size of the feature subset is used as the GA fitness measure.

Our belief is that such a hybrid learning system will identify significantly better feature subsets than those produced by existing methods for two reasons. First, we are exploiting the power of GAs to efficiently explore the non-linear interactions of a given set of features. Second, by using ID3 in the evaluation loop, we have an efficient mechanism for directly measuring classification accuracy.

In order to test our ideas we have implemented a prototype version of the system. For the GA component, we used without modification GENESIS [Grefenstette, 1991], a standard GA implementation. Similarly, we used without modification C4.5, a standard implementation of ID3 [Quinlan, 1986], to build up the decision trees for the evaluation procedure. For both components, standard default parameter settings from the literature were used. For the GA module, this resulted in a constant population size of 50, a crossover rate 0.6 and a mutation rate 0.001 For C4.5, the pruning confidence level was set to default 25%.

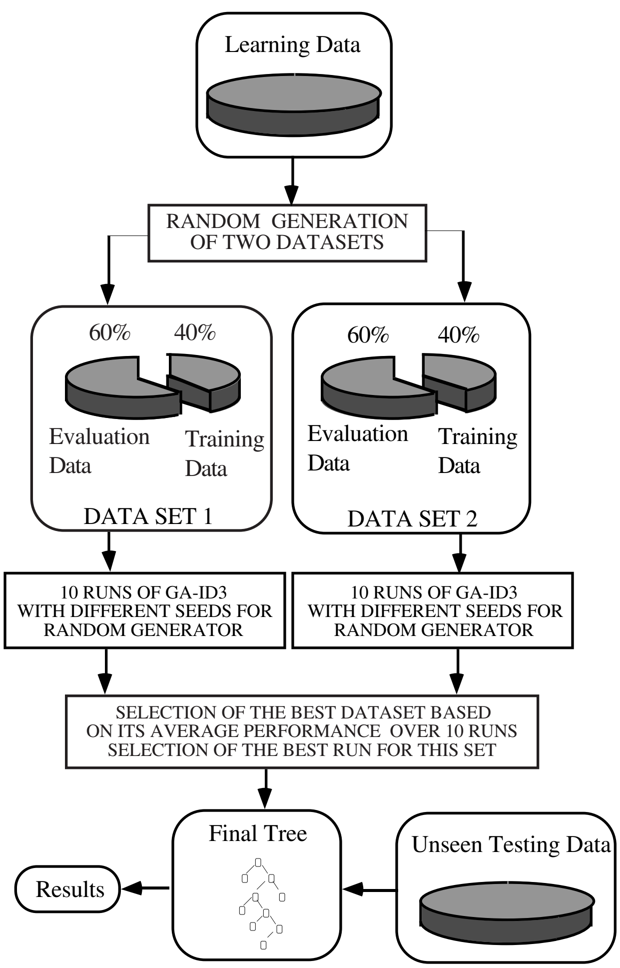


Figure 2: Setup for GA-ID3 Experiments

4 Experimental Results

An initial set of experiments has been performed to assess the performance of the hybrid GA-ID3 learning system. Subsets of optimal features for recognizing visual concepts in satellite and facial image data have been learned and compared against standard methods for feature selection. The error rates on unseen image data and the tree description complexity (measured as the number of nodes) have been used as the basis for comparison.

In order to apply cross validation of the learning process, the learning data was randomly shuffled to generate two sets (Figure 2). In each set 40% of examples were used for inducing decision trees and the other 60% for evaluation of the learned description. Ten experiments were performed on each set in order to find the best average performance. The set that produced the best result, i.e. the lowest error rate, was selected as the training set. The tree generated during the best run of that set (one of ten runs) is applied to the unseen test data.

Results obtained by the GA-ID3 system have been compared with two other sets of results. The first one was obtained by using all features (36 for the satellite date and 105 for facial data). To generate the second result a set with features reduced to the same number as the one produced by the GA-ID3 experiment was used. This reduction was achieved by an independent ranking of each feature using an information theory based entropy measure (infomax) to estimate which features are the most discriminatory. Features that lead to the greatest reduction in the estimated measure of the training examples are chosen. The exact criterion is to choose that feature vector X with values $\{xv_1, xv_2, ..., xv_m\}$ that minimizes the expression

over n classes, where x_i^+ is the number of examples in a given class with values xv_i, and x_i^- is the number of negative examples (all the examples not belonging to this class) with the value xv_i.

4.1 Experiments with the Satellite Data

The satellite image database consists of the multi-spectral values of pixels in 3x3 neighborhoods, and the classification associated with the central pixel in each neighborhood. The aim is to predict this classification, given the multi-spectral values. In the sample database used in experiments, the class of a pixel is coded as a number.

One frame of Landsat MSS imagery consists of four digital images of the same scene in different spectral bands. Two of these are the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each pixel is a 8-bit binary word, with 0 corresponding to black and 255 to white. The spatial resolution of a pixel is about 80m x 80m. Each image contains 2340 x 3380 such pixels.

The original data for the database used in our experiments was generated from data purchased from NASA by the Australian Center for Remote Sensing. One frame of imagery consists of four digital images of the same scene in different spectral bands. Two of these are in the visible region (corresponding approximately to green and red regions of the visible spectrum) and two are in the (near) infra-red. Each pixel is a 8-bit binary word, with 0 corresponding to black and 255 to white. Each example of data corresponds to a 3x3 square neighborhood of pixels completely contained within the 82x100 sub-area and it contains the pixel values in the four spectral bands (converted to ASCII) of each of the 9 pixels in the 3x3 neighborhood and a number indicating the classification label of the central pixel. Tables 1 and 2 characterize the data. Figure 3 shows average performances on the evaluation sets over 10 runs using the GA-ID3 system. Table 3 shows the results of the GA-ID3 experiment together with the corresponding performances for (i) the set with all the features and (ii) the set with features reduced to the number obtained in the GA-ID3 experiment and ranked using the (entropy) infomax measure.

CHARACTERISTIC	DESCRIPTION
NUMBER OF EXAMPLES	learning set of 4435 examples and test set of 2000 examples
NUMBER OF FEATURES	4 spectral bands * 9 pixels in neighborhood = 36 features
FEATURE VALUES	The feature value is numerical, in the range 0 to 255
NUMBER OF CLASSES	There are 6 decision classes: 1 to 6

Table 1: Characteristics of the Satellite Data

CLASS NAME	LEARNING SET	TESTING SET
Red soil	1072 (24.17%)	461 (23.05%)
Cotton crop	479 (10.80%)	224 (11.20%)
Gray soil	961 (21.67%)	397 (19.85%)
Damp gray soil	415 (09.36%)	211 (10.55%)
Soil with vegetation stubble	470 (10.60%)	237 (11.85%)
Very damp gray soil	1038 (23.40%)	470 (23.50%)

Table 2: Number of Examples in the Learning and Test Sets for Satellite Data

FULL FEAURE SET *36 Features*		REDUCED FEATURE SET BY THE GA-ID3 SYSTEM *17 Features*		REDUCED FEATURE SET BY FEATURE RANKING *17 Best Features Chosen*	
Error Rate	Tree Complexity	Error Rate	Tree Complexity	Error Rate	Tree Complexity
18.5%	99 nodes	16.9 %	72 nodes	20.6%	83 nodes

Table 3: Experiment Results for Satellite Data

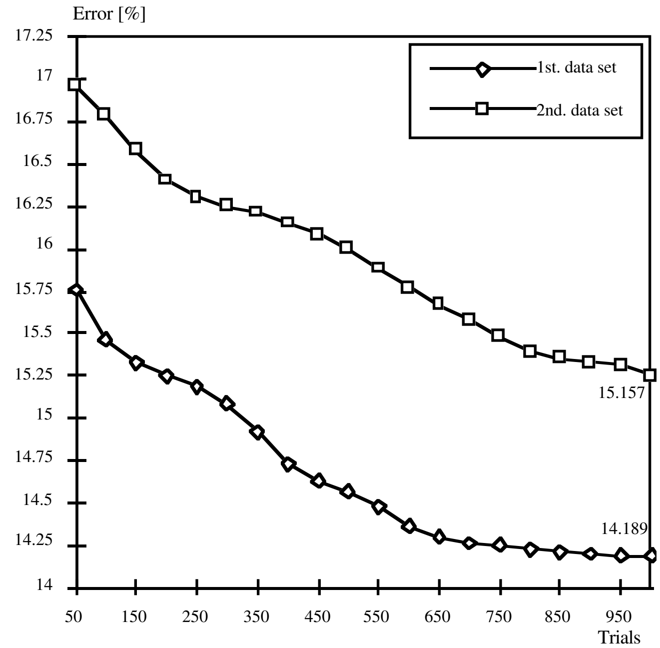


Figure 3: Average Error Rate for 10 Runs of Satellite Data Sets 1, 2

4.2 Experiments with Face Data

The ability to detect salient facial features is an important component of any face recognition system. Among the many facial features available it appears that the eyes play the most important role in both face recognition and social interaction.

Detecting the eyes serves first of all an important role in face normalization and thus facilitates further localization of facial landmarks. It is eye detection that allows one to focus attention on salient facial configurations, to filter out structural noise, and to achieve eventual face recognition. For these purposes, we applied the hybrid learning system to accomplish feature selection on eye detection problem.

102 eye, 102 nose and 102 other facial region examples are made available as learning and 52 eyes, 52 nose and 52 other facial region examples are used as test data in our experiments. The original eye image, nose image or the image of other regions has resolution 16 by 12 (column by row) pixels, which is cut from human face images (64x72 pixels). 105 features are produced for each example. The configuration of those features follows the rules listed in Table 4, while Table 5 gives the characteristics of the face data.

Figure 4 shows average performance on the evaluation sets over 10 runs using the GA-ID3 system. Table 6 shows the results of the GA-ID3 experiment together with the corresponding performance for (i) the set with all the features and (ii) the set with features reduced to the number obtained in the GA-ID3 experiment and ranked using the (entropy) infomax measure.

Figure 5 represents graphically a comparison of various results obtained in experiments. Both for the satellite and face data an improvement of recognition rate (lower error rate) has been observed for sets with features reduced by the GA-ID3 system. Our method has also reduced tree complexity for the satellite data. The reduction of complexity for the face data has not been observed. However, all generated trees for the facial data are fairly simple (about 10 nodes). The number of features was reduced by 60% for the facial data and by 52% for the satellite data.

FEATURES	DESCRIPTION
From x_1 to x_{35}	Obtained by averaging a 4x4 window which has 50% (2 pixels) overlap in each shift.
From x_{36} to x_{70}	They represent the standard deviation of a 4x4 window.
From x_{71} to x_{105}	They represent the entropy for a 4x4 window.

Table 4: Configuration of 105 features

CHARACTERISTIC	DESCRIPTION
NUMBER OF EXAMPLES	Learning set of 102 examples and test set of 52 examples.
NUMBER OF FEATURES	105 features.
FEATURE VALUES	The feature value is numerical, in the range 0 to 255.
NUMBER OF CLASSES	There are 3 decision classes: 1 to 3.

Table 5: Characteristics of the Facial Data

FULL FEATURE SET *105 Features*		**REDUCED FEATURE SET BY THE GA-ID3 SYSTEM** *41 Features*		**REDUCED FEATURE SET BY FEATURE RANKING** *41 Best Features Chosen*	
Error Rate	Tree Complexity	Error Rate	Tree Complexity	Error Rate	Tree Complexity
38.4%	8 nodes	27.5%	13 nodes	38.5%	11 nodes

Table 6: Experimental Results for Face Data

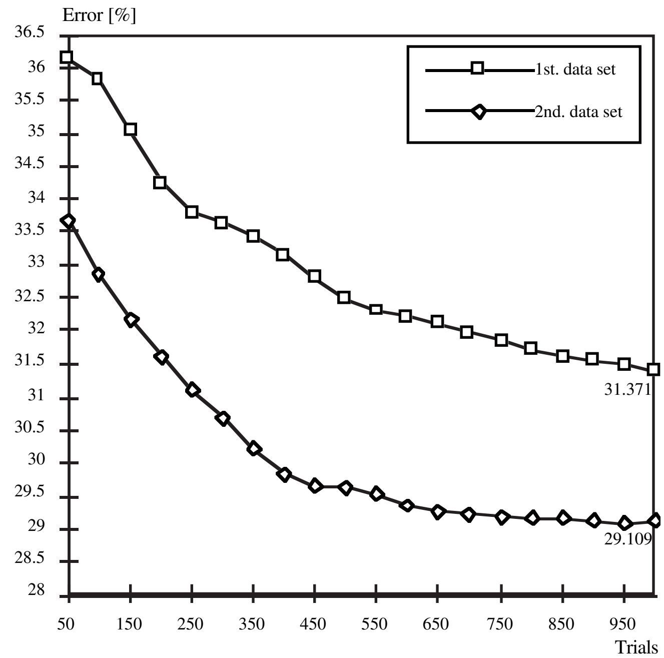
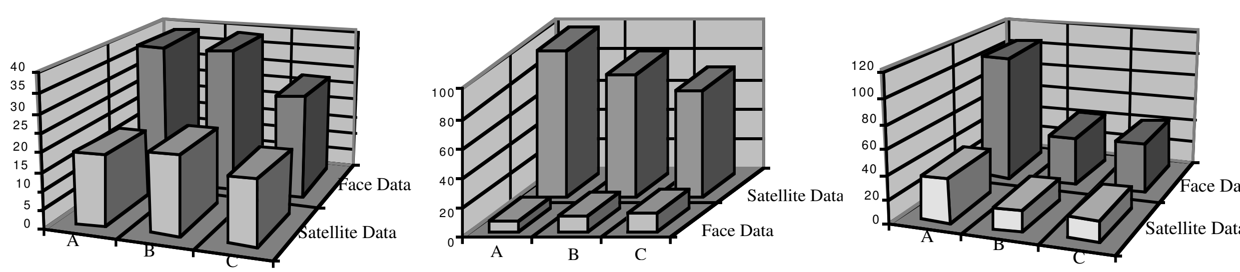


Figure 4: Average Error Rate for 10 Runs of Face Data Sets 1, 2



Recognition Error Rates in Percentages

Description Complexities Expressed as the Number of Nodes

Number of Features in Sets

A: Full Feature Set
B: Reduced by Independent Ranking
C: Reduced by the GA-ID3 System

Figure 5: Comparison of Results

5 Conclusions

This paper introduced a hybrid learning methodology that integrates genetic algorithms (GAs) and decision tree learning (ID3) for evolving optimal subsets of discriminatory features for robust pattern classification. Experimental results have been presented which illustrate the feasibility of our approach on difficult problems involving recognizing visual concepts in satellite and facial image data. The results have also shown significant improvements in classification performance and reduced description complexity when compared against standard methods for feature selection.

Clearly more work needs to be done. Although these two data sets are quite complex in comparison with symbolic machine learning data sets, they are still modest from an image processing point of view. We are currently involved in refining the system described here as we test it on larger and more complex problems.

An interesting extension to be explored is the possibility of additional feedback from ID3 concerning the evaluation of a feature set. Currently only classification accuracy is returned. However, there is potentially exploitable information with respect to which features were actually used to build the decision tree and their relative positions in the tree.

Acknowledgments

The Army Research Laboratory (ARL) has partially supported J. Huang and H. Wechsler under the DAAL01-93-R-9286 grant on 'Face Recognition'. We would like to thank Eric Bloedorn and Ali Hadjarian for useful comments on experiments, and to Bob Henery for explanations on satellite data.

References

[Bala *et al*, 1996] J. Bala, P. Pachowicz, and K. De Jong, *Multistrategy Learning from Engineering Data by Integrating Inductive Generalization and Genetic Algorithms*, in Machine Learning: A Multistrategy Approach, Vol. IV, R.S. Michalski and G. Tecuci (Eds.), Morgan Kaufmann, San Mateo, CA., pp. 121-138, 1994.

[Gruau and Whitley, 1993] F. Gruau and D. Whitley, *Adding Learning to the Cellular Development of Neural Networks: Evolution and the Baldwin Effect,* Evolutionary Computation, Vol.1, No.3, pp. 213-234, 1993.

[Grefenstette *et al*, 1991] J. Grefenstette, L. David and D. Cerys, *Genesis and OOGA: Two Genetic Algorithms System*, TSP: Melorse, MA, 1991.

[Hinton and Nowlan, 1987] G. E Hinton and S. J. Nolan, *How Learning Can Guide Evolution,* Complex Systems, Vol.1, pp. 495-502, 1987.

[Linsker, 1988] R. Linsker, *Self-Organization in a Perceptual Network*, Computer, Vol. 21, No. 3, pp. 105-117, 1988.

[Michalski, 1994] R. Michalski, *Inferential Theory of Learning: Developing Foundations for Multistrategy Learning,* in Machine Learning: A Multistrategy Approach, Vol. IV, R.S. Michalski and G. Tecuci (Eds.), Morgan Kaufmann, San Mateo, CA., pp. 3-61, 1994.

[Muhlenbein and Kinderman, 1989] H. Muhlenbein, and J. Kinderman, *The dynamics of Evolution and Learning. Toward Genetic Neural Networks*, in R. Pfeifer, Z. Schreter, F. Fogelman-Soulie, and L. Steels, Connectionism in Perspective, Elsevier Science, pp. 173-197, 1989.

[Quinlan, 1986] J. R. Quinlan, *The Effect of Noise on Concept Learning,* in Machine Learning: an Artificial Intelligence Approach, R.S. Michalski, J.G. Carbonell and T.M. Mitchell (Eds.), Morgan Kaufmann publishers, San Mateo, CA, pp. 149-166, 1986.

[Vafaie and De Jong, 1994] H. Vafaie and K. De Jong, *Improving 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., pp. 453-469, 1994.

[Wechsler, 1993] H. Wechsler, *A Perspective on Evolution and the Lamarckian Hypothesis Using artificial Worlds and Genetic Algorithms*, Revue Internationale de Systemique, Vol. 7, No. 5, pp. 573-592, 1993.

[Whitley *et al*, 1994] D. Whitley, V. S. Gordon, and K. Mathias, *Lamarckian Evolution, the Baldwin Effect, and Function Optimization*, in Y. Davidor, H.P. Schwefel, and R. Manner (Eds.), Parallel Problem Solving from Nature - PPSN III, Springer Verlag, pp. 6-15, 1994.