Inductive Genetic Programming with Decision Trees

Nikolay I. Nikolaev[1] and Vanio Slavov[2]

[1] Department of Computer Science, American University in Bulgaria, Blagoevgrad 2700, Bulgaria, e-mail: nikolaev@nws.aubg.bg

[2] Information Technologies Lab, New Bulgarian University, Sofia 1113, Bulgaria, e-mail: vslavov@inf.nbu.acad.bg

Abstract. This paper proposes an empirical study of inductive Genetic Programming with Decision Trees. An approach to development of fitness functions for efficient navigation of the search process is presented. It relies on analysis of the fitness landscape structure and suggests measuring its characteristics with statistical correlations. We demonstrate that this approach increases the global landscape correlation, and thus leads to mitigation of the search difficulties. Another claim is that the elaborated fitness functions help to produce decision trees with low syntactic complexity and high predictive accuracy.

1 Introduction

Inductive concept learning is considered search for concept descriptions that model accurately given data. The inductive learning has to construct concept descriptions which: first, cover most of the data; second, feature by high extrapolation power; and, third, have reasonable length. Nowadays, stochastic complexity (minimum description length) measures [6],[7],[8] are used for achieving these qualities. The stochastic complexity measures keep a balance between the syntactic complexity and predictive accuracy of the inferred descriptions.

Evolutionary algorithms [1],[5] are random search methodology that could be employed for inductive learning. They build concept descriptions with reproduction, recombination and mutation operators. The evolutionary search could be viewed as navigation by the genetic operators on their landscapes determined by a fitness function. The fitness landscape structure is the only information for search navigation. The evolutionary algorithms perform well on fitness landscapes that contain relevant information about the search goal [4]. On such landscapes they can find global, or nearly-global solutions, reliably and quickly.

This research was motivated by the fact that the navigation difficulties are influenced by the fitness function, as they obviously depend on how the landscape has been created. An approach to developing fitness functions for efficient search upon a global fitness landscape analysis is presented. The approach helps to make fitness functions that provide enough information for locating the global optima on the landscape. The fitness distance correlation [4] measure is used for estimating the correspondence between the fitness of a point and its distance to a global optimum in the search space of the task.

The approach is illustrated with elaboration of a fitness function for efficient concept learning by inductive Genetic Programming with Decision Trees (GPDT) [9]. The claim is that a careful design of a stochastic complexity fitness function helps to achieve: first, mitigating the navigation difficulties by increasing the global landscape correlatation, and, second, keeping a balance between the parsimony and accuracy of the decision trees. The experiments show that the inductive GPDT is really useful when it continuously promotes genotypes with low syntactic complexity and high predictive accuracy.

This paper presents GPDT in section two. It includes the genetic decision tree-like programs, the reproduction, recombination and mutation operators, and the fitness function. The study of GPDT with a special statistical measure is given in section three. Finally, conclusions are derived.

2 Genetic Programming with Decision Trees

2.1 Genetic Decision Tree-like Programs

The Genetic Programming (GP) paradigm can be used for solving inductive concept learning tasks [3],[5] by manipulating a population of decision trees [7]. The search could be organized by modifying the size and shape of decision trees with recombination and mutation operators. The GP is appropriate for inductive learning as it allows the size of the trees to be discovered automatically.

The presented GPDT evolves genetic programs in the form of decision trees [7]. In context of the learning task and in terms of the chosen description language, the nodes of the decision tree are attributes of the concept features, and the leaves denote the class of the concept. Since a decision tree can be viewed as a representation of a composition of functions, it is easy to determine how the tree components serve as genetic material: the concept attributes could naturally be functional nodes, and the concept classes could be terminal leaves.

2.2 Reproduction, Recombination and Mutation Operators

GPDT breeds a population of decision trees with three genetic operators: reproduction, recombination and mutation. Currently, steady state reproduction is employed with fitness proportionate selection [1].

The recombination operator for GPDT performs crossover [5] by cutting and splicing two parent trees in randomly chosen nodes. The crossover operator maintains the closure property [5] and derives only syntactically correct decision trees. That is why the cross points in the parent trees are selected so that only offsprings having nonrepeating attributes on each branch can result.

An uniform replacement mutation operator has been developed especially for GPDT. The operator traverses a decision tree in depth-first manner and changes each visited node or leaf with a probability $Pm = x/length(DT)$, where: x is a parameter. When a functional node is encountered, it is replaced with equal probability by another randomly chosen functional or terminal node. A terminal

is replaced again with equal probability by a randomly chosen functional or terminal. This uniform mutation also preserves the closure property [6]. When a functional node is to be replaced with a functional node, the new one is chosen so that it does not appear in the above and below contexts.

2.3 The Stochastic Complexity Fitness Function

The purpose of inductive learners is to identify concepts that best model given examples. The stochastic complexity measures [6],[7],[8] provide sound criteria for data modeling. In terms of decision trees, these are criteria for isolation of parsimonious trees, with minimal syntactic complexity, and accurate trees, which model well the examples. A balance between accuracy and parsimony is very important for efficient evolutionary search [3]. We clarify this revelation is sense that the use of ready formulae does not necessarily lead to better results.

Experiments into inductive GPDT were carried out with the recent measure of Quinlan [8]. This measure [8] was originally prepared for pruning decision trees, but it is reasonable to employ it for growing decision trees also:

$$F(DT) = min\{I(DT) + I(e|DT)\}$$

$$I(DT) = n_f + n_l + n_f \times log_2(f) + n_l \times log_2(l)$$

$$I(e|DT) = tp \times (-log_2(P(e|tp))) + fp \times (-log_2(1 - P(e|tp))) + tn \times (-log_2(P(e|tn))) + fn \times (-log_2(1 - P(e|tn)))$$

It has been found, however, that during evolutionary learning distinct decision trees have been assigned equal stochastic complexity values. That is why, we modify the measure so that the relative frequencies of the examples are computed with the following conditional probabilities [9]:

$$P(e|tp) = P(e|tp_1 + tp_2 + ... + tp_{N_p}) = \sum_{i=1}^{N_p} P(tp_i) \times P(e|tp_i)$$

$$P(e|tn) = P(e|tn_1 + tn_2 + ... + tn_{N_n}) = \sum_{i=1}^{N_n} P(tn_i) \times P(e|tn_i)$$

where: n_f-functional nodes in DT; n_l-leaves in DT; f-possible functions; l- leaf classes; N_p- positive leaves; N_n- negative leaves; tp-true positive examples; fp-false positive examples; tn-true negative examples; fn-false negative examples.

This calculation of the relative frequencies of the examples with conditional probabilities has the advantage that it accurately evaluates a decision tree as a disjoint combination of conjunctive components formed along the branches.

3 Studies of Inductive GPDT

This research investigates the different fitness landscape structures that arise from different fitness functions employed in inductive GPDT. Such analysis is valuable as the navigation difficulties are monitored by the global structure of the fitness landscape, which is a global search characteristics of the task.

3.1 Fitness Distance Correlation

The fitness distance correlation (FDC) measure [5] provides evidences for the global correlation character of the fitness landscape:

$$FDC(F,D) = \frac{\sum_{i=0}^{M}(F_i-\bar{F})\times(D_i-\bar{D})}{\sqrt{\sum_{i=0}^{M}(F_i-\bar{F})^2}\times\sqrt{\sum_{i=0}^{M}(D_i-\bar{D})^2}}$$

where: M - is the number of the steps considered, and $M > 0$; F_i- is the fitness at step i ; D_i - is the distance at step i ; $\bar{F}= (1/(M+1)) \times \sum_{0}^{M} F_i$, and $\bar{D}= (1/(M+1))\times \sum_{0}^{M} D_i$ are weighted means.

The FDC is calculated with pairs (F, D) recorded during random walks on the landscape. The distance $D = Dist(DT, O)$ is defined as the number of one-point mutations needed to produce the optimal decision tree O from a particular decision tree DT. Since the stochastic complexity fitness function has a minimizing effect, FDC is 1 when the correlation is maximal.

3.2 The Fitness Landscape

The fitness landscape here is a methaphore that relates decision tree-like genotypes with their fitnesses, calculated with the stochastic complexity function. Such a view has the advantage that the fitness landscape could be reliably examined as special statistical measures for estimation of its characteristics exist. The landscape consists of points with fitnesses. The differences between the fitnesses of these points form the correlation structure of the landscape. The differences between the decision trees could be precisely identified as their fitnesses are points on different hills, valleys, and slopes of the correlation structure.

The FDC measure summarizes whether the global optima is accessible from a given point on the landscape. The distance to the global optimum should decrease with improving the fitness. Abstractly, this means that the global optima is visible and therefore the landscape is mountable.

3.3 Experimental Results

The elaborated stochastic complexity formula was built in GPDT as a fitness function. Two kinds of experiments were conducted: first, with groups of uniformly generated examples, and, second, with benchmark datasets.

The landscape analysis with FDC started with the measure of Quinlan [8]. Two groups each of 5 sets of examples for different decision trees with up to 6 attributes have been generated. The domains for the attributes consisted of 2 to 4 discrete values. The sampling decision trees were symmetric and anisymmetric. The first group included 5 symmetric trees, which left subtree mirrors the right subtree. The second group included 5 anti-symmetric trees with different depths: shallow, medium, and deep. The parameter x of the probability Pm was: $x = 0.05$. Walks on 1500 landscape points have been carried out.

Symmetric Decision Trees The results from the tests on each of the 5 sets are visualized by scatter plots and abbreviated FDC summaries. Figure 1 presents the scatter plot for the fitness/distance relationship derived from the 5^{th} symmetric set (the remaining four scatter plots almost coincid with it). It reveals that the fitnesses and the distances produced with the measure of Quinlan are uncorrelated. The fitness values vary within a small range nevertheless the distances to the globally optimal tree increase or decrease. Moreover, this has been noted for inductive tasks which are not very difficult. For all the 5 tasks the averaged FDC values from the measure of Quinlan are in $[-0.2 \div -0.025]$.

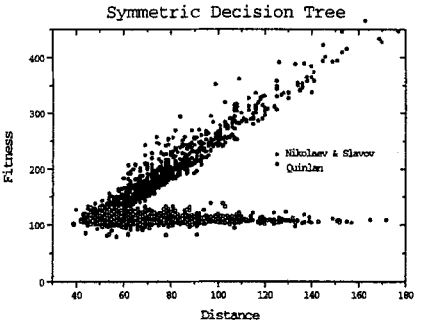

fig.1

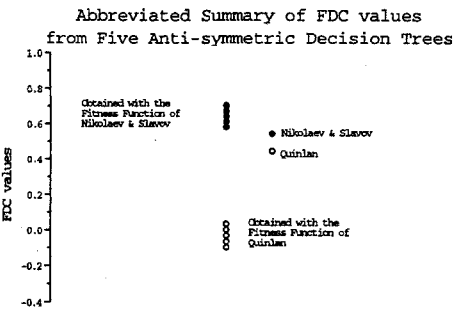

fig.2

Anti-symmetric Decision Trees The global landscape is obviously uncorrelated as the FDC values were near 0. Another series of tests with anti-symmetric decision trees have been conducted using the same parameters. Figure 3, with the scatter plots of FDC derived from the 10^{th} example set, shows that there is no clear correlation between the fitnesses and their distances to the optima. When evaluated with the formula of Quinlan, the points on the landscape are within a tight band parallel to the horizontal axis. The formula failed to evaluate differently trees, which are at different distances to the global optima.

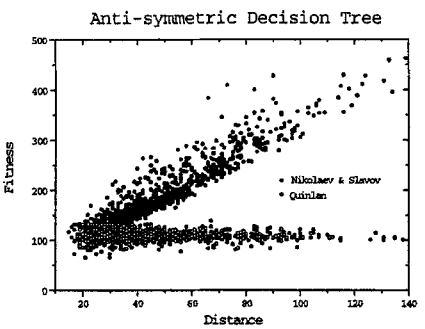

fig.3

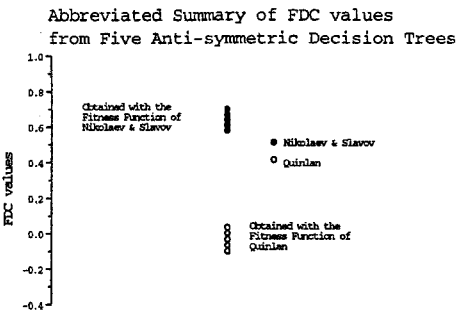

fig.4

The abbreviated summary of averaged FDC from 10 trials with the sets for anti-symmetric trees is given on Figure 4. The FDC values were near 0, this time higher within the interval $[0.035 \div -0.125]$, which says that there is virtually no correlation between the fitness and the distance to the global optima. In other words, the problems were difficult for the measure of Quinlan, which is against the common sense as they have been deliberately generated as easy. After a careful examination it was found that when evaluated with this fitness function some offspring decision trees have the same fitness as their parents.

Improving the Fitness Function Our objective was to make the formula to account for the number of the branches in a decision tree. The idea is to calculate the frequencies of the true positive and true negative examples with a conditional probability estimate that inherently measures how many concepts, acquired at the leaves, participate in the disjunctive concept description. The conditional probability estimate say that the smaller the branches that generate conjuncts, the better the compression of the examples. Hence, the stochastic complexity fitness function will assign lower values to more compact decision trees as they better compress the examples.

Experiments with the Repaired Fitness Function The experiments have been repeated using the modified stochastic complexity fitness function. The FDC scatter plots are given on figures 1 and 3. They reveal that the improved formula maintains a high fitness/distance relation. This is proven by the distribution of almost all points, derived with the novel formula, around a line that is at 45 degrees between the horizontal and the vertical axes.

The abbreviated summaries with average FDC values from 10 trials are given on figures 2 and 4. While the FDC values derived with the formula of Quinlan are within the range $[-0.2 \div 0.035]$, those derived with the improved formula are in the interval $[0.57 \div 0.845]$. That is why, the FDC values from Figure 2 and Figure 4 allow to conclude that the landscape has been tuned from misleading to straightforward, according to the original definitions [5]. Thus, the repair of the formula made the global fitness landscape more correlated.

Experiments with Classical Benchmark Datasets The improved stochastic complexity function was tested also by running GPDT on two classical benchmark datasets : the Iris data [7]; and the Multiplexor-11 data [7]. The accuracy and complexity of the evolved decision trees were compared with the ones produced by C4.5 [7] after learning from exactly the same data sets and testing on exactly the same testing sets. We used the pruned decision trees produced by C4.5 with confidence level $CF = 25\%$.

The Iris data set consists of 150 examples from 3 classes. For training 100 examples were used. They include 4 continuous attributes, which we discretized in advance. The parameters were: $PopulationSize = 50$, $Generations = 100$, $P_{Crossover}$=0.8, and $P_{Mutation}$=0.2. Tests with the formula of Quinlan and the improved one, using values of x: 0.5, 1, 1.5, 2, and 2.5, were conducted. On Figure 5 the average fitnesses from 10 runs with the Iris data are plotted. The best fitnesses, visualized on Figure 6, are obtained only with the improved formula and values of x: 0.5, 0.8, 1, 1.2, 1.5, 2, 2.5, 2.95, 3, and 3.5.

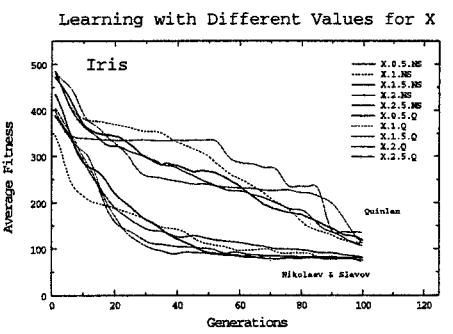

fig.5

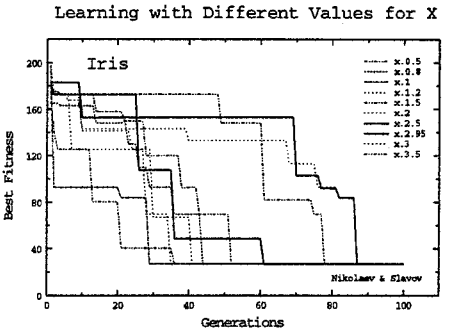

fig.6

The plots on Figure 5 reveal that the evolutionary learner with the improved formula solves the Iris task more accurately. As evidences the more stable curves serve. With the Quinlan's formula the evolution was periodically misleaded. Figure 6 shows that all of the runs with the improved fitness formula converged to the same decision tree. The size of the tree, to which all the runs of GPDT converged, was 4, exactly the same as the decision tree generated by C4.5. The classification accuracy on unseen test data was also the same 98.7%. Note that this is an indication for a slightly overpruned decision tree, but this is because of the simplistic manual discretization of the continuous attributes.

Figures 7 and 8 represent results derived when learning concepts from the Multiplexor-11 [7] data set. They describe an 11-bit multiplexor with 3 attributes for the address bits, and 8 for the data bits. There were randomly generated 100 examples for training and 1000 for testing. The decision tree produced by C4.5 for the training set was with length 39, and the decision tree produced by GPDT with length 7. The accuracy of the tree of C4.5 on unseen data was 71.63% and the accuracy of the GPDT tree was 81.54%. The difficulty of this inductive task can be understood from the small band in which the curves on Figure 7 appear. The band of curves produced with the formula of Quinlan is more rugged and implies unstable evolutionary search.

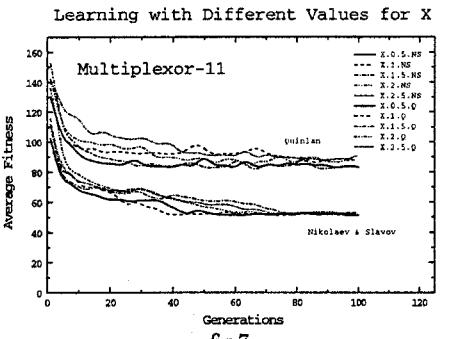

fig.7

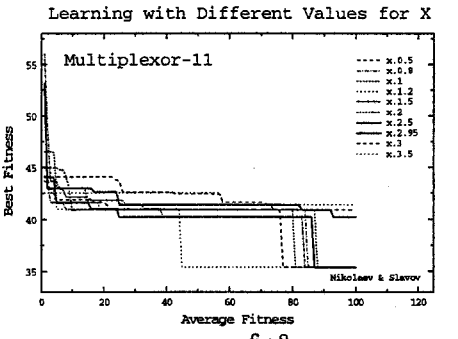

fig.8

4 Discussion

The presented study had two aspects. First, it demonstrated an approach to elaboration of fitness functions for evolutionary learning, but the approach could be used for improving the heuristic functions for symbolic learners also.

Second, the study demonstrated that evolutionary concept learners can be successfully applied for solving inductive tasks. They can produce concept descriptions with accuracies that could be compared favorably with the results generated by some of the best symbolic learners. The evolutionary learning process, however, is critically influenced by the selection of parameter values for the genetic operators. Note that these aspects were exploited with a system which uses decision tree-like genotypes, but they can help also to improve GP systems with other genetic program representations [3].

5 Conclusion

This paper continues the study of the navigation and the structure of the search carried out by inductive learners. Statistical correlations were used to measure the quality of the solutions in order to make the search more efficient. The contribution of the paper is twofold: first, it developed a general stochastic complexity formula for eastimating decision trees appropriate for incremental and batch learners; and, second, a relationship between the performance and the representation in inductive evolutionary learners has been found.

References

1. J. Holland, *Adaptation in Natural and Artificial Systems*, A Bradford Book, The MIT Press, Cambridge, MA, 1992.
2. J. Horn and D.E. Goldberg, 'Genetic Algorithm Difficulty and the Modality of Fitness Landscapes', in *Foundations of Genetic Algorithms*, L.D.Whitley and M.D.Vose (eds.), Morgan Kaufmann Publ., San Mateo, CA, 243-269, 1995.
3. H. Iba, H. de Garis and T. Sato, 'Genetic Programming using a Minimum Description Length Principle', in *Advances in Genetic Programming*, K.Kinnear (ed.), The MIT Press, 265-284, 1994.
4. T. C. Jones and S. Forrest, 'Fitness Distance Correlation as a Measure of Search Difficulty for Genetic Algorithms', in *Proc. Sixth Int. Conference on Genetic Algorithms*, L.Eshelman (ed.), 184-192, 1995.
5. J. R. Koza, *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, The MIT Press, Cambridge, MA, 1992.
6. M. Li and P. Vitanyi, *An Introduction to Kolmogorov Complexity and Its Applications, Springer-Verlag*, NY, 1993.
7. J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publ., San Mateo, CA, 1993.
8. J. R. Quinlan, 'MDL and Categorical Theories (Continued) ', in *Proc. Int. Conference on Machine Learning, ICML-95*, Tahoe City, CA, 1995.
9. V. Slavov and N. Nikolaev, 'Fitness Landscapes and Inductive Genetic Programming', in *Proc. Third Int. Conf. on Artificial Neural Networks and Genetic Algorithms, ICANNGA'97*, Springer, Vienna, 1997.