Information Processing Letters 46 (1993) 163–168
Elsevier

25 June 1993

Genetic algorithm for feature selection for parallel classifiers

Ludmila Kuncheva

Central Laboratory of Bioinstrumentation and Automation, Bulgarian Academy of Sciences, Acad. G. Bonchev Street, Bl. 105, 1113 Sofia, Bulgaria

Communicated by L. Boasson
Received 17 April 1992
Revised 26 April 1993

Abstract

Kuncheva, L., Genetic algorithm for feature selection for parallel classifiers, Information Processing Letters 46 (1993) 163–168.

A way to select a combination of feature subsets serving as inputs for a parallel classifier is described. A genetic algorithm with a properly modified fitness function is used. Experimental results with three sets of real data from internal, neonatal and aviation medicine are reported.

Keywords: Parallel algorithms; pattern recognition; genetic algorithm; combinatorial problems

1. Feature selection for parallel classifier

Parallel classification is a pattern recognition paradigm which is a topic of renewed interest due to the increased possibilities for parallel computation. Neural network classifiers are a persuasive example of this tendency [10,11].

Here a parallel classification scheme is considered [7,9] which consists of r first-level decision makers (operating on different subsets of features) and a second-level "aggregator" of their classification decisions. The complication of the structure aims rather at higher classification accuracy than at facilitating the computation process. It can be formally proven that the voting scheme of independent classifiers outperforms the best one of them.

Let X be the set of n features describing the objects. The feature selection task for a conventional (one level) classifier consists in choosing the best subset of X in terms of a certain classification criterion.

The main difficulty in feature selection is that the searching surface is usually multimodal, and the criterion can hardly be described analytically. This fact precludes the application of the classical searching tools based on criterion derivatives. It is widely accepted that, in spite of the advanced computational means and technologies, the exhaustive search amongst all subsets of features is an extremely hard combinatorial problem. Different suboptimal algorithms have been developed [3,16] which help to avoid the exhaustive search and which are supposed to lead to a suboptimal decision.

The problem of feature selection becomes even more cumbersome in the parallel classification scheme because a set of subsets $S=\{S_1,\dots,S_r\}$, $S_i \subseteq X$, is to be chosen instead of a single subset. Let $F(X)$ be the set of all subsets of X. The

Correspondence to: L. Kuncheva, Central Laboratory of Bioinstrumentation and Automation, Bulgarian Academy of Sciences, Acad. G. Bonchev Street, Bl. 105, 1113 Sofia, Bulgaria.

exhaustive search in $F(X)$ for choosing the best single subset requires 2^n calculations of the criterion value while the same operation for choosing the best combination of subsets implies 2^{2^n} calculations. The conspicuous impossibility to perform this search in real problems determines the accent of the paper: to formulate a feature selection technique for parallel classifier.

2. Genetic algorithm for feature selection

The attractive properties of genetic algorithms are recently being rediscovered in connection with increased technical capabilities for parallel computation [5]. Along with simulated annealing, neural networks, tabu search, etc., genetic algorithms are attributed to the wide palette of heuristic searching tools [4]. They mimic the mechanisms of natural reproduction processes. Possessing a variety of assets, genetic algorithms have demonstrated good performance in different optimization fields [1,2,12,14,15]. In contrast to many trivial optimization techniques, genetic algorithms can operate on discontinuous, noisy, multidimensional and multimodal surfaces.

Each point in the search space is represented as a string (chromosome). A necessary component of a genetic algorithm is the way of encoding solutions on chromosomes. It is pointed out in [2] that although any alphabet is acceptable, there is some evidence that the binary representation (alphabet {0, 1}) is in a certain sense optimal. The search among feature subsets naturally fits this framework. In this case the chromosome directly expresses a feature subset S_k as a bit vector. Value "0" at the ith position denotes the absence of the ith feature from the current subset and value "1" its presence, respectively.

A "fit" criterion $J(S_k)$ is defined to evaluate each chromosome. In the case of feature selection this may be any measure of discrimination power of the respective subset.

The algorithm works on a set of chromosomes simultaneously (population set $\Pi = \{S_1, \dots, S_r\}$) imitating the stages of evolution. A simple genetic algorithm, which has been reported to perform well in a variety of circumstances, is described below with special emphasis on its implementation characteristics for feature selection. The template from [15] is applied here:

1. *Formulation of the initial population set* $\Pi = \{S_1, \dots, S_p\}$.
2. *Formulation of mating set* M (the set of potential parents). The principle of the biased roulette wheel [5] is used on the subset of the population whose criterion value exceeds the average. A chromosome is put into the mating pool as much times as is the ratio of its fitness to the average one calculated over the whole population. The parents are then picked up at random from the mating set.
3. *Crossing-over of parent's chromosomes*. Each randomly selected couple of parents produces two offspring chromosomes with a probability P_c by exchanging parts of their strings. In this way the set of offsprings $O = \{S_{p+1}, \dots, S_{2p}\}$ is formed.
4. *Mutation*. Each bit in the offspring's vector switches to the opposite value with previously defined mutation probability.
5. *Combination*. A new population set is constituted. There are a number of strategies for this procedure. A "generation gap" is used in [6] which controls the percentage of the population to be replaced during each generation. The model of surviving accepted here is referred to as "elitist" model [1,6] due to which the best individuals are preserved from generation to generation. Formally, the new population set is constructed by
$$\Pi = \{S_k \mid S_k \in \Pi \cup O,\ J(S_k) > J(S_l),\ k \in I,\ \forall l \notin I\},$$
where I is an index set with $\operatorname{card}(I) = p$. The generation gap in this case is a variable and its value depends on the rates of the offsprings.
6. *Starting the next generation process*. Steps 2 to 5 are repeated until a certain stop criterion is fulfilled.

It should be emphasized that due to mutation, each future has the potential chance to be included in the subset under consideration at any iteration step. The lack of this opportunity has always been notified as one of the main drawbacks of other feature selection techniques.

Since the chromosomes are treated simultaneously the algorithm allows for parallel computation of the most time-consuming value: the criterion for each chromosome.

Suppose that the set S^* of the best L $(L > r)$ subsets has already been chosen by an exhaustive search on $F(X)$:

$$S^* = \{S_k \mid S_k \in F(X),\ J(S_k) > J(S_l),\ k \in I,\ \forall l \notin I\},$$

where card $(I) = L$.

Let $J_p(S_1, \ldots, S_r)$ denote the criterion assessing the classification accuracy of the parallel classifier whose first-level decision makers use the subsets $S_1, \ldots, S_r$. After the exhaustive search, the best combination of r out of top L subsets can be obtained, e.g., for $r = 3$:

$$J_p(S_{k1}^*, S_{k2}^*, S_{k3}^*) = \sup_{S_{k1}, S_{k2}, S_{k3} \in S^*} J_p(S_{k1}, S_{k2}, S_{k3}).$$

It is doubtful, however, that the classifier based on these subsets will surpass significantly the best single-level decision since the subsets in S^* may be dependent. This expresses the fact that the best subset of features is hardly isolated. It is probably surrounded with several "relatives", subsets with very similar structures. This case impedes searching for independent decisions which are to work in parallel, unless S^* is sufficiently large $(L \gg r)$.

The most appealing property of genetic algorithms in the context of parallel classifiers is that the search is performed on a set of alternatives simultaneously. Then the input subsets of features for the parallel classifier can be picked up directly from the population set. Due to the implied diversity of the population and its continuously increasing rating, it can be expected that high quality, independent (to some degree) subsets will be chosen. In this paper, along with the classical criterion "probability of correct classification", a modified criterion is suggested. It implies that the chromosomes which form the best combination of three subsets amongst the current population set, supplemented with the winners from the rest of the population and offsprings, remain in the renewed population set, i.e. Π is formulated by

$$\Pi = \{S_{k1}^*, S_{k2}^*, S_{k3}^*\} \cup \{S_k \mid S_k \in \Pi \cup O \setminus \{S_{k1}^*, S_{k2}^*, S_{k3}^*\},\ J(S_k) > J(S_l),\ k \in I,\ \forall l \notin I\}, \quad (1)$$

where card$(I) = p - 3$.

This corresponds to the "elitist" model which in this case, together with the best individual, preserves the best triple of chromosomes.

Let $\Phi(F(X))$ denote the set of all combinations of subsets of X. Applying genetic algorithm to it we can search for a solution at a higher level of abstraction. The task, however, becomes too complex because of the following reasons:

The coding of points in Φ is not trivial. Suppose that a simple presentation is chosen when each subset is represented as a bit in a string. Additional criteria should be included to shorten the chromosome size which in this case will be of card$(F(X))$ bits (e.g., for 7 features, card$(F(X)) = 127$).

The execution time will increase enormously because the number of 1s in the string may appear large. To avoid this, artificial restriction must be imposed on the chromosome, e.g., the "active" bits to be less than 7. Moreover, the frequent case of two-classes and majority principle of aggregation at the second level implies an odd number of first-level decision makers (classifiers). This should be somehow taken into account in formulating the chromosome.

Making allowances for all these factors, a heavy feature selection paradigm of doubtful use can be obtained in result. The search on $F(X)$ with the modified criterion (1) seems more computationally simple and effective.

3. Experimental study

3.1. Data sets

Three data sets were used as described below.

(1) The first set of data was taken from [13].

The set consists of 77 patients suffering from duodenal ulcer which are treated by Highly Selective Vagotomy (HSV). Each one is described by 11 preoperative features transformed previously into qualitative form. The two-class problem is considered in dependence on the complications after the HSV treatment. The problem is to preclassify the patient, i.e. to make a prognosis about the postoperative result and to supply advice to the physician.

(2) Data for hyaline membrane disease for 78 preterm newborn infants was used. All the children suffer from Respiratory Distress Syndrome (RDS) caused by different disorders one of which is Hyaline Membrane Disease (HMD). The early recognition of the disease (in the first hours after delivery) is extremely important since the type of the lung ventilation depends on the etiology of RDS. Data consists of 10 features from anamnesis, laboratory tests and clinical observations which could be measured during the first several minutes after delivery. Two classes have been formed according to the presence/absence of HMD as evaluated by experts.

(3) Data for hypoxic resistance of 200 healthy male pilots examined in barocamera [8] was used. The feature set consists of 7 features expressing the systolic blood pressure of the pilot in 7 different time moments during the examination. Two classes have been formed: good resistance to hypoxia and bad resistance, respectively.

The availability of three data sets from different fields of medicine provides an opportunity to generalize in some degree, or prevents inappropriate generalization of the experimental conclusions.

3.2. *Statement of the experiment*

Four experiments were carried out with each data set:

(E) Exhaustive search in $F(X)$ for obtaining the set S^* of the best 10 feature subsets.

(G1) Genetic algorithm with classical criterion: probability of correct classification. The estimate was calculated as the percent of correctly classified objects from the sample (leave-one-out method). The classification method was k-Nearest Neighbors (k-NN). The size of population set was 10, although there exist some considerations in favor of larger ones [12]. This restriction, caused on the one hand by our acceptation that the population size should depend on card $(F(X))$, and on the other, by the limited computational resources, led to some changes in the other parameters. Although the parameters' values differ significantly from the natural situation, here the experimental template advocated in [15] was ac-

Table 1
Probability of correct classification [%]

Data set	First level decision makers	Exhaustive search (**E**)	Random search (**R**)	Genetic algorithm (**G1**)	Genetic algorithm (**G2**)
1	1	81.8	79.7	81.8	81.7
	3	85.7	87.4	86.4	87.9
	5	87.8	84.3	86.5	88.2
	7	85.7	82.3	85.7	87.3
2	1	88.5	86.7	87.7	87.7
	3	92.3	90.4	91.4	93.2
	5	92.3	90.6	91.6	93.2
	7	91.0	90.4	91.3	92.3
3	1	89.5	88.9	89.5	89.5
	3	91.5	91.1	91.2	92.0
	5	92.5	91.6	91.6	92.4
	7	92.0	91.4	91.4	92.4

cepted, due to which the crossover probability P_c was stated to 1.0 and the mutation rate was chosen 0.15. For the first data set the population size chosen presents 0.49%, for the second 0.98%, and for the third 7.8% of the respective card($F(X)$). For the first and for the second data sets 25 generations were considered which means that at most 12.2% and 24.4% of all possible subsets were checked, respectively. For the third data set the generations were 10–78% of card($F(X)$).

(**G2**) Genetic algorithm with the same statement as described above but with the modified criterion (1).

(**R**) Random search. For each data set the same number of subsets were checked as in the genetic algorithms. The difference was that the formulation of the chromosomes was purely random.

For experiments **G1**, **G2**, and **R**, after the constitution of each new generation, the best parallel classifier of 3, 5, and 7 decision makers was detected by exhaustive search on the population set.

Since each pass of the genetic algorithm is unique, 10 experiments **G1**, **G2**, and **R** were carried out for each data set. The best two-level classification result during each experimental run were detected and an average value was calculated for **G1**, **G2**, and **R**, respectively.

4. Results and discussion

Table 1 presents the results of the experiments. The following conclusions may be drawn on their basis:

(i) The parallel classifier surpasses the best single classifier amongst the first-level components of the scheme. The improvement, even not very large, exists in all experiments.

(ii) Both **G1** and **G2** found better second-level classifiers than **E**. This fact supports the hypothesis that the set of the best 10 chromosomes may consist of "relatives" which would provide dependent first-level classification decisions.

(iii) The worst results were obtained by random search which succeeded neither in detecting the best single chromosome nor in finding sufficiently high result from parallel classifier.

(iv) A steady tendency of **G2** to surpass **G1** can be observed. Although the differences are not drastic they exist in almost all experiments.

(v) In the case of comparatively large population size (e.g. 7.8%) a sufficient number of independent classifiers may be contained in S^*. The greatest difference between two-level results obtained by **E** and **G2** is observed in the case with the smallest population size (0.48%). This emphasizes the fact that the use of genetic algorithm for small-scaled problems is ineffective.

References

[1] A.S. Bickel and R.W. Bickel, Determination of the near-optimum use of hospital diagnostic resources using the "GENES" genetic algorithm shell, *Comput. Biology Medicine* **20** (1990) 1–13.

[2] K. De Jong, Adaptive system design: A genetic approach, *IEEE Trans. Systems Man Cybernet.* **10** (1980) 566–574.

[3] P.A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach* (Prentice-Hall, Englewood Cliffs, NJ, 1982).

[4] F. Glover and H.J. Greenberg, New approaches to heuristic search: A bilateral linkage with artificial intelligence, *European J. Oper. Res.* **39** (1989) 119–130.

[5] D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning* (Addison-Wesley, Reading, MA, 1989).

[6] J.J. Grefenstette, Optimization of control parameters for genetic algorithms, *IEEE Trans. Systems Man Cybernet.* **16** (1986) 122–128.

[7] J. Jozefczyk, Determination of optimal recognition algorithm in the two-level system, *Pattern Recognition Lett.* **4** (1986) 413–420.

[8] K.H. Kunchev, Investigation on the resistance of the flying staff in respect to a moderate degree of hypoxia, Thesis, Sofia, 1981 (in Bulgarian).

[9] L.I. Kuncheva, Fuzzy multi-level classifier for medical applications, *Comput. Biology Medicine* **20** (1990) 421–431.

[10] R.P. Lippmann, Pattern classification using neural networks, *IEEE Comm. Mag.* (1989) 47–64.

[11] E. Masson and Y.-J. Wang, Introduction to computation and learning in artificial neural networks, *European J. Oper. Res.* **47** (1990) 1–28.

[12] D.J. Montana and L. Davis, Training feedforward neural networks using genetic algorithms, in: *Proc. 11th Internat. Joint Conf. on Artificial Intelligence*, Detroit, MA (1989) 762–767.

[13] Z. Pawlak, K. Slowinski and R. Slowinski, Rough classification of patients after highly selective vagotomy, *Internat. J. Man-Machine Studies* **24** (1986) 413–433.

[14] E. Sanchez, Genetic algorithms, neural networks and fuzzy logic systems, in: *Proc. 2nd Internat. Conf. on Fuzzy Logic and Neutral Networks*, Iizuka, Japan (1992) 17–19.

[15] W. Siedlecki and J. Sklansky, A note on genetic algorithms for large-scale feature selection, *Pattern Recognition Lett.* **10** (1989) 335–374.

[16] S. Stearns, On selecting features for pattern classifiers, in: *Proc. 3rd Internat. Conf. on Pattern Recognition* (1976) 71–75.