Map-Elites Algorithm for Features Selection Problem

Brenda Quiñonez[a], Diego P. Pinto-Roa [a], Miguel García-Torres[b], María E. García-Diaz [a], Carlos Núñez-Castillo[a] and Federico Divina [b]

[a]Facultad Politécnica - Universidad Nacional de Asunción [1]

[b]Division of Computer Science - Universidad Pablo de Olavide [2]

Abstract

In the High-dimensional data analysis there are several challenges in the fields of machine learning and data mining. Typically, feature selection is considered as a combinatorial optimization problem which seeks to remove irrelevant and redundant data by reducing computation time and improve learning measures. Given the complexity of this problem, we propose a novel Map-Elites based Algorithm that determines the minimum set of features maximizing learning accuracy simultaneously. Experimental results, on several data based from real scenarios, show the effectiveness of the proposed algorithm.

Keywords: Feature Selection, Map-Elites, Combinatorial Optimization, Machine Learning, Data Mining

1 Introduction

Recently, the available data has increased explosively in both the number of samples and the dimensionality in different machine learning applications, such as text mining, artificial vision and bio-medical. Our interest is mainly focused on the high dimensionality of the data. The large amount of high-dimensional data has imposed a great challenge on existing machine learning methods. The presence of noisy, redundant and irrelevant dimensions can make the learning algorithms very slow and can also generate difficulties when interpreting the resulting models [3]. In machine learning and statistics, the feature selection is the process of selecting a subset of relevant characteristics to use in building the model. Attribute selection methods greatly influence the success of data mining processes by reducing computational time and improving learning metrics, for this reason we propose a new attribute technique selection based on Illumination Algorithm [1].

This paper is organized as follows. Section 2 introduces to the features selection problem. Then, we describe the Illumination search algorithms in Section 3, specifically the Map-Elites algorithm. The section 4 contains the proposal of this paper and, finally, in the last section there is a brief discussion of the results obtained so far.

[1]{bquinonez,dpinto,mgarcia,nunez}@pol.una.py

[2]{mgarciat,fdivina}@upo.es

2 The Feature Selection Problem

A feature selection algorithm basically is the combination of a search technique to propose new subsets of features, with an evaluation measure that qualifies the different subsets. The simplest algorithm is to test every possible subset of features to find the one that minimizes the error rate. This is an exhaustive search of space, and it is computationally intractable except for the smallest feature sets; i.e. for n attributes, there are 2^n solutions. The choice of the evaluation metric has a great influence on the algorithm, and they can distinguish among three main categories: wrap methods, filter methods and embedded methods [3]. Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model. Filters are similar to Wrappers in the search approach, but instead of evaluating against a model, a simpler filter is evaluated. Embedded techniques are embedded in and specific to a model [3].

Many popular search approaches use greedy hill climbing, which iteratively evaluates a candidate subset of features, then modifies the subset and evaluates if the new subset is an improvement over the old. Evaluation of the subsets requires a scoring metric that grades a subset of features.

Search approaches applied to the feature selection include: exhaustive, best first, simulated annealing, genetic algorithm, greedy forward selection, greedy backward elimination, particle swarm optimization, targeted projection pursuit, Scatter Search, Variable Neighborhood Search [2, 4].

Genetic algorithm (GA) [5] method due to the capability to evolve new features of the selected features and a vast exploration of the search space for new fitter solutions. GA includes a subset of the growth-based optimization methods aiming at the use of the GA operators such as selection, mutation and recombination to a population of challenging problem solutions. GA has been effectively applied to several optimization problems such as classification tasks and pattern recognition. The GA's stochastic component does not rule out excitedly dissimilar solutions, which may give the better result. This has the advantage that, given sufficient time and a well bOlli1ded problem, the algorithm can discover a global optimum. It is well suited to feature selection problems because of the above reason. In the next section it will describe the MAP Elites algorithm based on genetic algorithms.

3 MAP Elites

The Multi-dimensional Archive of Phenotypic Elites (MAP-Elites) [1] algorithm illuminates search spaces, which produces a large diversity of high-performing, yet qualitatively different solutions, which can be more helpful than a single, high-performing solution. Interestingly, because MAP-Elites explores more of the search space, it also tends to find a better overall solution than state-of-the-art search algorithms. This is because MAP-Elites illuminates the relationship

between performance and dimensions of interest in solutions. MATP-Elites returns a set of high-performing and improves the state-of-the-art for finding a single, best solution, it will catalyze advances throughout all science and engineering fields.

MAP-Elites is quite simple. First, the user chooses a performance measure f(x) that evaluates a solution x. Second, the user chooses N dimensions of variation of interest that define a feature space of interest to the user. Each dimension of variation is discretized based on user preference or available computational resources. Given a particular discretization, MAP-Elites will search for the highest performing solution for each cell in the N-dimensional feature space. The search is conducted in the search space, which is the space of all possible values of x, where x is a description of a candidate solution [1].

4 MAP-Elites for Feature Selection

In this paper we propose to use the Map-Elites algorithm as an innovative technique for features selection in the automatic learning process. The challenge of this problem is that the inputs variable are binary one, whereas the basic Map-Elites was designed for real numbers. Therefore, the main objective of this proposal will be to create a search space for a MAP-Elites for the binary variables given the feature selection is a combinatorial problem.

To face this challenge, we represent a set of solutions as a vector with the indexes of the selected features. Algorithm 1 shows the pseudo-code of the Combinatorial MAP-Elites proposed. To create the map that allows us to distribute the solutions in the search space, we define the number of cells as the algorithm input parameter *NC*. Subsequently, this parameter is used to calculate a number of fixed features per cell *NFF*, which are used as cell identifiers and help determine which cell of the map each solution will be associated with (1). We also use two more input parameters for the algorithm that are typical of genetic algorithms such as the number of iterations I and the number of initial genomes G.

Algorithm 1 Combinatorial Map-Elites algorithm for feature selection

```
1:  procedure MAP-ELITES(NC, I, G)
2:      NFF ← log2(NC)
3:      MAP ← loadCell(NC, NFF)                          ▷ (1)
4:      for iter = 1 → I do
5:          if iter < G then
6:              MAP ← randomSolution()                   ▷ (2)
7:          else
8:              MAP ← randomVariation()                  ▷ (3)
9:          end if
10:     end for
11:     return MAP                                       ▷ feature-fitness map
12: end procedure
```

MAP-Elites starts by randomly generating G genomes (solutions coded) and determining the performance and features of each (2). In a random order, those genomes are placed into the cells to which they belong in the feature space (if multiple genomes map to the same cell, the highest-performing one per cell is retained). At that point the algorithm is initialized, and the following steps are repeated until a termination criterion is reached I. A cell in the map is randomly chosen and the genome in that cell produces an offspring via mutation and/or crossover (3). The features and performance of that offspring are determined, and the offspring is placed in the cell if the cell is empty or if the offspring is higher-performing than the current occupant of the cell, in which case that occupant is replaced by the new solution. The algorithm returns a map with the best solution found for each cell along with the corresponding fitness.

5 Discussion

The proposed Map-Elites tries to determine the minimum set of features that maximizes learning accuracy. This preliminary experiment was conducted using different data sets from real scenarios obtained from [2]. Table 1 presents performance of the algorithm where solutions were evaluated using Bayes Classifier and 2-fold cross validation [2]. In this experiment the input parameters for the algorithm were 5,000 iterations (I), a map of 8 cells (NC) and the number of initial genomes (G) equal to 500. In addition, we able to see the accuracy obtained by the proposed algorithm is promissory and at the same time it reduces the number of features.

Table 1: Map-Elites result experiment with Bayes Classifier

Dataset	All features	Fitness	Selected features
ionosphere	34	92.02 ± 2.38	12.6 ± .89
glass	9	70.12 ± 4.27	6.0 ± 1.00
anneal	38	96.44 ± 1.60	7.6 ± 1.34
tokyo1	44	92.91 ± 1.08	10.6 ± 2.51
spambase	57	91.76 ± .60	10.6 ± .89
kr-vs-kp	36	90.43 ± 1.46	3.0 ± .00
corral	6	86.90 ± 2.22	5.0 ± .00
breast-cancer	9	71.34 ± 3.95	3.6 ± .89
hypothyroid	29	96.66 ± .29	1.0 ± .00
labor	16	91.21 ± 11.14	5.0 ± .71
vote	16	95.63 ± 1.26	1.0 ± .00

Currently, the performance of Map-Elites is being tested and compared with the competitive algorithms of the-state-of-the-art.

References

[1] Jean-Baptiste Mouret and Jeff Clune. “Illuminating search spaces by mapping elites”. CoRR. 2015.vol. abs/1504.04909

[2] Félix García López, Miguel García-Torres, Belén Melián Batista, José A. Moreno Pérez, and J. Marcos Moreno-Vegatitle. “Solving feature subset selection problem by a Parallel Scatter Search”. European Journal of Operational Research. 2006.

[3] Miao, Jianyu Niu, Lingfeng. (2016). A Survey on Feature Selection. Procedia Computer Science. 91. 919-926. 10.1016/j.procs.2016.07.111.

[4] M. Garcia-Torres, F. Gomez-Vela, B. Melian, J.M. Moreno-Vega. High-dimensional feature selection via feature grouping: A Variable Neighborhood Search approach, Information Sciences, vol. 326, pp. 102-118, 2016.

[5] Sindhiya, S Selvaraj, Gunasundari. (2015). A survey on genetic algorithm based feature selection for disease diagnosis system. Proceedings of ICCCS 2014 - IEEE International Conference on Computer Communication and Systems. 164-169. 10.1109/ICCCS.2014.7068187.