

# EMO Shines a Light on the Holes of Complexity Space

Núria Macià, Albert Orriols-Puig, Ester Bernadó-Mansilla  
Grup de Recerca en Sistemes Intel·ligents  
La Salle - Universitat Ramon Llull  
C/ Quatre Camins, 2 08022 Barcelona (Spain)  
nmacia@salle.url.edu, aorriols@salle.url.edu, esterb@salle.url.edu

## ABSTRACT

Typical domains used in machine learning analyses only partially cover the complexity space, remaining a large proportion of problem difficulties that are not tested. Since the acquisition of new real-world problems is costly, the machine learning community has started to give importance to the automatic generation of learning domains with bounded difficulty. This paper proposes the use of an evolutionary multi-objective technique to generate artificial data sets that meet specific characteristics and fill these holes. The results show that the multi-objective evolutionary algorithm is able to create data sets of different complexities, covering most of the solution space where we had no real-world problem representatives. The proposed method is the starting point to study data complexity estimates and steps forward in the gap between data and learners.

## Categories and Subject Descriptors

D.2.8 [Software Engineering]: Metrics—*complexity measures, performance measures*

## General Terms

Algorithms

## Keywords

Data complexity, evolutionary multi-objective optimization, artificial data sets

## 1. INTRODUCTION

In supervised learning, the performance of new learning techniques is usually analyzed by comparing the new approaches with the state-of-the-art methods over a collection of data sets that come from different real-world domains. Conclusions extracted from these types of comparisons are often built upon a table of results which contain a measure or several measures of performance of the methods across the data sets; then, statistical tests are applied to support hypotheses about the excellence of the learners [2]. Although this enables the derivation of conclusions about the general behavior of a given learner, the particular problem complexities that affect the different techniques are poorly analyzed and understood. Hence, it is usual to find that the best

method on average provides the poorest results in specific data sets. This fact has awakened the interest in estimating the complexity of data and using these estimates to identify which problem difficulties affect different learning systems.

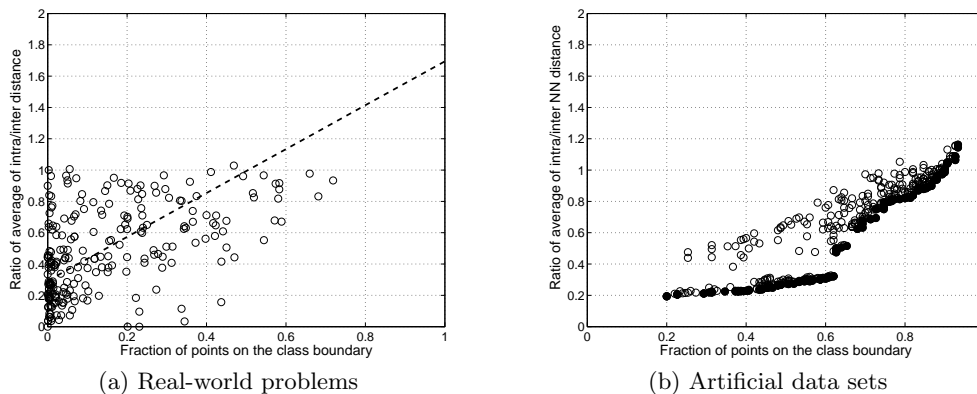
Learner accuracy depends on both the intrinsic characteristics of data and the algorithm design which includes the knowledge representation and the generalization process employed by the algorithm. Recent studies have derived metrics to evaluate different sources of complexity [3] and linked data complexity to the performance of particular learners. Despite the initial promising results, these works also detected two important challenges that need to be addressed to advance in the field of data complexity. First, it emphasized that the complexity metrics designed could not fully capture all the sources of complexity of classification problems. This has resulted in some efforts to develop new complexity metrics. Second, experimental results over a large collection of real-world data sets from public repositories revealed that there were complexities that could not be found in the selected data sets. There were holes in the complexity measurement space that were not covered by any data set (see Figure 1(a)). This led to the idea of generating artificial data sets (ADS) with bounded difficulty that were a representative sample of regions of the complexity space and set the baseline to further study the gap between data complexity and learner performance.

The purpose of this paper is to design and implement a technique based on evolutionary multi-objective optimization (EMO) to generate ADS which satisfy different types and levels of complexity. The proposed approach is based on the NSGA-II algorithm [1] and optimizes some of the complexity metrics provided in [3].

The remainder of this paper presents the experiments performed which produce a diverse set of data sets that enables, for the moment, a few spots of the complexity space and discusses future directions.

## 2. TOWARD A MULTI-OBJECTIVE OPTIMIZATION PROBLEM

To cover the holes of the data complexity space, a collection of data sets with different complexity levels for several estimations provided by the complexity metrics is required. For this purpose, we define the following multi-objective optimization problem. We consider a set of  $n$  unlabeled examples  $\{e_1, e_2, \dots, e_n\}$  drawn from a random distribution or from a physical process. Then, the optimization problem consists in searching the combination of class labels  $\{c_1, c_2, \dots, c_n\}$  that satisfies the  $m$  objectives of the problem,



**Figure 1: Data complexity space characterized by two measures of class separability, fraction of points on the class boundary and ratio of average intra/inter class nearest neighbor distances.**

which correspond to  $m$  complexity metrics.

Figure 1(b) shows the complexity of the solutions obtained by the EMO approach. The plot illustrates all the possible solutions, identifying those that belong to the Pareto set with black circles. We observe that the method could obtain a Pareto front with a large number of solutions when the fraction of points on the class boundary was maximized while the ratio of average intra/inter nearest neighbor distances was minimized. These ADS spread across the feature space, achieving the objective of creating data sets diverse enough to fall in the blind spots detected with real-world problems. On the other hand, the generation of ADS could provide insights into the correlation and dependencies among complexity metrics. For instance, the Pareto front presents a linear increase that may indicate that both measures are strongly correlated.

### 3. LOOKING AHEAD

The preliminary results have shown that the method holds promise, being able to generate problems with certain difficulties that could be rarely found in real-world problems [3]. These results encourage us to continue on advancing on the use of evolutionary computation for ADS generation. In future work, we aim at using boundedly-difficult problems to (1) study the structure of these synthetic problems with different complexities and (2) refine complexity estimates and compare several classification techniques over groups of problems with similar characteristics.

A key point in the generation of ADS is to analyze whether the underlying distribution of the resulting data sets is similar to the structure of real-world concepts. Our approach focuses on labeling a set of existing instances without controlling its distribution. We do not therefore prevent the system from evolving any type of solution. In order to obtain data sets with a predefined structure, we propose to (1) use physical processes or probability distributions that occur in nature to design the original collection of unlabeled examples, (2) include constraints in the optimization process to maintain a certain distribution of examples per class so that we resemble real-world domains, and (3) add an instance selection procedure.

The second future line is related to the use of ADS with similar characteristics to test classification techniques over.

In the current studies, performance of new algorithms is analyzed by comparing the new approach with a set of existing learning methods over data sets which are typically haphazardly selected from data repositories. In this kind of approach, we do not know anything about the complexity of the data sets and choosing a different collection of data sets may lead to contradictory conclusions. To avoid this effect, some comparisons have included a larger number of data sets. This practice may yield to more reliable conclusions since a larger variety of data sets is considered. Nevertheless, we still cannot guarantee that data sets with different characteristics are added. Actually, even considering a collection of 264 data sets, we were covering only partially the whole complexity space. In addition, it is worth mentioning that, as theoretically demonstrated by the no-free-lunch theorem [4], if we were able to include data sets with different characteristics so that we could cover all the complexity space, all the learning methods would perform the same on average. For this reason, we would like to use the EMO approach to establish different comparisons of learning techniques over collections of data sets with similar characteristics with the aim of identifying the sweet spot in which each algorithm actually outperforms the other ones.

### 4. ACKNOWLEDGMENTS

The authors thank the *Ministerio de Educación y Ciencia* for its support under the project TIN2008-06681-C06-05, *Fundació Crèdit Andorrà*, and *Govern d'Andorra*.

### 5. REFERENCES

- [1] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.
- [2] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.
- [3] T. K. Ho, M. Basu, and M. Law. Measures of geometrical complexity in classification problems. In *Data Complexity in Pattern Recognition*, pages 1–23. Springer, 2006.
- [4] D. H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.