

Beyond Homemade Artificial Data Sets

Núria Macià, Albert Orriols-Puig, and Ester Bernadó-Mansilla

Grup de Recerca en Sistemes Intel·ligents
La Salle - Universitat Ramon Llull
C/ Quatre Camins 2, 08022 Barcelona (Spain)
`{nmacia, aorriols, esterb}@salle.url.edu`

Abstract. One of the most important challenges in supervised learning is how to evaluate the quality of the models evolved by different machine learning techniques. Up to now, we have relied on measures obtained by running the methods on a wide test bed composed of real-world problems. Nevertheless, the unknown inherent characteristics of these problems and the bias of learners may lead to inconclusive results. This paper discusses the need to work under a controlled scenario and bets on artificial data set generation. A list of ingredients and some ideas about how to guide such generation are provided, and promising results of an evolutionary multi-objective approach which incorporates the use of data complexity estimates are presented.

Key words: Data complexity, artificial data sets, machine learning.

1 Introduction

Machine learning techniques have a practical application on a large variety of real-world problems. The diversity of domains—medicine, industry, learning—provides extremely disparate data sets regarding the type of features, volume of instances, and data distribution, among others. All of these characteristics have led to the implementation of different strategies to tackle each problem properly, since learner performance depends partly on the algorithm design. At present, the development of techniques has reached an advanced state of maturity offering thousands of methods, all of them very competitive, providing accurate models from data which are generalized from a sample of the problem at hand. Despite the headway progress in data classification, many questions remain unanswered such as how the intrinsic characteristics of the data sets affect learners. This, coupled with the little leeway for improvement and the uncertainty of the ability of techniques to fully capture the underlying knowledge of data, leads us to look toward other elements involved in the learning process. At this point, data steals the limelight from learners.

Some authors have started giving importance to the study of data complexity in supervised learning, and recent studies, compiled in [2], have shown that learner performance also depends on data complexity. This dependence has unleashed a new research line which is focused on the analysis of the nature of problems and whose aim is to characterize data and relate them to learner

properties. This link, based on complexity metrics, could help assess the ability of learners and recommend which learner should be applied to solve a specific problem. Nonetheless, again doubts arise about the reliability of these complexity estimates, since the characterization is made up of real-world problems and validated by machine learning techniques whose relationship is close but not yet well-defined [3]. Thereby, it highlights the need to create data sets of bounded difficulty in order to analyze complexity metrics and learners under a controlled scenario.

The purpose of this paper is to present a new technique to generate artificial data sets (ADS) diverse enough to provide a solid experimental framework where the analysis of learner behavior and learner performance can be carried out. There have been the first attempts to build ADS based on complexity estimates by using heuristic searches [10] and genetic algorithms [9]. Despite of the fresh aroma of these proposals, the approaches just optimize one complexity dimension. In this work, we explode this idea and use a multi-objective evolutionary algorithm to make data sets that meet different types and levels of complexity.

The remainder of the paper is organized as follows. Section 2 briefly reviews data complexity analysis. Then, Section 3 discusses the why, what kind, and how to generate ADS. The mix of some points dealt with in the previous section results in a “*nouvel* generator of ADS” whose design and results are presented in Sections 4 and 5 respectively. Finally, Section 6 concludes the work with some future directions.

2 Data Complexity

In order to define the relationship between data and learners, some studies investigated problem characterization by means of different estimates based on difficulty factors. Ho and Basu first identified the sources of problem difficulty [7] and proposed the following classification: (1) class ambiguity, (2) boundary complexity, and (3) sample sparsity and feature space dimensionality. *Ambiguity* refers to the situation when there are examples whose features cannot permit distinguishing their classes. Usually, this ambiguity is due to the problem formulation in which the concepts are intrinsically inseparable or the set of attributes is no adequate or sufficient to describe the concepts. *Boundary complexity* is related to the description of the class boundary. Class separability and problem linearity are based on the geometrical complexity of data structure. Finally, *sample sparsity and feature space dimensionality* are concerned with the difficulty layer that an incomplete or sparse sample add to the problem, enkindling the importance of the sample representativity.

Among these sources of problem difficulty, investigations carried out by Ho and Basu focused on boundary complexity because of the difficulty to determine the class ambiguity and the real sparsity of a training set. Thus, they designed a set of measures to estimate the class boundary [7]. These measures evaluate different aspects such as (1) overlaps in features values from different classes, (2) separability of classes, and (3) geometry, topology, and density of manifolds. In later studies, the data characterization built upon this set of metrics, which

provides a space of data complexity, permitted determining some relationships between certain estimates and certain learners from different paradigms [3].

Although preliminary results showed some correlations between these complexity estimates and learner accuracy, the link between data characteristics and learner properties is not mature enough. There are still too many relationships among data, complexity metrics, and learners, highly dependent and out of control [6]. Therefore, in order to avoid again getting partial conclusions through pieces of problems and apparent estimations, we have to resort to ADS. The generation of ADS is the procedure to establish the framework to study estimates of data complexity and learner performance.

3 The Why, What Kind, and How to Generate ADS

This section discusses ADS generation. We present a general picture of this incipient topic by answering why we need ADS, what kind of characteristics they should have, and how to provide ADS with these desirable requirements.

3.1 Why?

Over the last few decades, the machine learning community has designed and developed techniques to solve real-world problems and to extract knowledge from their data. To validate the efficiency of these techniques, the most usual methodology adopted by the community consists in testing new techniques on a collection of real-world problems and comparing the obtained accuracy with other learners. Nevertheless, this procedure may lead to inaccurate conclusions due to (1) real-world problems constraints and (2) data dependence of the learner.

Usually, learners are tested using real-world problems from public repositories. Even though sharing these problems benefits the obtaining of a common test bed for the experiments and facilitates the comparison between the own and the community results, these data sets may result in misleading conclusions. On the one hand, the current sets are composed of few problems whose independence is unknown, i.e, we ignore whether this set of problems is representative enough to cover the whole problem space. We cannot guarantee that these problems are diverse enough to test the learner limitations in an exhaustive way, since there are no studies that indicate what problems, regardless of the domain to which they belong, are structurally similar. On the other hand, the high cost of experiments, the difficulty of conducting them, or data privacy policies hinder data collection, resulting in complex data sets characterized by few instances, missing values, and imprecise data. The combination of these deficiencies in the data sample goes beyond our control, blurring our knowledge of to what extent the influence of these constraints negatively affects learner performance.

Empirical results show that there exist learning paradigms more suitable to solve a type of problems than others. Nevertheless, the learner dependence on opaque data is responsible for our lack of knowledge of the relationship between data characteristics and learner properties, limiting us in the progress of learning techniques. To overcome this, we need to work under a controlled scenario, with a certain kind of data, where complexity is known.

3.2 What kind?

Being aware of the need of artificial data sets to test learner performance, we have to define what kind of data set should be generated. To this end, we focus on the concepts for classification learning taking into consideration (1) data structure and (2) complexity factors.

Data structure present in real-world problems is significant in data analysis since these structures contain the underlying knowledge. Thus, we should force data sets to resemble real-world problems and attain such real structures in data. It means that data not only have to follow uniform or gaussian distributions but also have to include physic processes. Moreover, for classification problems, class labeling has to conform with clustering rules.

Complexity factors are related to the aforementioned aspects that are measured by the complexity metrics, such as the discriminative power of attributes, class separability, and geometry. Firstly, we have to generate well-defined problems with a known underlying concept and whose definition is complete and without ambiguity. After defining which characteristics have to describe data, constraints have to be introduced by varying their degree of difficulty to test different learner abilities. This implies relating difficulty factors to the type of performance that we want to assess, such as robustness, scalability, and predictive accuracy. For instance, noise, missing values, or ambiguity are suitable characteristics to test the learner robustness. Learner scalability would be tested by varying the number of features and the number of instances. By adding irrelevant or redundant attributes, a relevance analysis can be performed. Determining the number of classes of the problem adds another layer of difficulty, since some of the complexity factors have to be interpreted differently.

Generators of artificial data sets have to allow us to tune all these characteristics to test learner efficiency in particular cases and comprehend learner behavior in front of specific constraints.

3.3 How?

We pursue the generation of data sets which meet different complexity levels for different difficulty factors and whose structure resemble real-world problems.

The first requirement involves addressing the problem as an optimization problem in which each objective to minimize or maximize becomes a complexity metric. In this regard, multi-objective evolutionary algorithms (EMO) [4] are a natural support to conduct optimization problems with several objectives. In the problem definition, we consider a set of n unlabeled examples $\{e_1, e_2, \dots, e_n\}$, and the system proceeds to search for the combination of class labels $\{c_1, c_2, \dots, c_n\}$ that satisfies m predefined objectives, which correspond to m complexity metrics.

The second requirement, concerning the structure of the problems, could be achieved by generating the initial data set according to fixed distributions. Correlations among features can be set by users guided taking into account statistics extracted from real data [8]. The use of existing samples of real-world problems or of learning techniques such as instance selection and feature selection are alternatives to dynamically manage distributions. Regarding the class labeling,

some instance classes could be previously fixed or grouped following real-world problems labeling.

4 EMO-Made Artificial Data Sets

In this section, we propose a data set generator based on a multi-objective evolutionary algorithm. In particular, we used the *non-dominated sorting genetic algorithm* (NSGA-II) [5] to tackle the optimization of data set complexities. However, the implementation or election of this method can change since the interest lie on the multi-objective concept.

In what follows, we first describe the meta-information required by the algorithm and the knowledge representation. Then, we detail the process organization and the genetic operators employed in our approach. And finally, we present empirical results to illustrate the outcome of the system.

4.1 Meta-Information and Knowledge Representation

In our implementation, we aim at obtaining the class labeling for a data set that meets different degrees of values from the specified set of complexity metrics. For this purpose, we have to define (1) the meta-information that needs the system, (2) the genetic representation of the solution of the problem, and (3) the fitness function to evaluate each candidate solution.

Meta-information refers to data structure and data itself. The first step is to load a data set containing unlabeled examples whose dimensionality in terms of number of instances and number of attributes is predefined by the user. Each instance is defined by m continuous- or nominal-valued attributes, and samples can be randomly generated following any kind of distribution or using a real-world distribution directly.

As for the *genetic representation*, the EMO system evolves a population of N individuals. Each candidate solution represents a class labeling of the data set which is encoded by a k -ary array (k is a configuration parameter that indicates the maximum number of classes of the data set) where the position i corresponds to the class label of the i th instance. Note that the individual size is constant and is determined by the number of instances contained in the data set.

The EMO technique searches the best combination of labels that satisfies the required complexity for the specified metrics. The *fitness function* for each objective corresponds to the computation of different complexity metrics.

4.2 Process Organization and Genetic Operators

In the following, we briefly describe the process organization and how the genetic operators are combined.

The NSGA-II algorithm evolves a population P_t of N individuals which are initialized at random and evaluated. To avoid dealing with special cases in the first iteration of the algorithm, we also create an auxiliary population Q_t of N individuals whose individuals are initialized randomly and evaluated as well.

Then, the procedure iteratively applies the following steps. First, populations P_t and Q_t are joined into population R_t , which contains $2N$ individuals. R_t is ranked according to the fast non-dominated sorting approach, which divides solutions up into different fronts. Then, starting from the first front, all the solutions of each front i are introduced into the new population P_{t+1} provided that there is enough room to allocate all the solutions of the given front. Otherwise, the solutions with the highest crowding distance of the front i are introduced into P_{t+1} until filling all the population; thence, no more solutions of higher fronts are added to the population.

Then, the offspring population Q_{t+1} is created using classical GA operators—selection, crossover, and mutation. The individuals are chosen from the parent population by means of s -wise tournament selection. Pairs of these parents are selected without replacement, and they undergo crossover and mutation with probabilities χ and μ respectively. If neither of both operators is applied, the parents are directly copied in the new population Q_{t+1} . Then, both populations constitute the population R_{t+1} whose individuals are reevaluated. This process iterates until the stop criterion is met, i.e., the number of generations is reached.

The EMO approach includes two sorting concepts: (1) the fast non-dominated sorting and (2) the crowding distance assignment. Thanks to these mechanisms the system can optimize different objectives in a single simulation run. The former organizes the population into different fronts, and the latter estimates the density of the solutions surrounding a particular solution in the population. Concerning the classical genetic operators, we used: (1) s -wise tournament selection, where tournaments of s randomly chosen parents are held, and the best parent, according to the crowded-comparison operator, is selected for recombination; (2) two-point crossover, which, provided two parents, randomly generates two cut points and uses them to shuffle the information of both parents; and (3) bit-wise mutation, which flips the value of the bit selected for mutation.

5 Experimental Results

The purpose of the experiments is to show how the system is able to provide a diverse test bed composed of data sets with different complexity levels across the required difficulty factors.

To this end, we fed the system with two types of meta-information: (1) a data set that follows a uniform distribution and (2) the iris problem [1] with a real-world problem structure. For all the runs, the system was initialized with a population of 400 individuals which was evolved during 50 generations. The probabilities of crossover and mutation were 0.85 and $1/n$ respectively, where n is the individual size, i.e., the number of instances of the data set.

Figure 1 plots the data sets characterized by different complexity metrics. The x-axis and y-axis represent the demanded objectives. In particular, we focus on optimizing three complexity metrics of the aforementioned set proposed by Ho and Basu: (1) the fraction of points on the class boundary (N1), (2) the ratio of average intra/inter nearest neighbor (NN) distance (N2), and (3) the ratio of the maximum Fisher’s discriminant (F1).

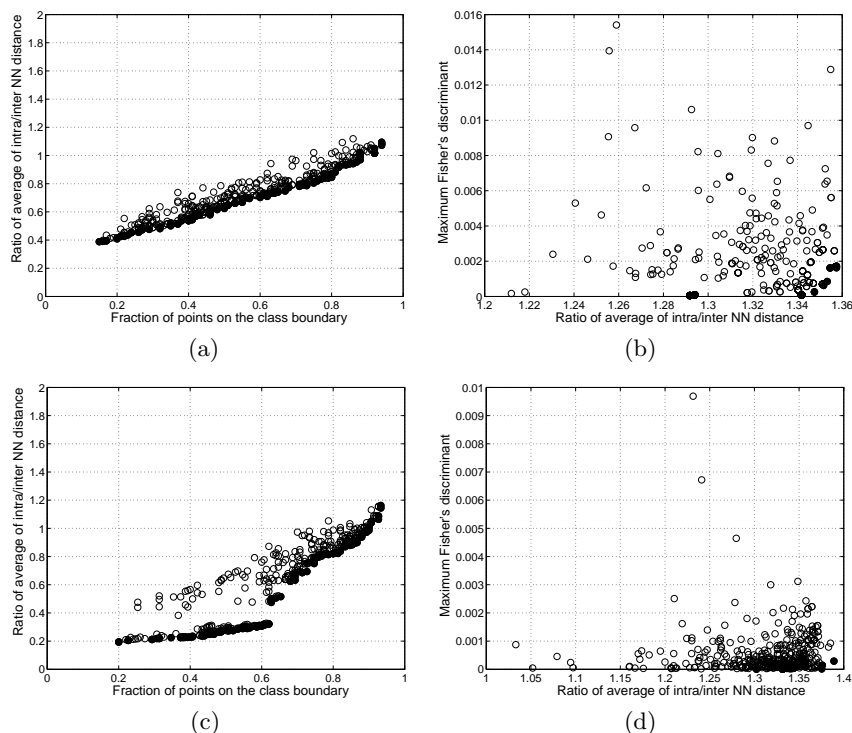


Fig. 1. Artificial data sets characterized by different complexity metrics. The upper plots refer to data sets whose structure follows a uniform distribution and, the lower plots correspond to the iris problem.

The solutions in the Pareto front are depicted with a black circle. We observe that, regardless of the distributions, the system finds solutions across the complexity space drawing the Pareto-optimal. Figures 1(a) and 1(c) draw the optimization of two class separability measures, the maximization of $N1$ and the minimization of $N2$. In this case, we obtained data sets whose complexity estimated by $N1$ and $N2$ was ranged in $[0.11, 0.96]$ and $[0.32, 1.17]$ respectively. The greater the value of these measures is, the higher the complexity is. Thus, this variability permits analyzing the learner ability according to the density of the class boundary. In Figures 1(b) and 1(d), one measure of class separability $N2$ was maximized and one measure of feature overlap was minimized, in particular $F1$. Thanks to these data sets, we could test the learner robustness to the class separability for data sets in which all the attributes are relevant, since the lower the value of $F1$ is, the lower the discriminative power of the attributes is.

The results show that our proposal can build data sets with different characteristics defined under a set of difficulty factors to test specific learner properties. However, we have to enhance the system and provide it with mechanisms to control the class balance and to enable dynamic distributions.

6 Conclusions

This work has presented a new approach of an old practice. Design and implementation of ADS are familiar to any practitioner that has once created data sets to test specific properties of learning techniques or to highlight the discovery of a concrete behavior. However, the main drawback of these homemade data sets is that they are *ad hoc* to the given problem, and often because of their no formal design, they could be in doubt. Our proposal, based on evolutionary learning, attempts to go one step further by providing data set generation with a theoretical basis, which allows us to produce generic data sets whose characteristics are customized by means of complexity estimates. Hence, data sets generated can satisfy different complexity levels at the same time.

The discussion and the proposed approach could be a turning point in ADS generation. Nevertheless, there is still a long way until achieving a set of benchmark problems since data characteristic definition is subject to the maturity of the study of data complexity. Attaining a complete experimental platform may imply including new complexity factors or the redefinition of the existing ones.

Acknowledgments. The authors would like to thank the *Ministerio de Educación y Ciencia* for its support under the project TIN2008-06681-C06-05. They also acknowledge *Fundació Crèdit Andorrà* and *Govern d'Andorra*.

References

1. A. Asuncion and D. Newman. UCI machine learning repository, 2007.
2. M. Basu and T. K. Ho. *Data Complexity in Pattern Recognition*. Springer-Verlag, 2006.
3. E. Bernadó-Mansilla, T. K. Ho, and A. Orriols-Puig. Data complexity and evolutionary learning. In *Data Complexity in Pattern Recognition*, pages 115–134. Springer, 2006.
4. C. A. Coello, G. B. Lamont, and D. A. V. Veldhuizen. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Springer, New York, 2nd edition, 2007.
5. K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE TEC*, 6:182–197, 2002.
6. T. K. Ho. Data complexity analysis: Linkage between context and solution in classification. In *Proceedings of the Joint IAPR International Workshops on Structural and Syntactic Pattern Recognition (SSPR 2008) and Statistical Techniques in Pattern Recognition (SPR 2008)*, 2008.
7. T. K. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Transactions on PAMI*, 24(3):289–300, 2002.
8. D. R. Jeske, B. Samadi, P. J. Lin, and L. Ye. Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems. In *11th International Conference on Knowledge Discovery in Data mining*, pages 756–762, 2005.
9. N. Macià, E. Bernadó-Mansilla, and A. Orriols-Puig. Preliminary approach on synthetic datasets generation for classification. In *2008 International Conference on Pattern Recognition*, volume 5342/2008, pages 986–995, 2008.
10. N. Macià, A. Orriols-Puig, and E. Bernadó-Mansilla. Genetic-based synthetic data sets for the analysis of classifiers' behavior. In *Proceedings of the 2008 Hybrid Intelligent Systems Conference*, pages 507–512, 2008.