7 THE LANDSCAPE

Summary. This chapter sketches out a landscape composed of artificial data sets which cover some gaps in the complexity space and provide an extended test bed. This measurement space is used to evaluate the performance of particular classifiers as well as a basis of a competence domain contest. Interestingly, the results of the contest point out the existence of benchmark-flavoured data sets.

7.1 INTRODUCTION

The analysis of the performance of learning algorithms should rely on a known, controlled testing framework due to the dependence between the capabilities of learners and the intrinsic complexity of data observed in previous chapters. By means of the evolutionary multi-objective optimisation approach presented in Chapter 6, artificial data sets are generated to cover reachable regions in different dimensions of data complexity space. *The landscape* provides a configurable framework to evaluate supervised classification techniques and detect their limitations. Systematic comparison of a diverse set of classifiers highlights their merits as a function of data complexity. Detailed analysis of their comparative behaviour in different regions of the space gives guidance to potential improvements of their performance.

The purpose of this chapter is to introduce a synthetic landscape that combines real-world problems and artificial ones, and offer a broad range of data sets. Preliminary experiments pursue (1) to support the belief that rather than a unique and globally superior classifier, there exist local winners, (2) to highlight the critical role of the test framework, and (3) to envisage how this space may help to understand the limitations of classifiers and offer guidance on the design of improvements that can push the boundaries of their domains of competence.

In the following, we state the objectives of the landscape, describe the generated artificial data sets, and present the results of different classifier systems. Finally, we discuss the danger of benchmarks.

7.2 OBJECTIVES OF THE LANDSCAPE

Problems that provide a good coverage of the data complexity space are necessary to perform exhaustive studies of learners. This section sets the general goals as well as the specific ones of our landscape.

7.2.1 *General goals*

We aim to design a wide set of problems that covers all the spectrum of complexity to:

1. Thoroughly assess learners' performance.
2. Identify domains of competence of learners.
3. Provide the scientific community with complete, complex benchmarks and escape from toy problems.
4. Generate specific data sets to test learners' limitations.

7.2.2 *Landscape potential*

Indeed, we aim to study learners over a large set of problems and be able to identify patterns in their behaviours, domains of competence. To this end, we designed a contest, held before the 20th International Conference on Pattern Recognition in 2010[1], that used a collection of problems selected on the basis of their complexity characteristics. Evaluation of the participating algorithms with this collection of problems provided the landscape featuring the domains of competence of each algorithm in the data complexity space. With this initiative, we looked for gaining visibility in the community and having the possibility of testing a diverse set of algorithms.

7.3 DESIGN OF THE CONTEST

In order to prepare the data for the contest, four data set collections are created. This section details the design.

First, we test (1) the learner behaviour over the problem space (S1, S2, and S3) and then (2) the learner local behaviour in their domain of competence (S4). The four data set collections are summarised as follows:

S1. Collection of data sets covering the reachable complexity space, designed for training the learner. All the instances are duly labelled.

S2. Collection of data sets covering the reachable complexity space, designed for testing the learner. No class labelling is provided.

S3. Collection of data sets with no class labelling, like S2, that was used in a live test that was run for a limited period of time (one hour). This follows the idea of a fair validation using unknown data and avoid overfitting from the classifiers.

S4. Collection of data sets with no class labelling, which covers specific regions of the complexity space where each learner dominates. Its design is aimed at determining the stability of dominance over the local neighbourhood.

7.4 THE RESULTING DATA SETS OF THE LANDSCAPE

To build the four collections, we generate 80,000 data sets running the EMO approach over five seed problems: Checkerboard, Spiral, Wave Boundary, Yin Yang, and Pima, explained in Chapter 6. This section describes the generation procedure.

The five seed data sets are evolved for different objective configurations, plus all the combinations of the optimisation of three complexity measures at each time; In particular, F2L1T2, F2L2L3, F2N1T1, F2N4T1, F3N2T1, F3N3L3, L1L2T1, L1L3T2, L1T1T2, N1L3T1, N1N2N3, N1N4L1, N1N4T2, N2N4L1, and N3L1T2. The selection of three complexity measures results in eight experiments which consist in maximising (1) or minimising (0) each dimension, (i.e. 000, 001, ..., 111). Each combination puts together the two least correlated measures with respect to a third one. The entire generation process uses eleven of the twelve initial complexity measures, with the omission of the maximum Fisher's discriminant ratio (F1)—since its extension to m-class is under study.

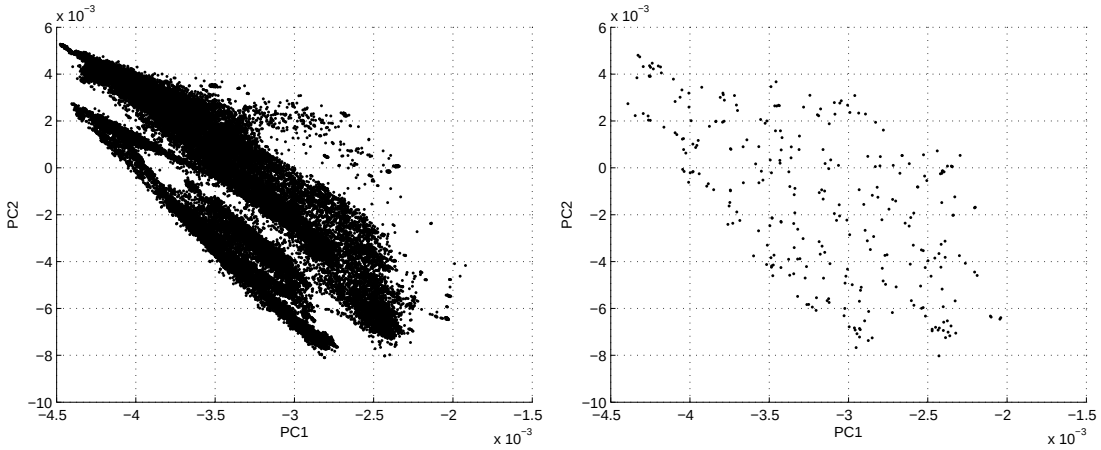
7.5 COVERAGE OF THE MEASUREMENT COMPLEXITY SPACE

Following the analysis performed in Ho and Basu [2002], we calculate the Singular Value Decomposition (SVD) over all the complexity measures and build a space with the first two principal components (see Fig. 7.1a), assuming Gaussian distribution in our data. This section presents the coverage of the landscape.

1 http://www.salle.url.edu/ICPR10Contest/

Table 7.1: Singular value decomposition of F1v-T2 over the landscape collection. *PC* refers to the principal components and *Std dev* to the standard deviation. *Variance* corresponds to the proportion of variance and *Cumulative* to the cumulative proportion.

PC	Std dev	Variance	Cumulative
PC1	13.1970	0.93820	0.93820
PC2	3.34290	0.06020	0.99830
PC3	0.44700	0.00108	0.99943
PC4	0.24744	0.00033	0.99976
PC5	0.11871	0.00008	0.99983
PC6	0.10236	0.00006	0.99989
PC7	0.09201	0.00005	0.99994
PC8	0.08173	0.00004	0.99997
PC9	0.05791	0.00002	0.99999
PC10	0.03747	0.00001	1.00000
PC11	0.01696	0.00000	1.00000
PC12	0.01510	0.00000	1.00000
PC13	0.00447	0.00000	1.00000



(a) (b)

Figure 7.1: Projection of the problems on the first and second principal components extracted from the complexity measurement: (a) the entire collection composed of 80,000 data sets and (b) 300 cherry-picked training data sets for use in the contest.

Table 7.1 shows the SVD of the complexity measures, specifically from F1v to T2, over the collection composed of 80,000 data sets synthetically generated. We can see that only with the first two principal components, we reach the cumulative proportion of 0.99.

To limit the size of the contest, we decide to select a sample from the collection. We divide the space into 100 cells and pick five data sets at random from each cell. Fig. 7.1b plots the distribution of the 300-data set sample from the generated collection. These are the 300 data sets used in the contest for S1; S2, S3, and S4 also contain 300 data sets with a similar distribution over the complexity space.

7.6 LANDSCAPE VS. UCI AND PSP

The coverage analysis performed in Chapter 5 reveals some gaps in the complexity space. This section shows how the landscape fills some of them.

Fig. 7.2 depicts the complexity of the landscape collection composed of 80,000 data sets and compares the variability obtained with only five seeds with the variability offered by the UCI

repository. In our artificial landscape, values of L2, L3, N2, N3, and N4 have a greater range than in the UCI collection and, as a consequence, provide more diversity. For the rest, the spread is equivalent, noting that our collection is 1,000 times bigger. However, despite the large number of data sets that contains the artificial landscape, the values reached are encouraging since there are plenty of seeds that can be transformed with the EMO and which may lead to a complete coverage of this space.

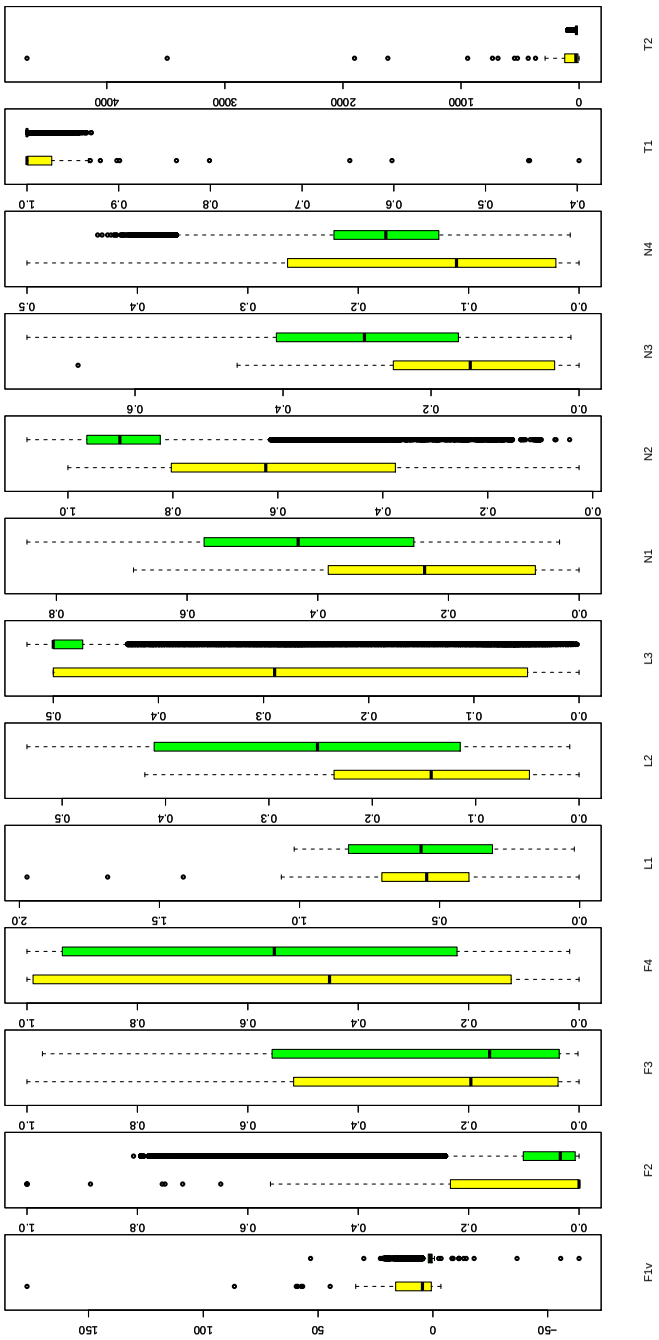


Figure 7.2: Comparison between the UCI repository (yellow boxplot) and the artificial landscape (green boxplot).

Regarding the comparison between the UCI repository and the PSP repository in Fig. 5.6 (Chapter 6), the gaps observed for L2 and L3 have been slightly shrunk by moving the limits from 0.42 to 0.534 and from 0.511 to 0.525 respectively. For N1 the maximum complexity has been considerably extended from 0.682 to 0.845 as well as for N3 improving from 0.47 to 0.746. For N1 its margins vary nearly in the same range.

Once the landscape and its complexity described, we proceed to assess different learners using this test bed.

7.7 STANDARD CLASSIFIERS PERFORMING ON THE LANDSCAPE

We perform a test run of the contest using six widely-used classifiers belonging to different learning paradigms: C4.5 [Quinlan, 1995], IBk [Aha et al., 1991], Naive Bayes (NB), PART [Frank and Witten, 1998], Random Tree (RT) [Breiman, 2001], and SMO [Platt, 1999]. All these methods are run using the Weka package with the following configurations: (1) $k = 3$ for IBk and (2) the rest of the parameters are set to their default value. The performance of each technique is evaluated with the test classification accuracy, estimated using stratified ten-fold cross-validation. This section summarises the results of two experiments (1) standard training and (2) test.

7.7.1 *Ten-fold cross-validation.*

Figures 7.3 and 7.4 represent the test accuracy over the complexity measurement space projected to the first two principal components. The x-axis refers to the first principal component and the y-axis refers to the second principal component. The colour bar shows the gradation of the test accuracy; the darker the colour, the lower the accuracy. For clarity of the plots, the accuracies are shown with a truncated scale from 25% to 100%. In Figs. 7.3 and 7.4, we can see the results obtained by C4.5, IB3, NB, PART, RT, and SMO over the data sets generated using the five different seeds. Each column refers to each seed problem, namely, Checkerboard, Pima, Spiral, Wave Boundary, and Yin Yang.

We observe that C4.5 achieves a good performance except for a small group of problems located in the upper left corner, whereas IB3 behaves correctly for only half of the collection. According to the gradation of the accuracy, we believe that this measurement space is able to distinguish to some extent between easy and difficult problems. As all the learners involved in the experimentation fail learning the concept of the data sets located in the upper left corner, we can conclude that those data sets refer to difficult problems. On the other hand, regarding the algorithms based on decision trees, C4.5 and RT behave similarly whereas this pair differs from the PART results. This may indicate that we should relate the data complexity with the knowledge representation used by the learners instead of the learning paradigms.

7.7.2 *Results with the reserved test sets.*

Figures 7.5a, 7.5b, and 7.5c show the training accuracy of C4.5 over S1 and the test accuracies of C4.5 over the collections S2 and S3. These two collections, S2 and S3, are generated based on the same partition of the space used to generate S1. We use a larger random selection of problems in order to match each problem contained in S1 with problems generated with the same seed problem and are of comparable data complexity. Thus, problems contained in S2 and S3 have structurally similar counterparts in S1.

In general, we observe that the accuracies obtained during training remain in the same range as the accuracies attained during testing. However, for problems with similar complexity, the comparative advantages between classifiers are sometimes reversed. The problems seemingly easy in S1 could result in low accuracies in S2 and S3. This means that, for apparently easy problems, the accuracies are less consistent across different sample problems of the same complexity. This leads to a note of caution that data complexity alone is not sufficient to ensure similar classifier performances if the training and testing data may differ structurally. Additional

Table 7.2: Contestant description.

Team 1	*Contestants:*	Joaquín Derrac, Salvador García, and Francisco Herrera
	Affiliation:	Universidad de Granada and Universidad de Jaén
	Contribution:	IFS-CoCo in the landscape contest: Description and results
Team 2	*Contestants:*	Robert P.W. Duin, Marco Loog, Elzbieta Pekalska, and David M.J. Tax
	Affiliation:	Delft University of Technology and University of Manchester
	Contribution:	Feature-based dissimilarity space classification
Team 3	*Contestants:*	Luiz Otávio Vilas Boas Oliveira and Isabela Neves Drummond
	Affiliation:	Universidade Federal de Itajubá
	Contribution:	Real-valued Negative Selection (RNS) for classification task

measures of the structural similarity between training and testing data are needed to project classification accuracy. An extreme example is as follows: data sets with either a vertical linear boundary or a horizontal linear boundary can have the same geometric complexity; but if the learner is trained with a data set containing a vertical boundary and tested on another data set containing a horizontal boundary, the classification accuracy will be low. For data generated from the same seed problem, such large differences in structure are unlikely, but not impossible at local scales, especially when the samples are sparse.

7.8 CONTESTANT CLASSIFIERS PERFORMING ON THE LANDSCAPE

The contest data set was released on March 31, 2010. Initially ten teams indicated their interest in participating in the contest. However, a combination of difficulties caused most teams to drop out over the next few months. At the end, three teams submitted their final results for the entire collection S2 by the June 1st due date, and all of them participated in the live test. This section introduces the contestants and gathers the most important results from the contest.

7.8.1 *Contestants*

Table 7.2 summarizes the information of each team. The approaches they used include (1) a classifier based on a co-evolutionary algorithm [Derrac et al., 2010], (2) a set of classifiers defined in a feature-based dissimilarity space [Duin et al., 2010], and (3) a classifier based on real-valued negative selection [Oliveira and Drummond, 2010].

7.8.2 *Contest description*

The contest was divided into two phases: (1) offline test and (2) live test.

For the offline test, participants ran their algorithms over two sets of problems, S1 and S2, and reported their results. In particular, we assessed (1) the predictive accuracy (i.e. test rate of correctly classified instances) applying a ten-fold cross-validation using S1 and (2) the class labelling of the test collection S2.

A live test took place during the conference. There, collections S3 and S4 were presented. S3 covered the data complexity space comprehensively, like S2, whereas S4 was generated according to the preliminary results submitted by the participants in order to determine the relative merits in each algorithm's respective domain of competence.