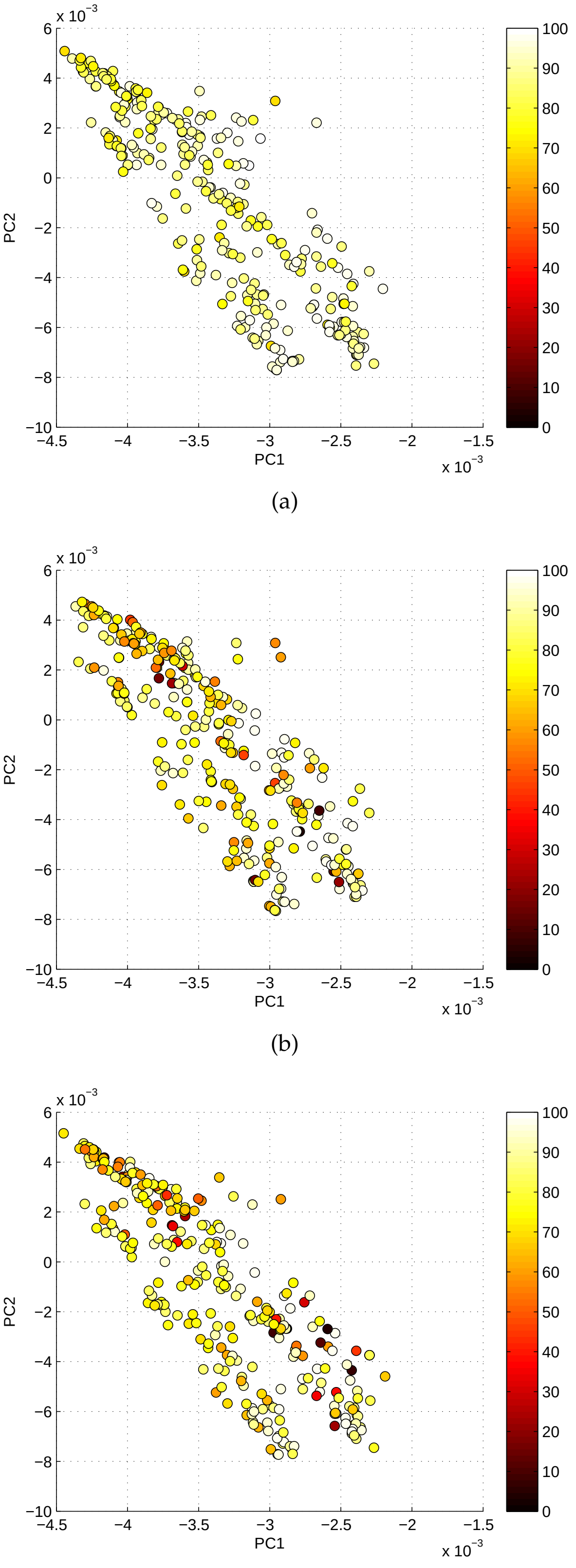
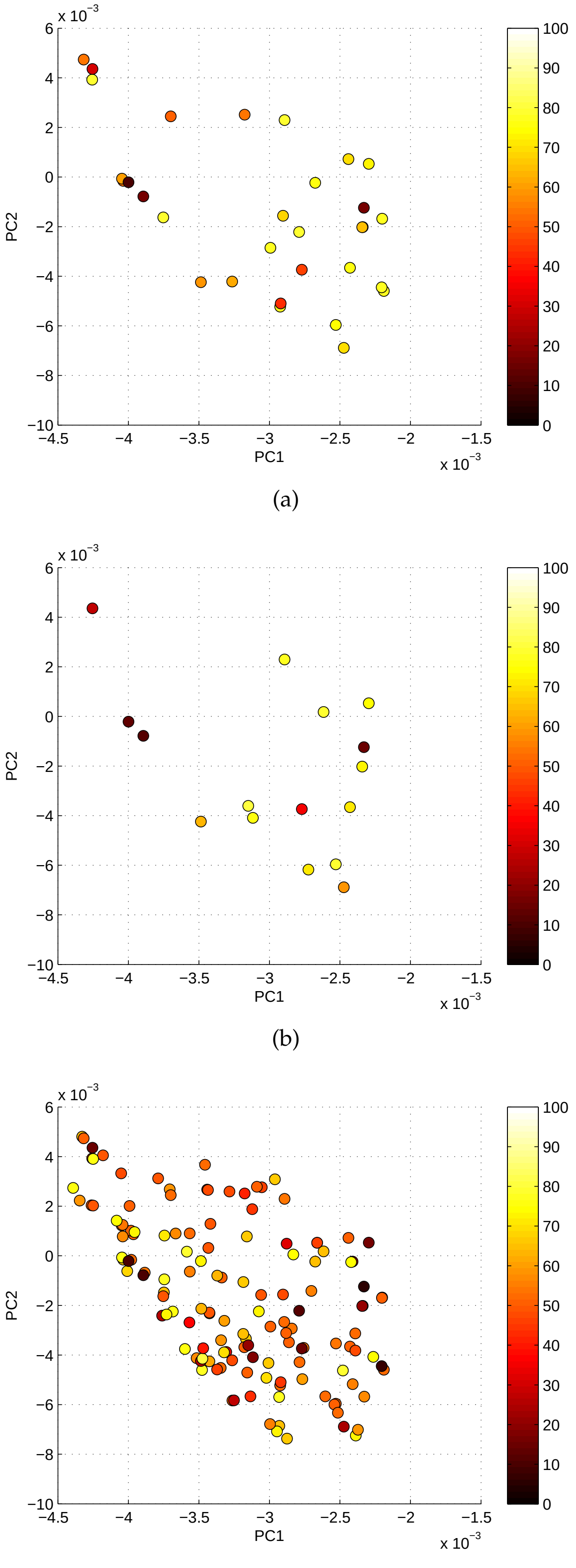


(a)

(b)

(c)

Figure 7.5: C4.5 accuracies over the (a) training collection S1, (b) test collection S2, and (c) test collection S3.



(a)

(b)

(c)

Figure 7.6: Classifier accuracies of the contestant classifiers over the test collection S2: (a) Team 1, (b) Team 2, and (c) Team 3.

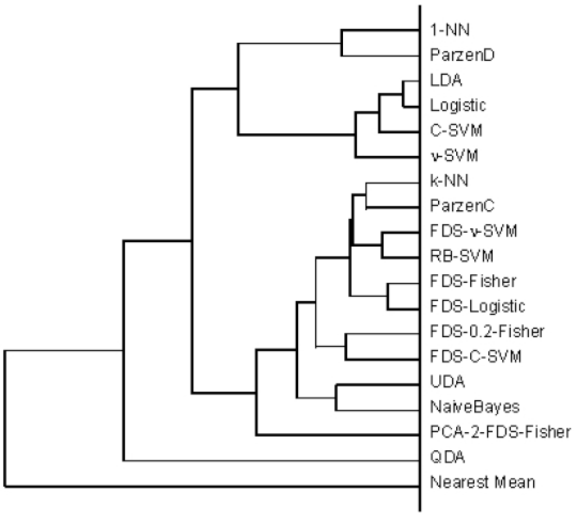


Figure 7.7: Clustering of the different techniques according to their domains of competence from [Duin et al., 2010].

7.8.3 *Analysis of results*

In analysing the contestants' results, we notice that the methods of Team 1 and Team 2 behave similarly. For the majority of the problems they score the same accuracies. In a win/loss comparison over the 300 data sets, Team 1 outperforms the others in 121 problems and Team 2 outperforms in 61 problems. For the other 118 problems, either all three techniques achieve the same score, or just Teams 1 and 2 come to a draw. A paired t-test shows that the difference between Teams 1 and 2 is not statistically significant, whereas their differences with respect to Team 3 are. The accuracies of Team 3 are far below its rivals'; its average accuracy is 76% while the others' are about 92%.

Fig. 7.6 plots the results with the reserved test set S2. For clarity, it plots only those data sets for which the learners achieved accuracies lower than 80%. In general, the number of correctly classified instances is high. Nonetheless, Figs. 7.6a and 7.6b show some spots where the accuracies are extremely low and suggest performing an in-depth study with these specific data sets. Interestingly they are not located in the same region of the complexity space. For both learning paradigms, there is a common set of problems that cause degradations to their performances. These problems, despite belonging to different zones of the space, share the same underlying concept: a wave-shaped boundary (Wave Boundary). For Team 2, a Checkerboard distribution poses some difficulties too. This points out the significant role of the *seeding* learning concept.

Regarding the domains of competence of classifiers, Team 2 applied a cluster analysis to the matrix of all the classification errors of nineteen classifiers from PRTool[2], a Matlab based toolbox that implements a collection of learning techniques for pattern recognition [van der Heijden et al., 2004] over the 301 data sets. Fig. 7.7 depicts the resulting dendrogram, where reasonable relations appear. For instance 1-NN and ParzenC [Parzen, 1962] were linked, which makes sense since both techniques are density estimators. Linear Discriminant Analysis (LDA), logistic and linear SVMs were also put together forming the group of linear classifiers. And the different variant of Feature-based Dissimilarity Space (FDS) were connected to radial basis SVM (RB-SVM) making the group of nonlinear classifiers. This relation confirms the interest of looking into domains of competence and credits the approach taken so far: complexity measures and generation of ADS based on these estimates, which may be insightful in the experimental methodology redefinition.

[2] http://prtools.org/

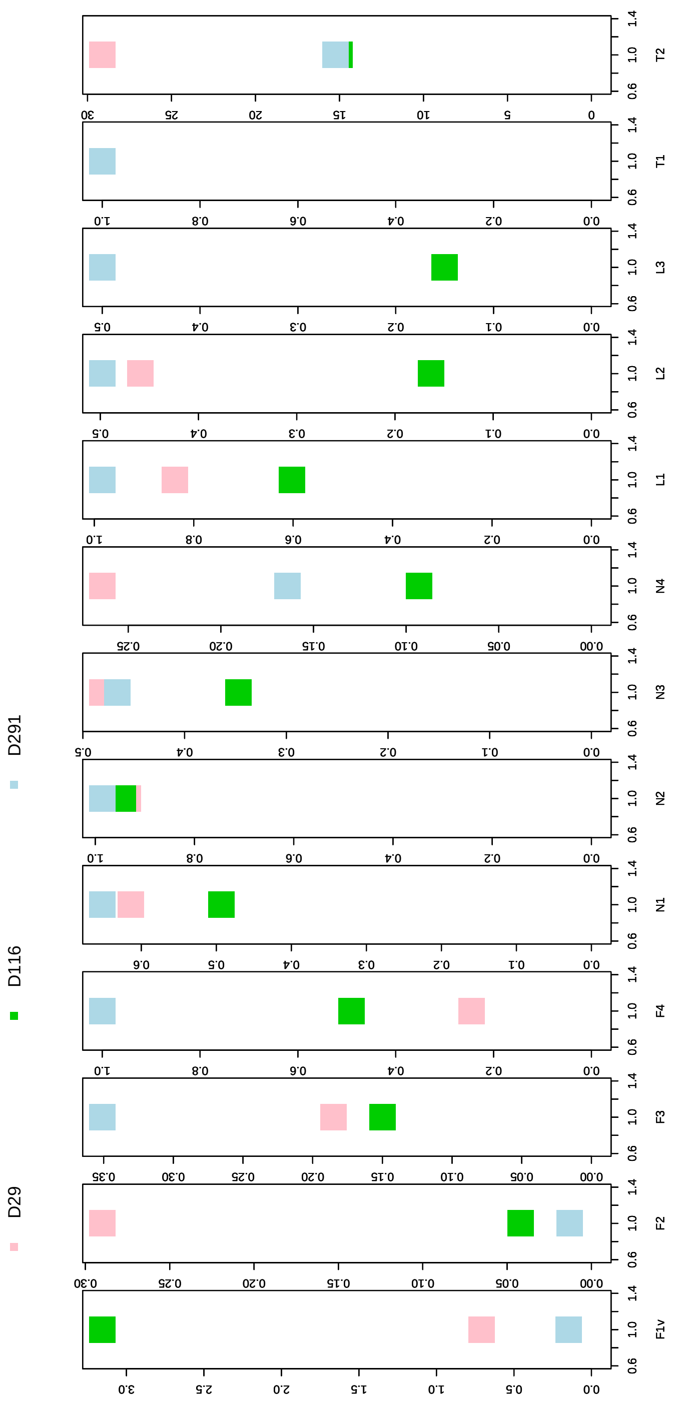


Figure 7.8: Intrinsic complexity of the nineteen favoured data sets.

Table 7.3: Nineteen favoured data sets.

Learner	D242	D47	D200	D168	D116	D291	D180	D100	D298	D82	D183	D171	D292	D286	D97	D5	D29	D24	D218
1NN	**.120**	.328	.083	.250	.350	.321	.049	.235	.611	.302	.074	.166	.049	.101	.164	.530	.526	.416	.060
kNN	.160	**.220**	.175	.254	.173	.272	.046	.126	.425	.300	.056	.166	.049	.101	.128	.513	.379	.296	.047
ParzenC	.144	.254	**.068**	.284	.207	.289	.051	.162	.449	.296	.052	.366	.051	.126	.125	.533	.453	.322	.047
ParzenD	.132	.323	.135	**.222**	.357	.403	.060	.235	.618	.279	.069	.186	.051	.086	.145	.560	.440	.382	.056
Nearest Mean	.277	.427	.425	.401	**.167**	.580	.114	.490	.385	.378	.511	.419	.521	.373	.178	.537	.435	.330	.509
UDA	.179	.267	.099	.274	.213	**.075**	.034	.126	.425	.253	.078	.284	.087	.220	.092	.527	.366	.270	.073
LDA	.177	.263	.310	.252	.197	.557	**.020**	.129	.412	.250	.069	.304	.049	.207	.095	.517	.366	.279	.047
QDA	.170	.272	.120	.277	.280	.164	.071	**.123**	.495	.260	.069	.282	.067	.202	.132	.503	.362	.335	.056
NaiveBayes	.158	.293	.117	.240	.207	.170	.049	.123	**.203**	.265	.069	.294	.056	.188	.115	.507	.371	.305	.043
Logistic	.172	.263	.304	.254	.197	.554	.034	.129	.409	**.236**	.069	.301	.054	.220	.099	.513	.379	.279	.052
FDS-0.2-Fish	.219	.310	.175	.285	.277	.216	.034	.132	.498	.281	**.048**	.337	.067	.200	.102	.550	.440	.352	.047
FDS-Fish	.158	.289	.182	.307	.217	.167	.031	.126	.462	.289	.056	**.133**	.041	.086	.095	.533	.466	.365	.047
FDS-Logistic	.130	.289	.086	.240	.217	.167	.031	.123	.462	.277	.056	.137	**.036**	.081	.095	.543	.466	.365	.047
FDS-C-SVM	.170	.280	.188	.262	.187	.226	.031	.123	.422	.289	.065	.142	.054	**.072**	.089	.493	.457	.343	.043
FDS-ν-SVM	.157	.246	.123	.270	.193	.236	.029	.126	.432	.289	.061	.210	.051	.136	**.066**	.487	.388	.288	.043
PCA2-FDS-Fish	.158	.319	.093	.302	.240	.197	.040	.123	.302	.352	.061	.161	.044	.099	.086	**.120**	.427	.403	.060
C-SVM	.200	.250	.286	.242	.187	.430	.029	.129	.412	.248	.061	.308	.056	.202	.086	.517	**.332**	.292	.043
ν-SVM	.200	.254	.286	.267	.193	.528	.037	.179	.415	.272	.061	.335	.056	.244	.092	.490	.366	**.249**	.043
RB-SVM	.160	.254	.125	.270	.197	.174	.031	.129	.445	.279	.061	.419	.054	.136	.095	.503	.362	.288	**.039**

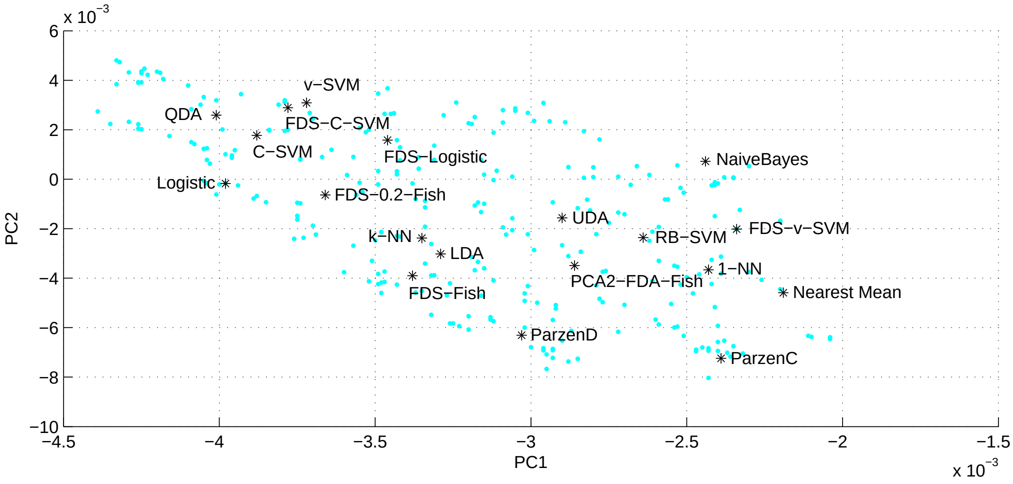


Figure 7.9: Data sets within the collection S1 that are favoured by the nineteen classifiers.

7.9 GOLDEN BENCHMARKS FOR THE ASSESSMENT OF MACHINE LEARNING TECHNIQUES

Duin et al. [2010] documented that among a set of classifiers, each technique found a most favourable data set from the collection S1, all distinct. This section traverse such idea which verifies the representativeness of our artificial sample.

They performed a comparison with nineteen classifiers and observed that each of the nineteen classifiers they tried has a unique data set for which it is the best—a very interesting result that is subject for further study. Table 7.3 gathers the nineteen favoured data sets (highlighted in red), and Fig. 7.8 plots the complexity of three of them (D29 in pink, D116 in green, and D291 in blue); we distinguish very different curves along the complexity measures.

We observe that D116 is judged as a very simple problem since Nearest Mean is the best classifier. In terms of complexity, this problem was generated by minimising the complexity of F2, L2, and L3 over a waved boundary to promote linearity and discriminant attributes. For data set D291, all linear classifiers fail and perform close to random, while UDA (Naive Gaussian) is very good. In deed, the underlying concept in that problem is an spiral—non-linear problem—and the complexity measures based on the SVM are medium-high, especially for L1 $[L1 = 0.984, L2 = 0.498, L3 = 0.500]$. Finally, data set D29 shows a nearly random performance for all classifiers and inspired to Team 2 to include the two-dimensional subspace classifier PCA2-FDS-Fish.

Fig. 7.9 shows the location of these nineteen data sets in the PC1-PC2 space. As we can see, their position is well spread across the space. Fig. 7.10 details the complexity of each data set for each measure. Again, we observe that this set of problems cover a wide region, although the ranges of F2, L2, L3, and F4 only reach medium complexities. Despite not having a complete space, the proposed technique to generate ADS may help to reinforce the use of prototypical data sets to test machine learning techniques.

7.10 DISCUSSION

A big drawback experienced along this research has been the need of resources in terms of storage and computational capacity. We find an implicit combinatorial problem—number of complexity characteristics to evolve and the granularity of the scale of each characteristic—which increases the number of experiments exponentially. However, the performed experiments have shown interesting results that point out the importance of different kind of data sets and also the problem structure. In order to better understand the learners' behaviour or to make some guidelines to choose the right learner any time, we have to address in detail three key points related to the construction of the testing framework: (1) complexity measures, (2) structural

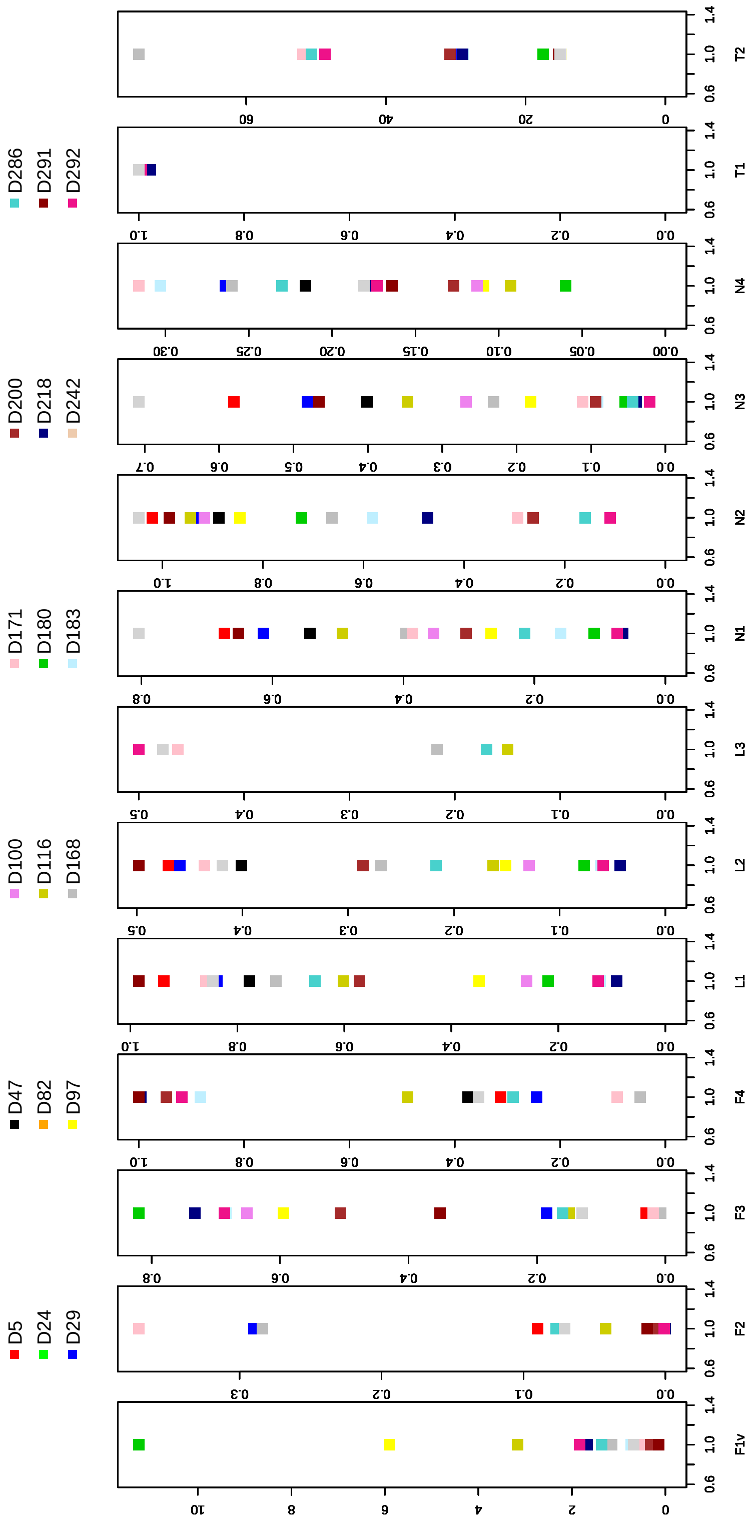
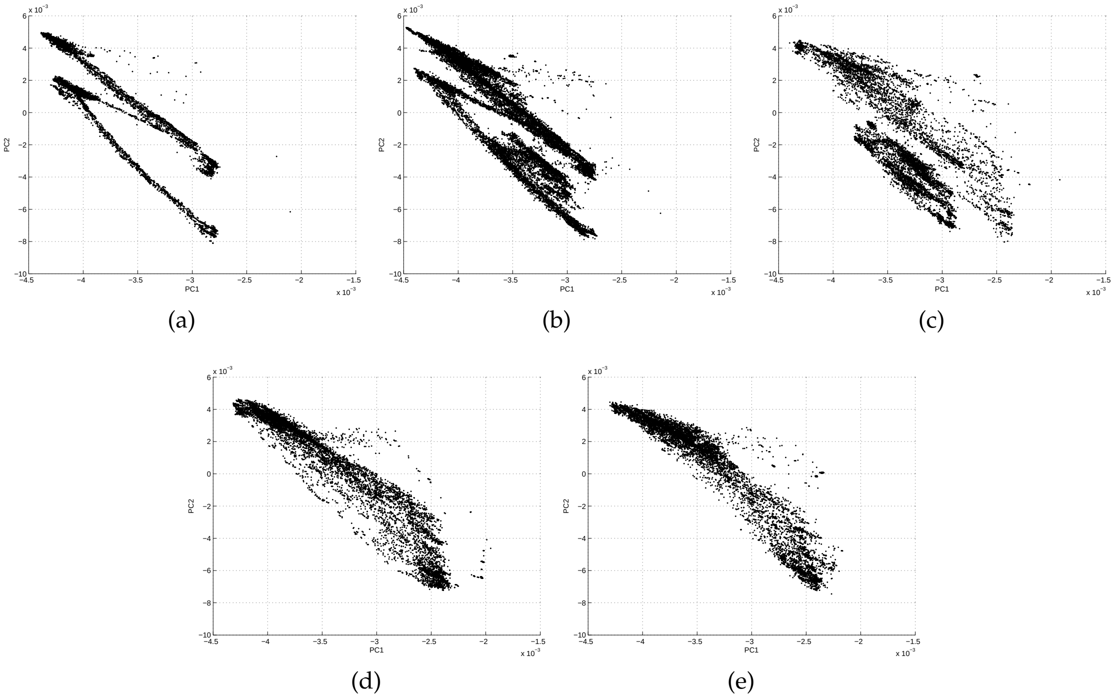


Figure 7.10: Intrinsic complexity of the nineteen favoured data sets.

dimension, and (3) completeness. This section elaborated each one of these aspects and envisage some other issues related to resources.

Complexity measures. The proposed landscape is built using eleven of the fourteen complexity measures. However, it is important to analyse the contribution of each measure and how the problem distribution may be modified depending on the insertion of more complexity measures



(a) (b) (c)

(d) (e)

Figure 7.11: Coverage derived from each seed data set: (a) Checkerboard, (b) Pima, (c) Spiral, (d) Wave Boundary, and (e) Yin Yang.

or the deletion of some. Moreover, it would be interesting to determine whether this space suffices to provide some guidelines that link data characteristics to learner properties, or whether we have to carry out an individual study for each complexity measure.

Structural limits from seeding data. The coverage of the proposed space is based on the difficulty of the problems originating from only five seed data sets each representing a different pair of class concepts. The nature of the seed distributions may influence the resulting testing framework. Fig. 7.11 shows how each seed data set leads to the coverage of a different region of the space, with some overlapping. Further work should be planned to determine the effect of the seed data on the resulting coverage, and whether coverage originating from different seed data would have any significant difference.

Completeness. The two aforementioned aspects lead to the concern about whether and how the completeness of the space could be guaranteed. What is the minimum number of dimensions needed to fully represent the difficulty of a problem? Which of these dimensions are most suitable? What would be the proper seed data that have the furthest reach over the space?

7.11 DANGER OF REPOSITORIES AND BENCHMARKS

Although public repositories such as the UCI repository have contributed to the maturation of experimental methodology, we have seen that there is no foundation to rely on such collection of problems to make strong claims, the sample is not representative at all. Actually, many of these data sets need to be preprocessed which is not indicated in the majority of contributions and, consequently, wrongly considered as standards and a basis for fair learner comparisons. The details of the data cleaning process of the UCI repository is found in [Macià and Bernadó-Mansilla, 2011]. On the other hand, there is a misuse/abuse of these data sets involved in comparisons just because they provide matrix of values—forgetting about the actual interest: the extraction of useful knowledge or hidden patterns. Then, synthetic generation appears a reliable mechanism and the previous step to consolidate new benchmarks for the community.

However, compulsion for outstanding in performance based on benchmark comparison can suppress creative work. This section warns about the risks of repositories and benchmarks.

Machine learning is meant to be the solution for mining huge amount of data or few amount of data really complex. But, during the development of techniques, a switch in the prime goal takes place. The validation of the techniques, the ones that have to solve challenging problems, is made over well-known test problems that have few attributes, few instances, few classes, simple boundaries, regular structures... No real peculiarities, i.e. peculiarities that strongly influence the classifiers, are tested. This style of experimentation results in learners performing well—overfitting—in a toy sample and on top of that accidentally biased the designer from the very beginning—just focused on passing the exam. Obviously, we are not against benchmarks since we have proposed an approach to generate ADS and leads us to the creation of benchmarks. We pursue a change of mentality and update the experimental methodology for machine learning techniques. This methodology would be aligned with the ideas proposed in [Prechelt, 1994], the tuning of the methods should be done on training and once the optimal measure is obtained, measure the performance of testing—this is basically the cross-validation method. However, he suggested not to have previous knowledge of the data sets to use for validation. Besides, we suggest to use a set of synthetic problems to test the correctness of the technique and its limitations, and then validate with another test of problems with real-world structures.

Sadly, this methodology is doomed to fail because any attempt of benchmark has been suffocated by the inertia of the UCI repository. Therefore, we believe that any change has to be promoted from this well-known repository. This means that the UCI repository should present two sections: (1) synthetic problems, labelled with their corresponding complexity for testing learners' limitations, and (2) real-world problems to validate algorithms.

Contribution.

1. Proposal of a landscape composed of artificial data sets.
2. Study of real-world and synthetic problems.
3. Analysis of learners' behaviours over the landscape.
4. International contest.
5. Proposal of the data sets landscape as benchmarks for learner assessment.

8 SUMMARY, CONCLUSIONS, AND FUTURE WORK

Summary. This chapter ends the thesis with a summary of the work conducted, some conclusions, and future directions.

8.1 INTRODUCTION

This thesis has provided some insights into data complexity and the generation of artificial data sets through complexity estimates. The exploration of data complexity is a thriving field that can enrich machine learning progress and complement data mining.

The purpose of this chapter is to summarise the work of this thesis and conclude with some remarks and future work.

In the following, we wrap up the three main achievements—DCoL, Generator of Data Sets (GoDS), and the adjustment of the experimental methodology—and share some observations regarding the aim of research on machine learning. Are we looking for deep knowledge on cognitive systems or just implementing intelligent tools to perform daily tasks? In addition, we raise some challenges for further work on data complexity that basically require a global project.

8.2 SUMMARY

Over the last few decades, the machine learning community has designed and developed techniques to solve real-world problems and to extract knowledge from data. To validate the efficiency of these techniques, the most usual methodology adopted by practitioners consists in testing new techniques over 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 learners. This section resumes the work done to address this concern.

Learners are usually tested using real-world problems from public repositories. Even though sharing these problems benefits the obtention of a common test bed for the experiments and facilitates the comparison between the individual researchers' results and the community's, 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 characterised by few instances, missing values, and imprecise data. The combination of these deficiencies in the data samples goes beyond our control, blurring our knowledge of to what extent the influence of these constraints negatively affects learner performance. Thus, data complexity analysis is advised to build a complexity space that offers a set of problems adequate to test learners' limitations and validate their performance. Table 8.1 summarises the contributions of this thesis, where the main goals have been to:

1. Provide a common implementation of complexity measures: DCoL.
2. Artificially generate new data on the ground of complexity estimates to reach a good coverage of the problem space: GoDS.
3. Modify the experimental methodology by including data complexity analysis and testing-validation over benchmarks from synthetic generation and public repositories.

Table 8.1: Summary of the scientific contributions of this thesis.

Chapter	Main contributions
Chapter 2	Criticism of experimental assessment of learners based on arbitrary selected data sets.
Chapter 3	DCoL, update and implementation of the complexity measures to provide the research community with a common base.
Chapter 4	Recommender system based on complexity measures.
Chapter 5	Characterisation of the UCI and PSP repository.
Chapter 6	Taxonomy of complexity descriptors. GoDS, generation of ADS based on evolutionary multi-objective optimisation using complexity measures.
Chapter 7	Collection of 80,000 synthetic problems: the landscape. Identification of preliminary benchmarks.

In the following, we shortly describe each attainment.

DCoL. Ho and Basu [2002] proposed a set of geometrical descriptors to estimate the complexity of the class boundary. It is not easy to find a test by which to judge whether a complexity space is reasonable but we can at least ascertain that this set of measures can be a start. For this reason, we have revised the definition of each complexity measure and implemented an open source library [Orriols-Puig et al., 2010]. This provides a common implementation to perform further fair analyses.

GoDS. Real-world problems with truth labels are expensive to obtain and difficult to control. This, plus the lack of complexity resolution of real-world problems, has promoted the quest for search procedures to generate artificial data sets with a specific complexity. In [Macià et al., 2010b], we have proposed a multi-objective approach that can create synthetic problems whose structure resemble real-world problems and meet different criteria of complexity guided by the studied measures. This method has been utilised to build a diverse collection of problems and compare learners' abilities on different regions of the complexity space [Macià et al., 2010a]. Indeed, synthetic data sets are an interesting alternative; created with the least cost and time, they can be controlled to reach a good coverage of regions in the complexity space. Though, to employ this alternative, it is important to ensure that the generated data have sufficient resemblance to real-world problems and, at the same time, contain sufficient variety to represent different aspects of data complexity.

Experimental methodology. The two previous implementations connect to the experimental methodology and may be introduced as components of the experimental testing. We have proposed to use them to generate data to test the limitations of the learners and validate their performance.

On linking data complexity to classifier performance, we cannot claim that higher complexity necessarily leads to lower classifier performance—this relationship is far more complicated. Yet, steps towards characterising the competence domains of each classifier family is unhesitatingly believed to be an improvement over the current state-of-the-art and an advantage over the trial-and-error processes typically followed in practice.

8.3 CONCLUSIONS

Machine learning inspired by human mechanisms may be an incredible way to simulate an artificial scenario to explore our cognitive system. Nonetheless, this anthropological approach has fallen apart—at least in fundamental research—and has exploded into diverse tasks such as optimisation, prediction-classification, pattern identification, etc. So now, the major concern is *for what kind of problems does a new learning technique work well?* The answer should not lay in an infinite combinatorial parametrisation of algorithms using a specific, small set of problems but instead in the scientific progress. Note that, in the literature, many techniques are claimed to be a universal approximator of an arbitrary decision function. It is the belief of the creators that the techniques will work well in all kinds of problems. For those who hold these claims, the above question does not even exist. Therefore, this question needs to be preceded by providing counter-evidence to such claims. This section supports critical contributions and recommends researchers to step back and reconsider old proposals to reinvent the experimental methodology. Furthermore, the advances in the fields are not only on the methodology but also on the restating of the goals.

This thesis has followed the criticism of Salzberg [1997], whose message is to be exceedingly careful extracting conclusions regarding learners' performance when mining large databases composed of different problems. However, he mentions that the theoretical limitation launched by Wolpert [1992, 1996]—simple learners equally perform, on average, over the whole space of possible problems—may not be relevant if we just focus on a portion of the problem space, for instance only on real-world problems. Although our main purpose has been to try to offer a better coverage and manipulate any kind of complexity, such a reflection steers us to address daily problems and concentrate our efforts on designing goal-oriented techniques instead of trying to father the universal learner. In fact, although global differences may be minor, local differences matter a lot as solutions for particular problems require the best. This view matches the studies based on domains of competence and particular problems.

Experimental methodology has been overshadowed by the widespread heavy focus on implementation of algorithms. Once statistical tests have been included in the evaluation process, the validity of the methodology has been taken for granted. However, the current methodology remains quite rudimentary and original proposals have been mostly ignored. Jensen [1991] argued that machines should be built to compensate human weaknesses instead of duplicating human strengths. According to him, tools would not automate model generation, they would allow investigators to generate and guide the basic structure of new models themselves—this is their strength since they are the experts. Alternatively, tools would automate the testing of new models and assess the statistical significance—a human weakness. To this end, randomisation testing was proposed to create a large number of random data. By destroying the relationship between attributes and class, i.e. by reversing or changing the class of instances, he obtained a large battery of randomised data that formed enough distributions to reflect whether the scores of the best learner could be expected only by chance. With this in mind, we should reexamine our proposal in Macià et al. [2009], which varies the labelling within the problem until reaching the desired complexity while maintaining the values of the attributes.

We remind the reader, though, that we embarked upon a doubtful space since complexity measures are still under study. For this reason, there is large room for improvement with respect to the statement of common methodology to evaluate learning techniques.

8.4 FUTURE WORK

There is still some work to do on consolidating the use of complexity estimates. This section points out some research lines and projects based on the improvement of, and cross-fertilising, ideas. We present the future work following the chapter structure.

Chapter 3. Surely, some researchers are not convinced of the practicability of such estimates because they believe *that complexity of a classification task cannot be nicely captured by single numbers* or because data complexity computation cost is equal to the runtime of a considerably large battery of learners. In this regard, the DCoL has to be improved in terms of (1) efficiency (see some directions in Chapter 3) and (2) interpretability by attaching some explanations to the data characterisation.

It is also a pending task to extend the measure to deal with multi-instance multi-label problems, or at least to determine how to proceed to estimate the complexity of such kinds of problems.

Chapter 6. Our approach is based on evolutionary computation; a technique with a high computational cost and which is difficult to parametrise. It is necessary to perform a more detailed analysis to determine the best genetic operators and their configuration in order to speed up the convergence of the algorithm. Alternatives to evolutionary computation such as meta-heuristics to guide the search in the complexity space are appealing.

Chapter 7. Despite these issues, it is important to gain visibility as soon as possible and involve as many people as possible in the development of complexity measures to establish a benchmark environment. For this purpose, it would be great to start a global project which aim to study the failure of so many data generators and would-be benchmarks released, and developing a unique site that gathers all of these proposals revised under certain parameters—again, this utopian project collides with the difficulty to stop local research. The site would be a parameterisable test suite containing data sets with a broad set of characteristics to check how the learning methods react to each different scenario. It would help to reduce the amount of experimentation by proposing the right test bed. Indisputably, the interest would be to offer a collection of benchmarks and recommend which ones to use to test specific behaviours. It is still the responsibility for researchers to not limit their creativity and to not just design algorithms to fulfil the requirement of the existing test bed.

On the other hand, it seems desirable to exploit the aspect of neutral complexity estimators. In many contributions of machine learning to real-world applications, the data of the problems solved are not available, most likely because of non-disclosure agreements or because the difficulty of the task cannot be judged by the reviewers and readers. In such a case, we could use DCoL to give an abstract description of their problem so that the reader can appreciate the results.

Finally, after the experimental results, we acknowledge the need of providing the link between data complexity and learners' behaviours with a theoretical, formal framework.

EPILOGUE

PHILOSOPHICAL DOUBTS...

Summary. This chapter does something unusual: it examines the current research practices and casts doubt on the meaning of researchers' task in machine learning. Experimentation is a valuable tool in this field. The beloved random experimentation is, however, a danger. Inspired by the French writer Rousseau—*society corrupts the individual* [Rousseau, 1972]—, the conviction of this chapter is that the scientific community corrupts the researcher. Unwritten rules prevent the new generation of Ph.D. students from being creative thinkers.

INTRODUCTION

The scientific method has evolved throughout the centuries, and philosophers have had a distinguished role in that change since they have been questioning the beliefs and truths used to found discoveries. They have been challenging the veracity of theories, arguing against new ways of thinking, sometimes without providing any answer, for the pleasure of asking questions [Russell, 1997]. Unfortunately, during the research journey, many students have missed that interest, passion, awakened mind. The current computer science community, especially the empiricists, write large amounts of plain technical reports tracing experiments, oblivious to the beauty of essays, the excitement of sharing revolutionary, outlandish ideas. Going through the literature has often become a mechanical skimming/scanning task, seeking for the bold numbers that highlight those few decimals that the proposed techniques outperform their competitors by. These results sustain a sort of research, somehow, based on a mere, casual parameter tuning of established techniques.

The purpose of this chapter is to show the disenchantment that, sometimes or often, would-be researchers—and even some tenured researchers—suffer and to denounce the proliferation of some questionable practices that are killing innovation.

In the following, we briefly review the effect of the modern obsession for publishing and to what extent academic research has distorted the meaning of experimental science. We see what calls the current methodology for assessing learners into question and how different alternatives have simply been ignored or incomprehensibly revived at a later time. Finally, we end with a reminder of what the purpose of Ph.D. studies could be in the midst of such confusion.

THE PERVERSION OF THE COMMUNITY

Pure research is becoming less attractive nowdays. Many research groups have abandoned some research lines since investors are more interested in applications—in spite of the relevance of essential, fundamental investigation. Hence, the ones that have managed to subsist are because either their research is leading in an applicative domain or their volume of publication is *high*. What is behind these numbers? Parnas [2007] makes a strong point about them and guts, one by one, every single perversion of the scientific community—authorship in pacts, monthly instalments, tailor-made conferences or workshops—that have encouraged superficial research made from overly large groups, repetition, small and insignificant studies, half-baked ideas, and so on. This section describes how the famous *publish or perish* has wreaked havoc on daily research, at least in the halls of European academia.

Eventually, h-index, impact factors, or the number of citations are the fallacious indicators of good research, of good researchers. Thus, fresh Ph.D. students, willing to know about research, methodology, values, are burdened and frustrated junk writers after a couple of months, since they have learnt from the feelings in their labs that their career will be measured by these statistics, whose actual interpretation is: write as many papers as possible. The only way to fulfil such requirements is not to work for the long-term run at all, but to dissolve the research and present every partial, experimental result—it is dramatic to see these students in a rush for publishing when they have not even fully experienced research yet; and it is even worse when this pressure comes from senior researchers who are pushing so hard in this direction just to keep their CVs up-to-date or repay colleagues in the favour chain which promotes quantity over substance. This tradition of compulsive publishing has plagued conferences, journals, and the Internet with so many papers that it is getting difficult to track innovative ideas. The more one reads, the more one bumps into similar attempts, similar flavours, déjà vus—facts that slows down the learning curve and discourages further reading. This is the price to pay for the democratisation of research, which raises a sharp debate between top-notch research from recognisable researchers and results from obligatory projects that allow young people to participate [Duin, 2011]. Undoubtedly, high level research is still done. Good papers are still written, but they are hidden in vast amount of less relevant works from Ph.D projects. Showing off the abilities of regular methods to non-technical experts and cherry-picking results from much wider experimentation are the most common schemes. Both serve to bridge theory and practice. However, the function of empiricism has been abused and now entails repeated preliminary results with no further continuation.

machine learning: experimental science

Experimental computer science, defined as "an apparatus to be measured, a hypothesis to be tested, and systematic analysis of the data (to see whether its supports the hypothesis)" by Denning [1980], is recurrent in machine learning, algorithms, and software engineering. Nevertheless, experimental methodology has been twisted; instead of sustaining conjectures, experiments are run to provide material to decide them retroactively, to give birth to a posteriori theories or feed useless, partial conclusions. Machine learning, for instance, is based on trials with performance measures, learners, and data. The combination of these elements made Langley [1988] encourage practitioners to join empirical testing, as it contributes to the process of theory formation. This section comments on the subsequent chaos of such a call: competition testing—a term coined by Hooker [1995] in relation to heuristics.

Many years later, no new learning paradigm has been introduced, some progress in standards has been made, and micro-tuning of the existing techniques is the trendy research. Superiority of techniques is shown usually following a three-step procedure: select data sets, typically from public repositories, choose specific referenced learners to compare the new approach with, and extract performance conclusions supported by erroneous statistical tests. With a pessimistic but very realistic description of the current scene, Demšar [2008] has been able to cope with the concerns about the meaning of such experimentation and the way it should be. Conventional statistical models are designed to test single learners in isolation; they are ill-suited to perform multiple comparisons. Hypothesis testing is useful to say whether the probability of the apparent accuracy of a learner is only due to chance, but its power goes down as the number of models examined increases. Then, it is worth determining what the ideal size of the test set is, what problems have to be involved, and empowering the testing methodology by sufficient data analysis. These—old claims—are the kind of things that one expects to be delighted with when reading papers. Yet, there are complicated milestones and many negative results are derived from the studies. Although these are meaningful to lead progress as well, the community does not consider them. This forces researchers to move back to the classical developments—safe bets to reach the approval. In addition, groundless rejections cause frustration in new researchers, which is reflected in their subsequent reviews. In turn, after being taught that going against the current is not fruitful, they will be unwittingly stopping promising ideas, frustrating new generations again.

GAMING THE SYSTEM IN LIEU OF RESEARCH

Gascuel and Caraux [1992] gave an early indication that the Bonferroni-type approximation was adequate to compare results obtained from applying different systems on the same—or different—test samples. In a lapse of twenty years, the t-test, despite being the wrong method, has been the most popular. Why? Because of trends and because results looked statistically stronger; because of trends. This section shares a view about the clout of journals and reviewers and the inertia of the machine learning community as a society.

Current research is like politics—each tendency has its own press. No matter the thoroughness of the content, if the work submitted to a journal is not aligned with the thought of its staff, it will never get the green light. In his inspiring four-page paper Demšar [2008] suggests the (im)possible solution: web-to-peer review. This unlikely idea seems to offer favourable conditions for critical and fair evaluations of "correctness, interestingness, usefulness, beauty, novelty". Solution or not, this evidences the urge to adopt other measures of productivity and recognition to end with the fake tenure of rigour, truths, and biased opinions. The new peer-review process should give back credibility to publications, and researchers should not be able to game it. Dietz et al. [2007] studied the citation influences. In this work, the following paragraph may catch ones eye: "Papers can be cited as background, reading for politeness, fear of receiving an adverse review from an aggrieved reviewer, or as related work that was argued against." Indeed, the analysis of references is interesting from a social point of view and has a crucial role in the shallow statistics—impact factors and indexes. Everyone knows they provide the information for the productivity computation. Thus, self-citations, citations to friends and the community clique, or citations to publications from particular journals are some of the mechanisms to scale. Citing has lost its traditional purpose: guiding the reader to obtain the background necessary to understand the paper and enlarge his knowledge with valuable contributions; valuable knowledge for the basis of Ph.D. studies.

SETTING ANY THESIS IN CONTEXT

Pressure for publishing is also destabilising the framework of Ph.D. studies and deflecting them from their goal. This section tries to find its place nowadays.

The objective of any thesis is to master a subject—not randomly investigate. When applying for some grants, one of the ever-present sections is to justify the relevance of the topic and the profit of further achievements for society. Mostly, thesis investigations in computer science are too specialised, of interest to tiny communities. In general, it is about impractical idealism that just serves well in the ivory tower, not in the real world. For this reason, the community should accept unusual ideas with an open mind or significant experimental results to gain soundness across different knowledge areas. Saitta and Langley's words should be taken up again: machine learning is much more than running algorithms on apparent *ready-to-use* data sets [Saitta and Neri, 1998]. The machine learning discipline emerged to gain an insight into complex tasks such as reasoning, problem solving, and language processing [Langley, 2011]. On the other hand, the true essence of Ph.D. studies, and by extension of fundamental research, is to keep curiosity alive, to learn how to become a researcher, and to push boundaries. Basically, Ph.D. students are still in training to become the elite of technical advances, and this is what people seem to have forgotten. So putting aside the expertise, Ph.D. students should acquire other skills to be a great asset to industry too. Therefore, there is a lot to do in Ph.D. programmes to hone communication skills, critical assessment ability, leadership... under an intensive level of advisory—more than just passive supervision. This means that supervisors have to be dedicated to at most a few students at a time, and a generous amount of hours have to be booked into their work load to closely follow the evolution of the would-be researchers, to consolidate their knowledge, and remind them of the existence and importance of ethics. Researchers should publish their work only when it is mature, relevant enough, and the community can benefit from it. The rate of publications has to be slowed down in order to gain in quality.

APPENDIX

A

TABLES

Summary. This appendix contains all the large tables that summarise the detailed experimentation analysed in Chapter 5.

Table A.1: Description of the external characteristics of the UCI collection. Data sets are alphabetically sorted. *#Cl* is the number of classes, *#Inst* is the number of instances, and *#Att* is the number of attributes. *#Real*, *#Int* and *#Nom* indicate the number of real-, integer- and nominal-valued attributes respectively. *%missInst*, *%missAtt*, and *%missVal* corresponds to the percentage of instances with missing values, attributes with missing values, and the total percentage of missing values. Finally, *%Maj* is the percentage of instances of the majority class and *%Min* is the percentage of instances of the minority class.

Dataset	#Inst	#Att	#Real	#Int	#Nom	%missAtt	%mistInst	%missVal	%Maj	%Min
aba.2co	4177	8	7	0	1	0.00	0.00	0.00	99.98	0.02
adl	48842	14	0	6	8	21.43	7.41	0.95	76.07	23.93
ann.2co	898	38	6	0	32	0.00	0.00	0.00	99.11	0.89
asbestos	83	3	0	1	2	0.00	0.00	0.00	55.42	44.58
aud.2co	226	69	0	0	69	10.14	98.23	2.03	78.76	21.24
aut.2co	205	25	15	0	10	28.00	22.44	1.15	98.54	1.46
authors.2co	841	70	0	70	0	0.00	0.00	0.00	62.31	37.69
bal.2co	625	4	4	0	0	0.00	0.00	0.00	53.92	46.08
bankruptcy	50	5	5	0	0	0.00	0.00	0.00	50.00	50.00
benford.2co	9	6	0	0	6	0.00	0.00	0.00	88.89	11.11
bondrate.2co	57	10	0	4	6	10.00	1.75	0.18	89.47	10.53
bpa	345	6	6	0	0	0.00	0.00	0.00	57.97	42.03
briv1.2co	105	11	0	4	7	63.64	33.33	5.28	84.76	15.24
briv2.2co	105	11	0	1	10	63.64	33.33	5.28	84.76	15.24
car.2co	1728	6	0	0	6	0.00	0.00	0.00	70.02	29.98
cmc.2co	1473	9	0	2	7	0.00	0.00	0.00	57.30	42.70
col	368	22	7	0	15	95.45	98.10	23.80	63.04	36.96
crx	690	15	3	3	9	53.33	5.51	0.66	55.51	44.49
drm.2co	366	34	0	33	1	2.94	2.19	0.06	69.40	30.60
ech	74	11	6	2	3	72.73	17.57	2.83	67.57	32.43
euc.2co	736	19	7	7	5	47.37	12.91	3.20	75.54	24.46
flg.2co	194	28	0	10	18	0.00	0.00	0.00	79.38	20.62
fourclass	862	2	2	0	0	0.00	0.00	0.00	64.39	35.61
gls.2co	214	9	9	0	0	0.00	0.00	0.00	67.29	32.71
gru.2co	155	8	0	2	6	0.00	0.00	0.00	68.39	31.61
h-s	270	13	13	0	0	0.00	0.00	0.00	55.56	44.44
hab	306	3	0	3	0	0.00	0.00	0.00	73.53	26.47
hay.2co	159	4	0	0	4	0.00	0.00	0.00	59.75	40.25
hep	155	19	2	4	13	78.95	48.39	5.67	79.35	20.65
hrs	368	27	8	4	15	77.78	98.10	19.39	63.04	36.96
ion	351	34	34	0	0	0.00	0.00	0.00	64.10	35.90
irs.2co	150	4	4	0	0	0.00	0.00	0.00	66.67	33.33
krk.2co	28056	6	0	0	6	0.00	0.00	0.00	90.03	9.97
krkp	3196	36	0	0	36	0.00	0.00	0.00	52.22	47.78
liv	345	6	1	5	0	0.00	0.00	0.00	57.97	42.03
lng.2co	32	56	0	0	56	3.57	15.63	0.28	71.88	28.13
mag	19020	10	10	0	0	0.00	0.00	0.00	64.84	35.16
msh	8124	22	0	0	22	4.55	30.53	1.39	51.80	48.20
nrs.2co	12960	8	0	0	8	0.00	0.00	0.00	66.67	33.33
opt.2co	5620	64	0	64	0	0.00	0.00	0.00	90.14	9.86
pas.2co	36	22	15	6	1	0.00	0.00	0.00	66.67	33.33
pbc.2co	5473	10	4	6	0	0.00	0.00	0.00	89.77	10.23
pen.2co	10992	16	0	16	0	0.00	0.00	0.00	89.60	10.40
pim	768	8	8	0	0	0.00	0.00	0.00	65.10	34.90
pos.2co	90	8	0	1	7	12.50	3.33	0.42	97.78	2.22
seg.2co	2310	19	19	0	0	0.00	0.00	0.00	85.71	14.29
shs.2co	52	24	6	15	3	29.17	3.85	0.56	55.77	44.23
shu.2co	52	23	5	15	3	34.78	17.31	3.26	53.85	46.15
spa	4601	57	55	2	0	0.00	0.00	0.00	60.60	39.40
spect	267	22	0	0	22	0.00	0.00	0.00	79.40	20.60
spectf	267	44	0	44	0	0.00	0.00	0.00	79.40	20.60

TableA.1 – continued from previous page

Dataset	#Inst	#Att	#Real	#Int	#Nom	%missAtt	%mistInst	%missVal	%Maj	%Min
statlog-sgm.2co	2310	19	19	0	0	0.00	0.00	0.00	85.71	14.29
tae.2co	151	5	0	1	4	0.00	0.00	0.00	67.55	32.45
tao	1888	2	2	0	0	0.00	0.00	0.00	50.00	50.00
thy.2co	215	5	5	0	0	0.00	0.00	0.00	69.77	30.23
tis	13375	927	924	0	3	0.00	0.00	0.00	75.24	24.76
tnc	2201	3	0	0	3	0.00	0.00	0.00	67.70	32.30
trn	10	32	0	10	22	31.25	70.00	15.94	50.00	50.00
veh.2co	846	18	18	0	0	0.00	0.00	0.00	74.94	25.06
vot	435	16	0	0	16	100.00	46.67	5.63	61.38	38.62
wav21.2co	5000	21	21	0	0	0.00	0.00	0.00	66.86	33.14
wav40.2co	5000	40	40	0	0	0.00	0.00	0.00	66.16	33.84
wbcd	699	9	0	9	0	11.11	2.29	0.25	65.52	34.48
wdbc	569	30	30	0	0	0.00	0.00	0.00	62.74	37.26
whi.2co	63	31	27	0	4	0.00	0.00	0.00	60.32	39.68
win.2co	178	13	11	2	0	0.00	0.00	0.00	66.85	33.15
wne.2co	178	13	13	0	0	0.00	0.00	0.00	66.85	33.15
wpbc	198	33	33	0	0	3.03	2.02	0.06	76.26	23.74
yea.2co	1484	8	8	0	0	0.00	0.00	0.00	68.80	31.20
zoo.2co	101	16	0	0	16	0.00	0.00	0.00	59.41	40.59

Table A.2: Complexity measures on the UCI repository.

Data set	F1	F1v	F2	F3	F4	L1	L2	L3	N1	N2	N3	N4	T1	T2
aba.2co	14.032	33.605	0.000	1.000	1.000	0.002	2.394e-04	0.500	4.788e-04	0.026	2.394e-04	0.000	0.648	522.125
adl	0.346	2.725	0.233	0.030	0.035	1.065	0.166	0.275	0.290	0.446	0.201	0.269	0.996	3488.714
ann.2co	2.526	59.494	0.000	0.728	1.000	0.099	0.009	0.500	0.007	0.135	0.001	0.021	0.960	23.632
asbestos	0.316	1.549	0.885	0.012	0.012	0.581	0.193	0.229	0.398	0.562	0.229	0.199	0.602	27.667
aud.2co	0.022	16.957	0.000	0.407	0.872	0.462	0.212	0.500	0.235	0.607	0.119	0.175	1.000	3.275
aut.2co	2.061	14.036	0.000	0.985	1.000	0.034	0.015	0.500	0.049	0.285	0.020	0.010	1.000	8.200
authors.2co	1.877	19.663	3.858e-13	0.225	0.998	0.413	0.002	5.945e-04	0.012	0.772	0.004	5.945e-04	1.000	12.014
bal.2co	0.385	0.418	1.000	0.000	0.000	0.570	0.048	0.066	0.184	0.623	0.138	0.101	0.902	156.250
bankruptcy	1.121	1.673	0.001	0.620	0.960	0.922	0.160	0.220	0.240	0.632	0.140	0.090	0.920	10.000
benford.2co	73.143	9.130	0.000	1.000	1.000	0.236	0.111	0.500	0.556	0.831	0.333	0.444	1.000	1.500
bondrate.2co	0.904	1.386	9.068e-05	0.474	1.000	0.213	0.105	0.500	0.228	0.813	0.158	0.167	1.000	5.700
bpa	0.055	0.148	0.073	0.032	0.107	0.841	0.420	0.500	0.574	0.913	0.374	0.342	1.000	57.500
briv1.2co	3.234	57.275	0.000	0.714	0.714	0.377	0.019	0.014	0.095	0.264	0.048	0.005	1.000	9.545
briv2.2co	0.415	44.742	0.000	0.714	0.714	0.395	0.010	0.024	0.076	0.379	0.048	0.014	1.000	9.545
car.2co	0.024	1.308	0.250	0.333	0.556	0.857	0.136	0.153	0.234	0.902	0.175	0.472	1.000	288.000
cmc.2co	0.029	0.289	0.750	0.002	0.002	0.788	0.356	0.383	0.583	0.877	0.406	0.370	0.944	163.667
col	0.305	12.965	0.187	0.038	0.098	0.561	0.073	0.030	0.201	0.689	0.114	0.030	1.000	16.727
crx	0.364	8.962	0.001	0.032	0.068	0.290	0.145	0.154	0.290	0.563	0.181	0.117	0.999	46.000
drm.2co	7.789	3.031	0.000	0.541	0.989	0.323	0.000	0.000	0.016	0.443	0.005	0.000	0.981	10.765
ech	4.221	15.976	0.000	0.973	1.000	0.362	0.054	0.014	0.135	0.433	0.095	0.000	0.973	6.727
euc.2co	2.722	16.211	3.990e-07	0.170	0.471	0.559	0.132	0.217	0.192	0.336	0.113	0.158	0.999	38.737
flg.2co	1.545	4.211	0.000	0.093	0.232	0.444	0.206	0.500	0.294	0.770	0.175	0.157	1.000	6.929
fourclass	0.952	1.372e-04	0.649	0.239	0.278	0.720	0.215	0.314	0.009	0.091	0.000	0.229	0.452	431.000
gls.2co	0.649	0.579	3.648e-05	0.290	0.486	0.671	0.327	0.500	0.322	0.432	0.182	0.227	0.991	23.778
gru.2co	3.240	1.388	0.486	0.077	0.123	0.653	0.316	0.500	0.497	0.825	0.342	0.300	0.994	19.375
hab	0.185	0.002	0.718	0.029	0.033	0.530	0.265	0.500	0.539	0.754	0.353	0.368	0.931	102.000
hay.2co	0.025	0.686	0.296	0.082	0.195	0.805	0.403	0.500	0.384	0.442	0.258	0.450	1.000	39.750
hep	1.040	6.249	0.000	0.232	0.568	0.467	0.194	0.490	0.284	0.623	0.187	0.026	0.987	8.158
hrs	0.305	17.615	0.000	0.038	0.141	0.546	0.073	0.037	0.193	0.684	0.106	0.024	1.000	13.630
h-s	0.760	4.344	0.196	0.015	0.093	0.532	0.156	0.104	0.367	0.672	0.244	0.107	1.000	20.769
ion	0.614	3.728	0.000	0.191	0.994	0.453	0.117	0.150	0.231	0.631	0.131	0.162	0.946	10.324
irs.2co	16.822	25.477	0.005	0.573	0.573	0.310	0.000	0.000	0.013	0.096	0.000	0.000	0.453	37.500
krk.2co	0.011	0.053	1.000	0.000	0.000	0.200	0.100	0.500	0.545	0.846	0.462	0.487	1.000	4676.000
krkp	0.002	6.688	0.000	0.183	0.433	0.753	0.062	0.084	0.273	0.714	0.156	0.264	1.000	88.778
liv	0.055	0.148	0.073	0.032	0.107	0.841	0.420	0.500	0.574	0.913	0.374	0.322	1.000	57.500
lng.2co	0.655	18.432	0.000	0.281	0.969	0.546	0.281	0.500	0.469	0.962	0.375	0.000	1.000	0.571
mag	0.559	58.946	0.081	0.006	0.018	0.717	0.214	0.233	0.293	0.650	0.188	0.236	1.000	1902.000
msh	0.038	11.452	0.000	0.201	0.419	0.491	0.043	0.068	0.003	0.377	0.000	0.033	1.000	369.273

TableA.2 – continued from previous page

Data set	F1	F1v	F2	F3	F4	L1	L2	L3	N1	N2	N3	N4	T1	T2
nrs.2co	0.002	9.722	0.500	1.000	1.000	0.667	0.000	0.000	1.543e-04	1.000	7.716e-05	0.000	1.000	1620.000
opt.2co	4.729	0.050	0.000	0.471	0.865	1.685	0.002	0.003	7.117e-04	0.408	1.779e-04	0.003	0.991	87.812
pas.2co	1.628	15.391	1.093e-06	0.583	0.972	0.637	0.333	0.500	0.306	0.646	0.194	0.028	1.000	1.636
pbc.2co	0.511	176.800	2.100e-05	0.016	0.026	0.319	0.095	0.484	0.068	0.184	0.037	0.213	0.988	547.300
pen.2co	3.475	0.001	0.755	0.109	0.120	1.973	0.014	0.049	0.001	0.163	3.639e-04	0.019	0.899	687.000
pim	0.576	1.414	0.252	0.007	0.022	0.689	0.350	0.500	0.438	0.840	0.294	0.289	0.999	96.000
pos.2co	0.172	2.138	0.000	0.478	0.733	0.045	0.022	0.500	0.056	0.415	0.033	0.500	1.000	11.250
seg.2co	1.798	18.828	8.627e-14	0.734	1.000	0.484	0.143	0.500	0.010	0.114	0.003	0.038	0.933	121.579
shs.2co	1.127	-3.500e+00	2.705e-05	0.327	1.000	0.642	0.192	0.163	0.327	0.789	0.154	0.183	1.000	2.167
shu.2co	6.788	9.850	0.005	0.250	0.885	0.748	0.404	0.423	0.519	0.899	0.288	0.125	1.000	2.261
spa	0.347	12.078	2.533e-33	0.091	0.383	0.772	0.394	0.500	0.155	0.376	0.082	0.113	0.986	80.719
spect	0.017	2.114	0.000	0.142	0.210	0.580	0.206	0.500	0.337	0.785	0.270	0.476	1.000	12.136
spectf	0.553	6.159e-04	3.603e-19	0.296	0.993	0.423	0.206	0.500	0.337	0.803	0.251	0.099	1.000	6.068
statlog-sgm.2co	1.798	11.204	8.627e-14	0.734	1.000	0.484	0.143	0.500	0.010	0.114	0.003	0.037	0.933	121.579
tae.2co	6.316	0.814	0.559	0.079	0.146	0.653	0.325	0.500	0.450	0.378	0.199	0.427	0.960	30.200
tao	1.395	0.031	0.479	0.360	0.362	0.590	0.163	0.110	0.077	0.157	0.043	0.155	0.398	944.000
thy.2co	0.256	0.543	0.001	0.186	0.414	0.588	0.302	0.500	0.102	0.310	0.028	0.151	0.837	43.000
tis	0.472	7.410	0.000	0.033	0.234	1.415	0.070	0.084	0.345	0.853	0.260	0.013	1.000	14.428
tnc	-1[1]	0.754	1.000	0.000	0.000	0.448	0.224	0.304	0.682	0.033	0.677	0.500	1.000	733.667
trn	18.000	4.734	0.000	0.500	0.500	0.706	0.000	0.000	0.400	0.832	0.100	0.000	1.000	0.312
veh.2co	0.185	1.006	5.339e-04	0.037	0.223	0.504	0.251	0.500	0.365	0.712	0.248	0.353	0.999	47.000
vot	0.066	28.884	1.000	0.000	0.000	0.354	0.037	0.010	0.115	0.348	0.064	0.008	1.000	27.188
wav21.2co	1.182	0.150	0.036	0.123	0.183	1.020	0.141	0.086	0.238	0.795	0.170	0.109	1.000	238.095
wav40.2co	1.168	0.130	0.015	0.149	0.211	1.000	0.143	0.081	0.273	0.900	0.196	0.062	1.000	125.000
wbcd	3.568	0.068	0.248	0.119	0.232	0.457	0.034	0.012	0.059	0.335	0.041	0.030	0.801	77.667
wdbc	3.405	56.726	5.683e-11	0.517	0.998	0.539	0.049	0.022	0.072	0.558	0.047	0.022	0.998	18.967
whi.2co	2.188	6.654	1.019e-06	0.222	0.968	0.687	0.381	0.492	0.492	0.948	0.333	0.063	1.000	2.032
win.2co	4.290	22.879	3.962e-05	0.764	1.000	0.371	0.056	0.028	0.067	0.490	0.028	0.003	0.994	13.692
wne.2co	4.290	22.879	3.962e-05	0.764	1.000	0.371	0.056	0.031	0.067	0.490	0.028	0.022	0.994	13.692
wpbc	0.472	4.823	1.422e-06	0.177	0.990	0.485	0.237	0.500	0.424	0.914	0.278	0.217	1.000	6.000
yea.2co	0.214	0.736	0.056	0.127	0.177	0.624	0.312	0.500	0.450	0.710	0.296	0.340	1.000	185.500
zoo.2co	0.344	86.450	0.000	0.802	0.802	0.142	0.000	0.000	0.020	0.200	0.000	0.000	1.000	6.312

[1] Value -1 indicates that the complexity measure could not be calculated. For instance, it can happen for F1v if the algorithm for diagonalising the bi-diagonal form loops over the singular values and does not converge before reaching the maximum number of iterations.

Table A.3: Complexity measures on the PSP repository.

Data set	F1	F1v	F2	F3	F4	L1	L2	L3	N1	N2	N3	N4	T1	T2
aa-uf-2c-w0	0.001	0.005	1.000	0.000	0.000	0.975	0.511	0.493	0.386	-nan	0.386	0.486	1.000	257560.000
aa-uf-2c-w1	0.004	0.005	1.000	0.000	0.000	0.975	0.511	0.493	0.435	0.320	0.418	0.486	1.000	85853.336
aa-uf-2c-w2	0.004	0.009	1.000	0.000	0.000	0.976	0.497	0.486	0.535	0.947	0.422	0.483	1.000	51512.000
aa-uf-2c-w3	0.004	0.013	1.000	0.000	0.000	0.975	0.486	0.478	0.571	0.972	0.421	0.483	1.000	36794.285
aa-uf-2c-w4	0.004	0.016	1.000	0.000	0.000	0.974	0.476	0.469	0.561	0.981	0.422	0.482	1.000	28617.777
aa-uf-2c-w5	0.004	0.024	1.000	0.000	0.000	0.969	0.469	0.459	0.576	0.985	0.422	0.480	1.000	23414.545
aa-uf-2c-w6	0.004	0.027	1.000	0.000	0.000	0.967	0.465	0.455	0.571	0.987	0.421	0.477	1.000	19812.309
aa-uf-2c-w7	0.004	0.028	1.000	0.000	0.000	0.966	0.465	0.454	0.582	0.988	0.421	0.477	1.000	17170.666
aa-uf-2c-w8	0.004	0.028	1.000	0.000	0.000	0.965	0.464	0.452	0.582	0.989	0.423	0.474	1.000	15150.588
aa-uf-2c-w9	0.004	0.029	1.000	0.000	0.000	0.964	0.463	0.449	0.588	0.990	0.424	0.472	1.000	13555.789
aa-ul-2c-w0	0.002	5.295e-04	1.000	0.000	0.000	0.599	0.299	0.500	0.304	-nan	0.304	0.493	1.000	257560.000
aa-ul-2c-w1	0.004	8.899e-04	1.000	0.000	0.000	0.599	0.299	0.500	0.374	0.178	0.356	0.493	1.000	85853.336
aa-ul-2c-w2	0.004	0.003	1.000	0.000	0.000	0.599	0.299	0.500	0.467	0.909	0.363	0.492	1.000	51512.000
aa-ul-2c-w3	0.004	0.005	1.000	0.000	0.000	0.599	0.299	0.500	0.494	0.951	0.361	0.491	1.000	36794.285
aa-ul-2c-w4	0.004	0.010	1.000	0.000	0.000	0.599	0.299	0.500	0.489	0.969	0.361	0.489	1.000	28617.777
aa-ul-2c-w5	0.004	0.017	1.000	0.000	0.000	0.599	0.299	0.500	0.499	0.974	0.360	0.487	1.000	23414.545
aa-ul-2c-w6	0.004	0.021	1.000	0.000	0.000	0.599	0.299	0.500	0.497	0.978	0.360	0.484	1.000	19812.309
aa-ul-2c-w7	0.004	0.023	1.000	0.000	0.000	0.599	0.299	0.500	0.506	0.980	0.362	0.483	1.000	17170.666
aa-ul-2c-w8	0.004	0.025	1.000	0.000	0.000	0.599	0.299	0.500	0.504	0.982	0.361	0.480	1.000	15150.588
aa-ul-2c-w9	0.004	0.025	1.000	0.000	0.000	0.599	0.299	0.500	0.508	0.984	0.363	0.478	1.000	13555.789
pssm-uf-2c-w0	0.466	0.096	0.877	9.318e-05	1.009e-04	0.825	0.263	0.209	0.451	0.914	0.318	0.312	1.000	12878.000
pssm-uf-2c-w1	0.466	0.110	0.688	9.318e-05	1.553e-04	0.855	0.253	0.195	0.425	0.903	0.302	0.283	1.000	4292.667
pssm-uf-2c-w2	0.466	0.120	0.575	9.318e-05	1.864e-04	0.869	0.248	0.190	0.412	0.891	0.295	0.227	1.000	2575.600
pssm-uf-2c-w3	0.466	0.127	0.513	9.318e-05	1.941e-04	0.890	0.241	0.181	0.405	0.885	0.290	0.181	1.000	1839.714
pssm-uf-2c-w4	0.466	0.152	0.457	9.318e-05	2.019e-04	0.903	0.235	0.173	0.402	0.883	0.291	0.156	1.000	1430.889
pssm-uf-2c-w5	0.466	0.375	0.359	9.318e-05	2.640e-04	0.906	0.234	0.173	0.402	0.884	0.295	0.143	1.000	1170.727
pssm-uf-2c-w6	0.466	0.405	0.320	9.318e-05	2.718e-04	0.909	0.233	0.171	0.405	0.886	0.301	0.133	1.000	990.615
pssm-uf-2c-w7	0.466	0.169	0.285	9.318e-05	2.795e-04	0.910	0.232	0.171	0.408	0.889	0.310	0.129	1.000	858.533
pssm-uf-2c-w8	0.466	0.270	0.236	9.318e-05	3.223e-04	0.913	0.231	0.170	0.413	0.892	0.321	0.125	1.000	757.529
pssm-uf-2c-w9	0.466	0.090	0.197	9.318e-05	3.572e-04	0.913	0.231	0.170	0.414	0.895	0.325	0.121	1.000	677.787
pssm-ul-2c-w0	0.550	0.090	0.944	3.883e-06	3.883e-06	0.779	0.233	0.286	0.391	0.881	0.274	0.337	1.000	12878.000
pssm-ul-2c-w1	0.550	0.104	0.694	2.330e-05	8.153e-05	0.846	0.216	0.260	0.351	0.881	0.242	0.303	1.000	4292.667
pssm-ul-2c-w2	0.550	0.113	0.534	5.047e-05	1.902e-04	0.858	0.211	0.250	0.336	0.872	0.232	0.236	1.000	2575.600
pssm-ul-2c-w3	0.550	0.116	0.476	5.047e-05	1.980e-04	0.884	0.203	0.237	0.327	0.869	0.226	0.185	1.000	1839.714
pssm-ul-2c-w4	0.550	0.118	0.394	5.047e-05	2.524e-04	0.907	0.197	0.227	0.323	0.869	0.226	0.148	1.000	1430.889
pssm-ul-2c-w5	0.550	0.129	0.309	5.047e-05	3.028e-04	0.910	0.197	0.224	0.324	0.872	0.229	0.135	1.000	1170.727
pssm-ul-2c-w6	0.550	0.181	0.222	5.047e-05	4.426e-04	0.915	0.196	0.223	0.322	0.875	0.232	0.123	1.000	990.615
pssm-ul-2c-w7	0.550	0.132	0.186	6.212e-05	5.009e-04	0.921	0.195	0.222	0.324	0.878	0.238	0.124	1.000	858.533
pssm-ul-2c-w8	0.550	0.153	0.134	6.212e-05	6.134e-04	0.922	0.194	0.221	0.329	0.883	0.250	0.117	1.000	757.529
pssm-ul-2c-w9	0.550	0.139	0.098	6.212e-05	7.260e-04	0.924	0.194	0.221	0.315	0.886	0.241	0.112	1.000	677.789

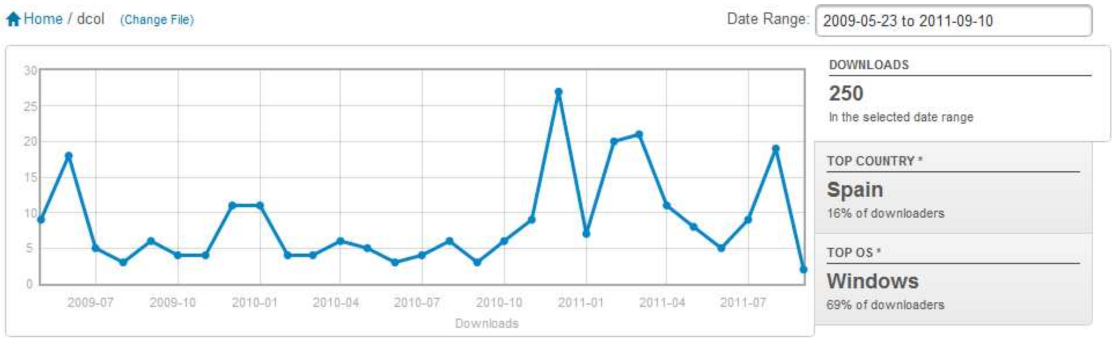
Table A.4: Complexity measures on the Checkerboard collection.

Data set	F1	F1v	F2	F3	F4	L1	L2	L3	N1	N2	N3	N4	T1	T2
Checkerboard4x4-I1000-R	0.0093	0.0707	0.9213	0.0091	0.0819	0.6710	0.9996	0.4968	0.2041	0.9443	0.4472	0.4300	1.0000	48.6000
Checkerboard4x4-I1000-STD0.1	0.0098	0.0752	0.0895	0.0092	0.0867	0.6591	1.0007	0.5020	0.2387	0.9773	0.4708	0.4658	1.0000	48.6000
Checkerboard4x4-I1000-STD0.2	0.0096	0.0739	0.0909	0.0091	0.0870	0.6597	1.0009	0.5031	0.2380	0.9779	0.4761	0.4728	1.0000	48.6000
Checkerboard4x4-I1000-STD0.3	0.0095	0.0737	0.0864	0.0093	0.0883	0.6597	1.0008	0.5029	0.2397	0.9774	0.4715	0.4655	1.0000	48.6000
Checkerboard4x4-I1000-STD0.4	0.0099	0.0715	0.0856	0.0094	0.0883	0.6579	1.0003	0.5013	0.2419	0.9771	0.4701	0.4681	1.0000	48.6000
Checkerboard4x4-I1000-STD0.5	0.0099	0.0677	0.0861	0.0094	0.0887	0.6606	1.0006	0.5032	0.2425	0.9772	0.4655	0.4638	1.0000	48.6000
Checkerboard4x4-I1000-STDR	0.0089	0.0617	0.0890	0.0090	0.0843	0.6549	0.9984	0.4897	0.2430	0.9807	0.4738	0.4692	1.0000	48.6000
Checkerboard4x4-I5000-R	0.0018	0.0136	0.9842	0.0019	0.0157	0.6695	0.9990	0.4975	0.3042	0.9763	0.4745	0.4700	1.0000	248.4000
Checkerboard4x4-I1000-STD0.1	0.0026	0.0147	0.1616	0.0019	0.0166	0.6506	0.9979	0.4920	0.3307	0.9829	0.4766	0.4723	1.0000	248.4000
Checkerboard4x4-I5000-STD0.2	0.0023	0.0146	0.1640	0.0018	0.0161	0.6488	0.9978	0.4898	0.3289	0.9826	0.4761	0.4725	1.0000	248.4000
Checkerboard4x4-I5000-STD0.3	0.0022	0.0156	0.1646	0.0017	0.0162	0.6499	0.9980	0.4912	0.3292	0.9816	0.4753	0.4706	1.0000	248.4000
Checkerboard4x4-I5000-STD0.4	0.0023	0.0159	0.1650	0.0018	0.0161	0.6504	0.9978	0.4916	0.3287	0.9811	0.4754	0.4735	1.0000	248.4000
Checkerboard4x4-I000-STD0.5	0.0025	0.0154	0.1536	0.0017	0.0162	0.6531	0.9978	0.4942	0.3302	0.9817	0.4747	0.4690	1.0000	248.4000
Checkerboard4x4-I5000-STDR	0.0018	0.0125	0.1679	0.0018	0.0169	0.6531	0.9981	0.4956	0.3286	0.9841	0.4768	0.4746	1.0000	248.4000
Checkerboard4x4-I5000-R	0.0014	0.0070	0.9917	0.0010	0.0076	0.6730	1.0005	0.5020	0.3404	0.9833	0.4828	0.4791	1.0000	498.6000
Checkerboard4x4-I10000-STD0.1	0.0011	0.0069	0.1781	0.0010	0.0081	0.6481	0.9961	0.4861	0.3619	0.9865	0.4830	0.4798	1.0000	498.6000
Checkerboard4x4-I10000-STD0.2	0.0010	0.0069	0.1965	0.0010	0.0080	0.6467	0.9958	0.4846	0.3608	0.9866	0.4833	0.4806	1.0000	498.6000
Checkerboard4x4-I10000-STD0.3	0.0011	0.0069	0.2085	0.0010	0.0079	0.6487	0.9958	0.4852	0.3596	0.9866	0.4837	0.4797	1.0000	498.6000
Checkerboard4x4-I10000-STD0.4	0.0011	0.0077	0.1985	0.0010	0.0080	0.6508	0.9965	0.4877	0.3600	0.9854	0.4832	0.4780	1.0000	498.6000
Checkerboard4x4-I10000-STD0.5	0.0012	0.0074	0.1894	0.0010	0.0080	0.6496	0.9963	0.4869	0.3614	0.9856	0.4835	0.4801	1.0000	498.6000
Checkerboard4x4-I10000-STDR	0.0008	0.0059	0.1792	0.0010	0.0083	0.6475	0.9960	0.4846	0.3631	0.9878	0.4839	0.4814	1.0000	498.6000

B
DCOL STATS

Summary. This appendix describes some tracked statistics about the implemented data complexity library.

The DCoL was released in March, 2009 on SourceForge.net. Since then, the library has accumulated 250 downloads as shown in Fig. B.1. However, it was not properly publicised until December, 2010 when version 1.1 was released and when the site registered its maximum activity—in ten months it reached 129 downloads.


Home / dcol (Change File)

Date Range: 2009-05-23 to 2011-09-10

DOWNLOADS
250
In the selected date range

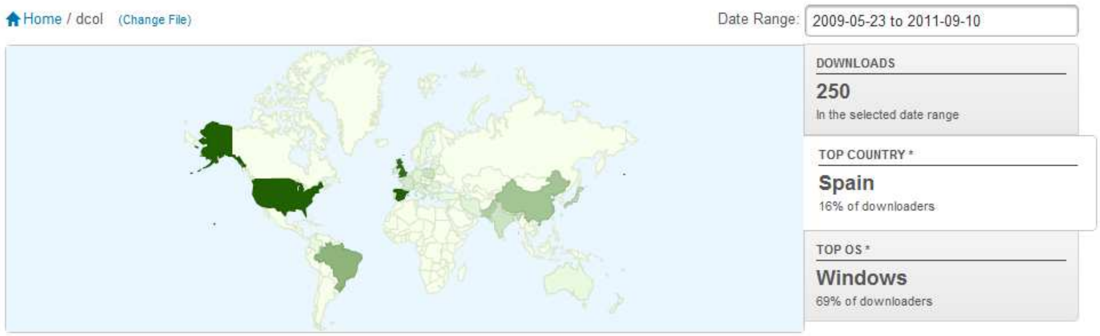
TOP COUNTRY *
Spain
16% of downloaders

TOP OS *
Windows
69% of downloaders

Figure B.1: DCoL downloads.

Fig. B.2 profiles the user demographic. In the second period of the DCoL life, the top five user countries are:

1.	Spain	22 downloads
2.	United States	22 downloads
3.	United Kingdom	21 downloads
4.	Brazil	13 downloads
5.	Pakistan	9 downloads


Home / dcol (Change File)

Date Range: 2009-05-23 to 2011-09-10

DOWNLOADS
250
In the selected date range

TOP COUNTRY *
Spain
16% of downloaders

TOP OS *
Windows
69% of downloaders

Figure B.2: DCoL user demographic.

BIBLIOGRAPHY

Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6(1):37–66.

Alcalá-Fdez, J., Fernández, A., Luengo, J., Derrac, J., García, S., Sánchez, L., and Herrera, F. (2011). Keel data-mining software tool: Data set repository and integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3):255–287.

Alcalá-Fdez, J., Sánchez, L., García, S., del Jesus, M. J., Ventura, S., Garrell, J. M., Otero, J., Romero, C., Bacardit, J., Rivas, V. M., Fernández, J. C., and Herrera, F. (2008). KEEL: A software tool to assess evolutionary algorithms to data mining problems. *Soft Computing*, 13(3):307–318.

Bacardit, J. (2007). *GAssist Source Code: http://www.asap.cs.nott.ac.uk/~jqb/PSP/GAssist-Java.tar.gz*.

Bacardit, J. and Krasnogor, N. (2006). Biohel: Bioinformatics-oriented hierarchical evolutionary learning. Technical report, University of Nottingham.

Bacardit, J. and Krasnogor, N. (2008). The infobiotics PSP benchmarks repository.

Bader, D. and Cong, G. (2006). Fast shared-memory algorithms for computing the minimum spanning forest of sparse graphs. *Journal of Parallel and Distributed Computing*, 66.

Basu, M. and Ho, T. K., editors (2006). *Data complexity in pattern recognition*. Springer.

Baumgartner, R. and Somorjai, R. L. (2006). Data complexity assessment in undersampled classification of high-dimensional biomedical data. *Pattern Recognition Letters*, 27(12):1383–1389.

Bernadó-Mansilla, E. (2004). Complejidad del aprendizaje y muestreo de ejemplos en sistemas clasificadores. In *Proceedings del III Congreso Español de Metaheurísticas, Algoritmos Evolutivos y Bioinspirados*, pages 203–210.

Bernadó-Mansilla, E. and Ho, T. K. (2005). Domain of competence of XCS classifier system in complexity measurement space. *IEEE Transactions on Evolutionary Computation*, 9(1):82–104.

Bernadó-Mansilla, E., Ho, T. K., and Orriols-Puig, A. (2006). Data complexity and evolutionary learning: Classifier's behavior and domain of competence. In *Data Complexity in Pattern Recognition*, pages 115–134. Springer.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Brazdil, P., ao Gama, J., and Henery, B. (1994). Characterizing the applicability of classification algorithms using meta-level learning. In *Proceedings of the European Conference on Machine Learning*, pages 83–102.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brewer, E. A. (2000). Towards robust distributed systems. In *Proceedings of the nineteenth annual ACM symposium on Principles of distributed computing*, page 7.

Casillas, J., Orriols-Puig, A., and Bernadó-Mansilla, E. (2008). Toward evolving consistent, complete, and compact fuzzy rule sets for classification problems. In *3rd International Workshop on Genetic and Evolving Fuzzy Systems*, pages 89–94.

Chang, J., Luo, J., Huang, J. Z., Feng, S., and Fan, J. (2010). Minimum spanning tree based classification model for massive data with mapreduce implementation. *2010 IEEE International Conference on Data Mining Workshops*, pages 129–137.

Coello, C. A., Lamont, G. B., and Veldhuizen, D. A. V. (2006). *Evolutionary algorithms for solving multi-objective problems (Genetic and evolutionary vomputation)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Dean, J. and Ghemawat, S. (2010). Mapreduce: A flexible data processing tool. *Communications of the ACM*, 53(1):72–77.

Deb, K. D., Pratap, A., Agarwal, S., and Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

Demšar, J. (2008). On the appropriateness of statistical tests in machine learning. In *3rd Workshop on Evaluation Methods for Machine Learning*.

Denning, P. (1980). What is experimental computer science? *Communication of ACM*, 23:543–544.

Derrac, J., García, S., and Herrera, F. (2010). IFS-CoCo in the landscape contest: Description and results. In *ICPR 2010*, volume 6388 of *Lecture Note in Computer Science*, pages 56–65. Springer.

Devroye, L. (1988). Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(4):530–543.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning*, pages 233–240.

Dong, M. and Kothari, R. (2003). Feature subset selection using a new definition of classifiability. *Pattern Recognition Letters*, 24(9-10):1215 – 1225.

Duda, R. O., Hart, P. E., and Stork, D. G. (2000). *Pattern classification*. Wiley-Interscience, 2nd edition.

Duin, R. P. (2011). Personal communication.

Duin, R. P. W., Loog, M., Pekalska, E., and Taz, D. M. J. (2010). Feature-based dissimilarity space classification. In *ICPR 2010*, volume 6388 of *Lecture Note in Computer Science*, pages 46–55. Springer.

Duin, R. P. W. and Pekalska, E. (2006). Object representation, sample size, and data set complexity. In *Data Complexity in Pattern Recognition*, pages 25–47. Springer.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, 56(263):52–64.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Ferri, C., Hernández-Orallo, J., and Modroiu, R. (2009). An experimental comparison of performance measures for classification. *Pattern Recognition Letters*, 30:27–38.

Fisher, R. A. (1959). *Statistical methods and scientific inference*. Oliver and Boyd, Edinbrugh, 2nd edition.

Frank, A. and Asuncion, A. (2010). UCI machine learning repository.

Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. In *Proceedings of the 15th International Conference on Machine Learning*, pages 144–151. Morgan Kaufmann, San Francisco, CA.

Friedman, J. H. and Rafsky, L. C. (1979). Multivariate generalizations of the wald-wolfowitz and smirnov two-sample tests. *Annals of Statistics*, 7(7):697–717.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, 32:675–701.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *Annals of Mathematical Statistics*, 11:86–92.

Gamberger, D. and Lavrac, N. (1997). Conditions for occam's razor applicability and noise elimination. In *Proceedings of the 9th European Conference on Machine Learning*, pages 108–123, London, UK. Springer-Verlag.

García, S., Cano, J. R., Bernadó-Mansilla, E., and Herrera, F. (2009). Diagnose of effective evolutionary prototype selection using an overlapping measure. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(8):1527–1548.

García, S. and Herrera, F. (2008). An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. *Journal of Machine Learning Research*, 9:2677–2694.

Gascuel, O. and Caraux, G. (1992). Statistical significance in inductive learning. In *Proceedings of the 10th European Conference on Artificial Intelligence*, pages 435–439.

Giraud-Carrier, C. and Martinez, T. (1995). An efficient metric for heterogeneous inductive learning applications in the attribute-value language. In *Intelligent Systems (Proceedings of GWIC'94)*, pages 341–350. Kluwer Academic Publishers.

Goldberg, D. E. (2002). *The design of innovation: Lessons from and for competent genetic algorithms*. Kluwer Academic Publishers, 1 edition.

Ho, T. K. (1995). Random decision forests. In *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, pages 278–282.

Ho, T. K. (1998). Nearest neighbors in random subspaces. In *Proceedings of the 2nd International Workshop on Statistical Techniques in Pattern Recognition*, pages 640–648.

Ho, T. K. (2001). Data complexity analysis for classifier combination. In *Proceedings of the 2nd International Workshop on Multiple Classifier Systems*, pages 53–67.

Ho, T. K. and Basu, M. (2002). Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300.

Ho, T. K., Basu, M., and Law, M. (2006). Measures of geometrical complexity in classification problems. In *Data complexity in pattern recognition*, pages 1–23. Springer.

Hoag, J. E. and Thompson, C. W. (2007). A parallel general-purpose synthetic data generator. *ACM SIGMOD*, 36:19–24.

Hoekstra, A. and Duin, R. P. W. (1996). On the nonlinearity of pattern classifiers. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 4, pages 271–275, Washington, DC, USA. IEEE Computer Society.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11:63–90.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75(2):383–386.

Hooker, J. N. (1995). Testing heuristics: We have it all wrong. *Journal of Heuristics*, 1:33–42.

Hughes, E. J. (2005). Evolutionary many-objective optimisation: many once or one many? In *IEEE Congress on Evolutionary Computation*, pages 222–227.

Ishibuchi, H., Tsukamoto, N., and Nojima, Y. (2008). Evolutionary many-objective optimization: A short review. In *IEEE Congress on Evolutionary Computation*, pages 2419–2426.

Japkowicz, N. and Stephen, S. (2002). The class imbalance problem: A systematic study. *Intelligence Data Analysis*, 6(5):429–449.

Jensen, D. (1991). Knowledge discovery through induction with randomization testing. In *Proceedings of the 1991 Knowledge Discovery in Databases Workshop*, pages 148–159.

Jeske, D. R., Samadi, B., Lin, P. J., Ye, L., Cox, S., Xiao, R., Younglove, T., Ly, M., Holt, D., and Rich, R. (2005). 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.

Jiménez-Peris, R., no Martínez, M. P., Kemme, B., and Alonso, G. (2002). Improving the scalability of fault-tolerant database clusters. In *Proceedings of the 22nd International Conference on Distributed Computing Systems*, pages 477–484.

Karloff, H., Suri, S., and Vassilvitskii, S. (2010). A model of computation for mapreduce. In *Proceedings of the twenty-first Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 938–948.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems in Information Transmission*, 1:1–7.

Kotsiantis, S. B., Zaharakis, I. D., and Pintelas, P. E. (2006). Machine learning: A review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190.

Krikellas, K., Elnikety, S., Vagena, Z., and Hodson, O. (2010). Strongly consistent replication for a bargain. In *Proceedings of the 26th International Conference on Data Engineering*, pages 52–63.

Langley, P. (1988). Machine learning as an experimental science. *Machine Learning*, 3:5–8.

Langley, P. (2011). The changing science of machine learning. *Machine Learning*, 82:275–279.

Lebourgeois, F. and Emptoz, H. (1996). Pretopological approach for supervised learning. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 4, pages 256–260, Washington DC, USA. IEEE Computer Society.

Li, L. (2006). *Data complexity in machine learning and novel classification algorithms*. PhD thesis, California Institute of Technology.

Li, L. and Abu-Mostafa, Y. S. (2006). Data complexity in machine learning. Technical report, California Institute of Technology.

Li, M. and Vitanyi, P. M. B. (1993). *An introduction to Kolmogorov complexity and its applications*. Springer-Verlag, New York, NY, USA.

Luengo, J., Fernández, A., García, S., and Herrera, F. (2010). Addressing data complexity for imbalanced data sets: Analysis of SMOTE-based oversampling and evolutionary undersampling. *Soft Computing*.

Luengo, J. and Herrera, F. (2010). Domains of competence of fuzzy rule based classification systems with data complexity measures: A case of study using a fuzzy hybrid genetic based machine learning method. *Fuzzy Sets and Systems*, 161(1):3–19.

Macià, N. and Bernadó-Mansilla, E. (2011). Scrutiny of the UCI repository. Technical report, La Salle - Universitat Ramon Llull.

Macià, N., Ho, T. K., Orriols-Puig, A., and Bernadó-Mansilla, E. (2010a). The landscape contest at ICPR'10. In *ICPR 2010*, volume 6388 of *Lecture Note in Computer Science*, pages 29–45. Springer.

Macià, N., Orriols-Puig, A., and Bernadó-Mansilla, E. (2008). Genetic-based synthetic data sets for the analysis of classifier behavior. In *Proceedings of the 8th International Conference on Hybrid Intelligent Systems*, pages 507–512.

Macià, N., Orriols-Puig, A., and Bernadó-Mansilla, E. (2009). Emo shines a light on the holes of complexity space. In *Proceedings of the 10th annual Conference on Genetic and Evolutionary Computation*, pages 1907–1908. ACM.

Macià, N., Orriols-Puig, A., and Bernadó-Mansilla, E. (2010b). In search of targeted-complexity problems. In *Proceedings of the 11th annual Conference on Genetic and Evolutionary Computation*, pages 1055–1062. ACM.

Maciejowski, J. M. (1979). Model discrimination using an algorithmic information criterion. *Automatica*, 15:579–593.

Malina, W. (2001). Two-parameter fisher criterion. *IEEE Transactions on Systems, Man, and Cybernetics–Part B: Cybernetics*, 31(4):629–636.

Martorell, J. M. (2007). *Definició d'una metodologia experimental per a l'estudi de resultats en sistemes d'aprenentatge artificial*. PhD thesis, La Salle - Universitat Ramon Llull.

Melli, G. (1999). The datgen dataset generator. version 3.1.

Michie, D., Spiegelhalter, D. J., Taylor, C., and Campbell, J., editors (1994). *Machine learning, neural and statistical classification*. Ellis Horwood, Upper Saddle River, NJ, USA.

Mitchell, T. M. (1997). *Machine learning*. McGraw Hill.

Mitchell, T. M. (2009). Mining our reality. *Science*, 326:1644–1645.

Mollineda, R. A., Sánchez, J. S., and Sotoca, J. M. (2005). Data characterization for effective prototype selection. In *Proceedings of the 2nd Iberian conference on Pattern Recognition and Image Analysis, Part II*, pages 27–34.

Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26:394–395.

Neyman, J. and Scott, E. L. (1958). Statistical approach to problems of cosmology. *Journal of the Royal Statistical Society*, 20(1):1–43.

Oliveira, L. O. V. B. and Drummond, I. N. (2010). Real-valued Negative Selection (rns) for classification task. In *ICPR 2010*, volume 6388 of *Lecture Note in Computer Science*, pages 66–74. Springer.

Orriols-Puig, A. and Casillas, J. (2010). Fuzzy knowledge representation study for incremental learning in data streams and classification problems. *Soft Computing, in press*.

Orriols-Puig, A., Macià, N., and Ho, T. K. (2010). Documentation for the data complexity library in C++. Technical report, La Salle - Universitat Ramon Llull.

Parnas, D. L. (2007). Stop the numbers game. *Communication of ACM*, 50.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076.

Peng, Y., Flach, P. A., Soares, C., and Brazdil, P. (2002). Improved dataset characterisation for meta-learning. In *Proceedings of the 5th International Conference on Discovery Science*, pages 141–152.

Penrose, R. (1955). A generalized inverse for matrices. In *Proceedings of the Cambridge Philosophical Society*, volume 51, pages 406–413.

Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In *Advances in kernel methods: Support vector learning*, pages 185–208. MIT Press.

Prechelt, L. (1994). Proben 1–a set of benchmarks and benchmarking rules for neural network training algorithms. Technical report, Universität Karlsruhe, Fakultat fur Informatik.

Prechelt, L. (1996). A quantitative study of experimental evaluations of neural network learning algorithms: current research practice. *Neural Networks*, 9:457–462.

Prim, R. C. (1957). Shortest connection networks and some generalizations. *Bell Systems Technical Journal*, pages 1389–1401.

Quinlan, J. R. (1995). *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, California.

Rachkovskij, D. A. and Kussul, E. M. (1998). Datagen: A generator of datasets for evaluation of classification algorithms. *Pattern Recogn. Lett.*, 19:537–544.

Raudys, S. J. and Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):252–264.

Rissanen, J. (1997). Modeling by shortest data description. *Automatica*, 14(5):465–471.

Rissanen, J. (n.d.). An introduction to the MDL principle. Technical report, Helsinki Institute for Information Technology.

Rousseau, J.-J. (1972). *Les Confessions*. Librairie Générale Française.

Russell, B. (1997). *The problems of philosophy*. Oxford University Press.

Saitta, L. and Neri, F. (1998). Learning in the "real world". *Machine Learning*, 30:133–163.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, 1:317–328.

Sánchez, J. S., Mollineda, R. A., and Sotoca, J. M. (2007). An analysis of how training data complexity affects the nearest neighbor classifiers. *Pattern Analysis & Application*, 10:189–201.

Sheskin, D. J. (2000). *Handbook of parametric and nonparametric statistical procedures*. Chapman & Hall.

Singh, S. (2003). Multiresolution estimates of classification complexity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25.

Solomonoff, R. J. (2003). The kolmogorov lecture. the universal distribution and machine learning. *The Journal Computer*, 66(6):598–601.

Stanfill, C. and Waltz, D. (1986). Toward memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228.

Stonebraker, M., Abadi, D., J.DeWitt, D., Madden, S., Paulson, E., Pavlo, A., and Rasin, A. (2010). Mapreduce and parallel DBMSs: Friends or foes? *Communications of the ACM*, 53:64–71.

Stout, M., Bacardit, J., Hirst, J. D., and Krasnogor, N. (2008). Prediction of recursive convex hull class assignments for protein residues. *Bioinformatics*, 24(7):916–923.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Lawrence Erlbaum Associates, Mahwah, NJ, USA.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42(1):230–265.

van der Heijden, F., Duin, R. P. W., de Ridder, D., and Tax, D. M. J. (2004). *Classification, parameter estimation and state estimation - an engineering approach using Matlab*. John Wiley & Sons.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer Verlag.

Vapnik, V. N. (1998). *Statistical learning theory*. John Wiley & Sons, New York.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280.

Wilson, D. R. and Martinez, T. R. (1997). Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34.

Witten, I. H. and Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2nd edition.

Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2):241–259.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, 8(7):1341–1390.

Wolpert, D. H. and Macready, W. G. (1997). Self-dissimilarity: an empirically observable complexity measure. In *Proceedings of the International Conference on Complex Systems*, pages 625–643.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowledge and Information Systems*, 14(1):1–37.

DECLARATION

This is my thesis and my contribution to the scientific community.

Barcelona, October 2011

Núria Macià Antolínez