

# On the Dimensions of Data Complexity through Synthetic Data Sets

Núria MACIÀ, Ester BERNADÓ-MANSILLA, and Albert ORRIOLS-PUIG

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

**Abstract.** This paper deals with the characterization of data complexity and the relationship with the classification accuracy. We study three dimensions of data complexity: the length of the class boundary, the number of features, and the number of instances of the data set. We find that the length of the class boundary is the most relevant dimension of complexity, since it can be used as an estimate of the maximum achievable accuracy rate of a classifier. The number of attributes and the number of instances do not affect classifier accuracy by themselves, if the boundary length is kept constant. The study emphasizes the use of measures revealing the intrinsic structure of data and recommends their use to extract conclusions on classifier behavior and their relative performance in multiple comparison experiments.

**Keywords.** Data complexity, Classification, Dimensionality, Synthetic data sets

## Introduction

The analysis of data complexity [3] is an emergent research area that studies the characterization of data sets to understand their intrinsic structure and to identify to what extent useful patterns can be extracted from them. Recent investigations have characterized the complexity of data sets by a set of measures describing the geometry of classes around the feature space and have found correlations with the error of classifiers [8]. Also, preliminary studies [4] have tried to identify categories of classification problems according to complexity and relate optimal classifiers to each group. These kinds of studies may enhance our current understanding of classifier behavior such as its expected accuracy for a given problem and its comparative advantages with respect to other methods in certain types of domains.

Simultaneously with this line of study, there are many investigations trying to propose new classification algorithms or improve the existing ones. Such investigations are usually supported by experimental validation of the method of interest across a selection of several real-world data sets, and possibly comparing the given method with others to highlight its advantages. In these experiments, there are few cases where the given classifier outperforms the other(s) in all the domains. Instead, the method generally provides better performance in some problems and worse performance in others. A mandatory

conclusion of this approach is to identify in which cases the given method will be best and worst. Usually, there is no such understanding. Then, the desired result is that the method performs best more times than worst. This also has an important limitation because we cannot generalize whether the classifier will still behave best in a different set of problems, or whether a new problem will be properly classified by the given algorithm. Furthermore, given a new problem, one can never know whether there is a classifier that will perform better or whether we have already reached the maximum accuracy bound.

A classic investigation of this type involves a table of accuracy rates of different classifiers across different real-world problems (usually from public repositories). These data sets are barely characterized by the number of classes and the dimensionality. We aim to discuss why we cannot find correlations between classifier behavior and data set given this characterization. We study the influence of dimensionality on data complexity and see whether this can be somehow related with classifier performance. To perform this study, we design a set of synthetic data sets that allow us to vary the data dimensionality while maintaining a given class structure. We demonstrate that data dimensionality does not affect classifier error by itself. In summary, we emphasize the use of complexity measures that look at the intrinsic structure of the data set rather than the external appearance.

The remainder of the paper is organized as follows. Section 1 shows an example of a typical experimentation where no informational value can be extracted by relating classifier behavior to the external characterization of the data set. Section 2 briefly reviews the analysis of data complexity and describes one of the most important measures describing the inherent structure of data. Section 3 introduces this measure and finds correlations with classifier error. Then, we study the influence of data dimensionality on data complexity and classifier behavior. Finally, Section 4 presents the conclusions and future work.

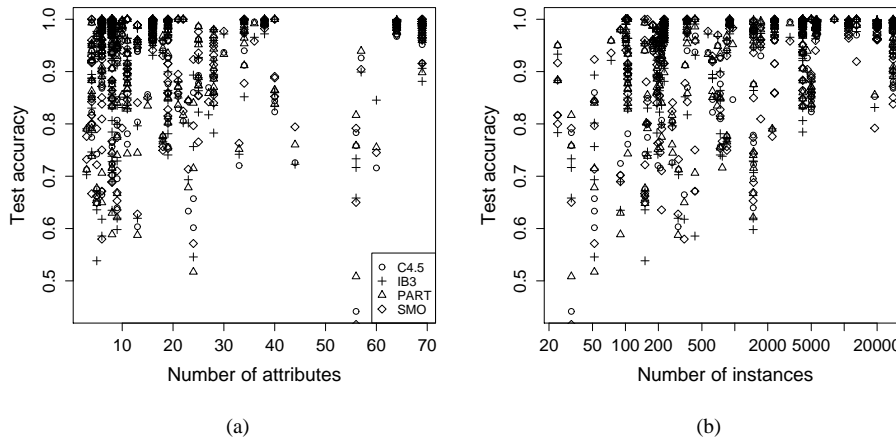
## 1. Motivation

Many studies claim the universality of a given classification method by testing it over an apparent large variety of problems. Table 1 shows an example of a typical experimentation framework, where the accuracies of several classifiers are compared on a set of problems extracted from the UCI repository [2]. Each problem is characterized by the number of classes (not depicted here, since we restricted the study to binary class problems), the number of attributes, and the number of instances. These two latter dimensions are frequently misunderstood as indicators of the problem complexity. Often one assumes that the higher the dimensionality, the higher the complexity. Certainly, these dimensions provide an estimation of the data volume and can have some relationship with data sparsity. However, there are other complexities hidden in the data sets that may be more influential. Then, we aim to study the following issue: what is the relationship, if any, between these dimensions and classifier performance?

To answer this question, we first carried out an experiment with an enhanced set containing 264 problems, which were obtained from a selection of 54 problems from the UCI repository transformed into binary class problems. Figures 1(a) and 1(b) depict the relation between the accuracy of several classifiers –an induction tree (C4.5) [10], an instance based learning IB3 [1], a rule learner (PART) [11], and a support vector ma-

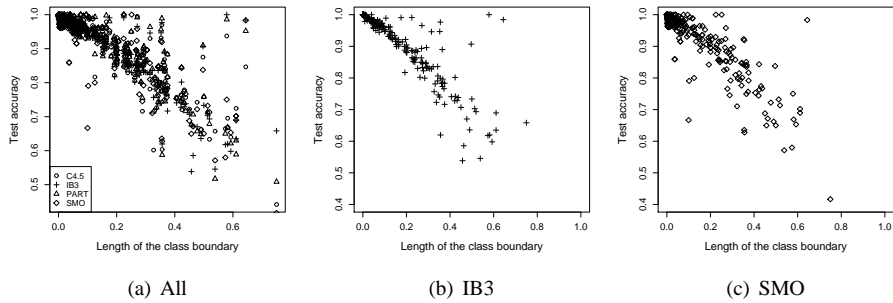
**Table 1.** Classical experimental framework

| Data set                      | #Attr | #Inst | C4.5   | IB3    | PART   | SMO    | B      |
|-------------------------------|-------|-------|--------|--------|--------|--------|--------|
| Abalone                       | 8     | 4177  | 0.9998 | 0.9998 | 0.9998 | 0.9998 | 0.0005 |
| Balance Scale                 | 4     | 625   | 0.8467 | 0.8866 | 0.8625 | 0.9297 | 0.2240 |
| Breast Cancer Wisconsin       | 30    | 569   | 0.9367 | 0.9719 | 0.9332 | 0.9771 | 0.0721 |
| Chess (King-Rook vs. King)    | 6     | 28056 | 0.9791 | 0.9010 | 0.9975 | 0.9003 | 0.1642 |
| Glass Identification          | 9     | 214   | 0.8071 | 0.8355 | 0.8407 | 0.7104 | 0.3224 |
| Heart Disease                 | 13    | 303   | 0.8037 | 0.7963 | 0.7444 | 0.8407 | 0.3667 |
| Hepatitis                     | 19    | 155   | 0.8008 | 0.7996 | 0.8020 | 0.8804 | 0.2839 |
| Ionosphere                    | 34    | 351   | 0.9120 | 0.8518 | 0.9118 | 0.8776 | 0.2308 |
| Iris                          | 4     | 150   | 0.9933 | 1      | 0.9933 | 1      | 0.0133 |
| Lenses                        | 4     | 24    | 0.8167 | 0.9333 | 0.8833 | 0.8000 | 0.1250 |
| Letter Recognition            | 16    | 20000 | 0.9960 | 0.9994 | 0.9965 | 0.9916 | 0.0016 |
| Lung Cancer                   | 56    | 32    | 0.7583 | 0.7333 | 0.7583 | 0.6500 | 0.5000 |
| Optical Recognition           | 64    | 5620  | 0.9939 | 0.9996 | 0.9957 | 0.9977 | 0.0007 |
| Pima Indians Diabetes         | 8     | 768   | 0.7566 | 0.7382 | 0.7161 | 0.7669 | 0.4375 |
| Statlog (Image Segmentation)  | 19    | 2310  | 0.9939 | 0.9952 | 0.9931 | 0.9965 | 0.0010 |
| Statlog (Vehicle Silhouettes) | 18    | 846   | 0.7695 | 0.7731 | 0.7766 | 0.7494 | 0.3652 |
| Thyroid Disease               | 5     | 215   | 0.9344 | 0.9485 | 0.9299 | 0.7907 | 0.1023 |
| Waveform Database Generator   | 21    | 5000  | 0.8290 | 0.8496 | 0.8360 | 0.8588 | 0.2384 |
| Wine                          | 13    | 178   | 0.9604 | 0.9833 | 0.9604 | 0.9889 | 0.0674 |
| Yeast                         | 8     | 1484  | 0.7223 | 0.7069 | 0.7156 | 0.6880 | 0.4501 |



**Figure 1.** Accuracy of classifiers with respect to (a) number of attributes and (b) number of instances in logarithmic scale

chine (SMO) [9]– and the number of instances and number of attributes of the data sets respectively. From these plots, we cannot observe any pattern nor any sort of correlation between the classifiers’ accuracy and these data set characteristics. Thus, it seems that it is necessary to find out other descriptors to characterize data and see if we can find better estimators of data set complexity. The best characterization would be the one able to provide a predictive estimation of learner behavior.



**Figure 2.** Accuracy of classifiers with respect to the length of the class boundary

## 2. Data Complexity

The analysis of data set complexity provides a general framework to characterize data sets and find explanations about classifier behavior. Previous studies by Ho and Basu [8] proposed a set of measures that described different aspects of complexity: a) the discriminative power of attributes, b) the separability of classes, c) the geometry of manifolds expanded by each of the classes, and d) the sparsity. These descriptors were found useful to estimate classifier performance [5]. Also, data set characterization was used to investigate the domains of competence of classifiers [4] and study the suitability of classifier ensembles [7]. In all these cases, measures of class separability were identified as the most relevant for the characterization of data set complexity. In particular, the *length of the class boundary* achieved high correlations with several algorithms' performance and thus, was identified as a good estimator of classifier error.

The length of the class boundary counts the proportion of points in the data set that lie near the class boundary. It is computed as follows. Given a data set, a *minimum spanning tree* (MST) is built with all the points of the data set, according to their Euclidean distances. Then, the number of edges connecting points of opposite classes is counted and divided by the total number of connections. This ratio is taken as the measure of boundary length. If the data set is highly interleaved, which means that points of opposite classes are very close to each other, the measure will be high (close to 1). Otherwise, if classes are well separated, the boundary length will be small.

We computed this measure to the 264 data sets to analyze whether it could provide better insights of classifier behavior than with measures based on the dimensionality of the data sets. The obtained results are depicted in figure 2. Plot 2(a) shows all the results of the different classifiers, plot 2(b) shows the most correlated classifier, which is IB3, and plot 2(c) the least correlated classifier, which corresponds to SMO. Even in the worst case, the boundary length presents almost a linear correlation with the accuracy of the learners. This means that the length of the class boundary could be an interesting data descriptor whose information could be used to predict classifier performance. Therefore, we considered the length of the class boundary as the main responsible for data complexity and studied how the data set dimensionality could alter classifier response to such complexity. The next section explains the procedure designed to perform this study.

### 3. Experimentation and Results

The previous section demonstrated that the boundary length is a good estimator of data complexity for several types of classifiers. However, we wonder whether the length of the class boundary is the only responsible for complexity. One of the relevant difficulties that classifiers may encounter is the sparsity of the training data set, i.e., the lack of representative examples. Also, a high number of attributes may hinder the performance of many classifiers. In general, the researcher knows that a low ratio between the number of instances and the number of features usually denotes sparsity and thus, a complex problem. However, this is not necessarily related to classifier accuracy as figure 1 demonstrated.

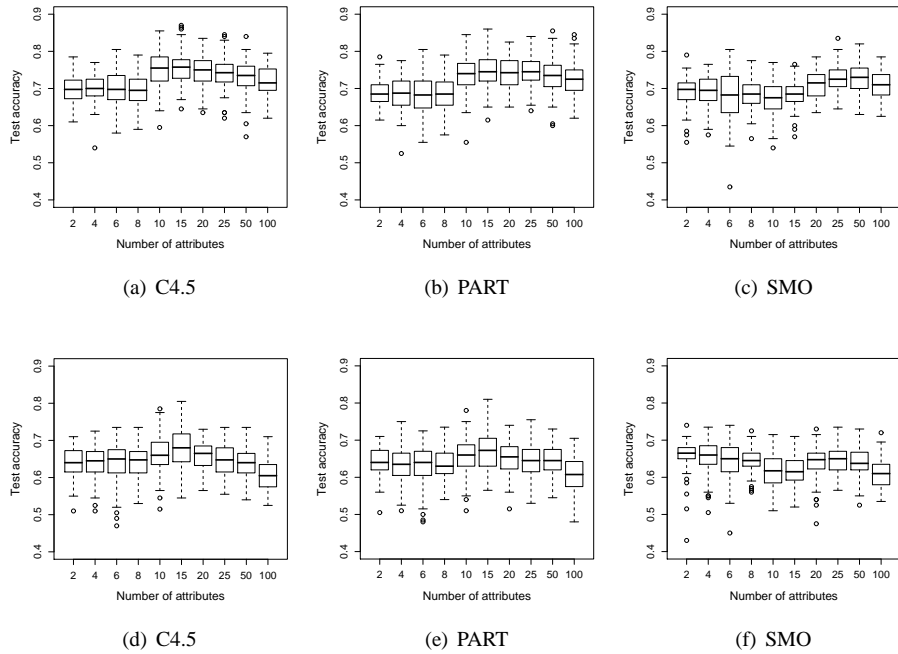
To perform such study, we could not rely on real-world data sets because we could not control the real complexity of the data set. We instead designed a set of synthetic problems that allowed us to vary these three dimensions of complexity (boundary length, number of attributes, and number of instances) independently. The data sets were built to study how a varying number of attributes and a varying number of instances influence classifier accuracy for a given fixed boundary length. For this purpose, the data sets were built according to the following procedure.

Each data set is characterized by the number of features  $m$ , the number of instances  $n$ , and desired boundary length  $b$ . Once these parameters were set, we generated  $n$  points with the values of the attributes distributed uniformly in the  $m$ -dimensional feature space. Then, we constructed the MST connecting the points according to Euclidean distances. To achieve the desired boundary length  $b$ , we label the classes of the points until there are  $b \cdot (n - 1)$  edges connecting points of different classes. In fact, for a given MST and boundary length  $b$ , there are  $2 \cdot \binom{n-1}{(n-1)-p}$  different labelings, where  $p$  is the number of edges joining different classes, i.e.,  $p = b \cdot (n - 1)$ . Among all these possible labelings, we perform an heuristic search that obtains a labeling that does not incur in class imbalances. The presence of class imbalances could also be a factor of complexity to some classifiers, so we did not aim to be disturbed by this possible complexity in our analysis. Future work will include this issue to further understand the complexity of data.

To analyze classifier behavior to increasing number of attributes, we prepared a collection of 2000 artificial data sets. We chose medium boundary length  $b = 0.3$  and  $b = 0.4$ , with 1000 data sets each. For a given boundary length, we built data sets with number of attributes  $m = \{2, 4, 6, 8, 10, 15, 20, 25, 50, 100\}$  and a fixed number of instances  $n = 200$ . For each value  $m$ , there were 100 data sets, where each data set contained a different distribution of points in the feature space and thus a different MST which was then labeled to achieve the given boundary length  $b$ . Thus, we ranged from a ratio instances/attributes of  $200/2$  to a ratio of  $200/100$ . Note that the latter represents a case where one would usually identify a difficult problem.

As the boundary length is based on the distance computation, we selected three classifiers that do not use distances in their learning process to avoid correlations with the metric. We chose three classifiers belonging to different learning paradigms: SMO, C4.5, and PART. We used a 10-fold cross-validation procedure [6] to estimate the classifier's accuracy with each data set. The algorithms were run with the software Weka [11] with their default configuration.

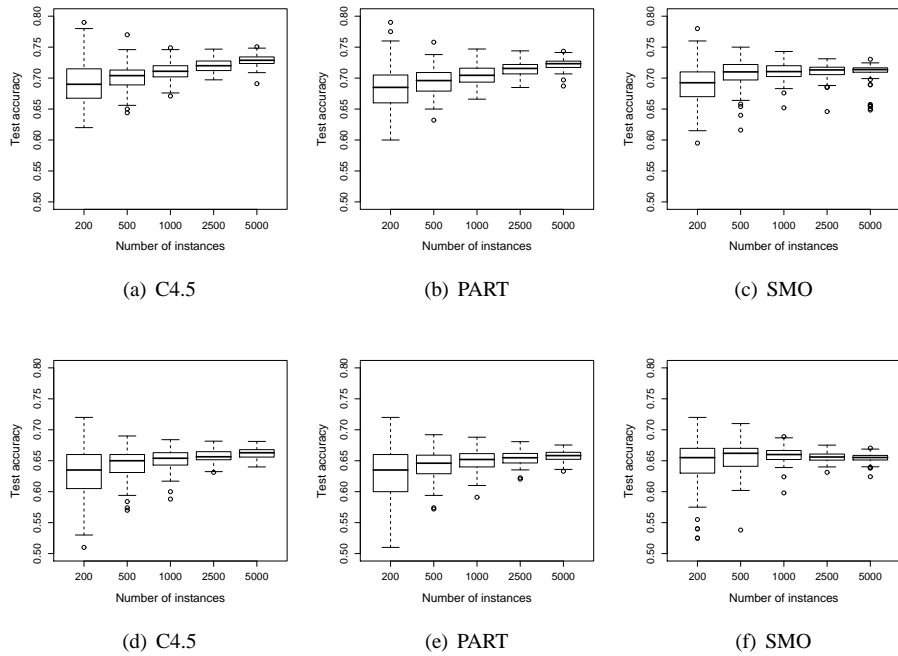
Figure 3 summarizes the results of the three classifiers, where the x-axis of each graph plots the number of attributes and the y-axis the accuracy rate. For a given number



**Figure 3.** Accuracy rate of different classifiers with a fixed boundary length and increasing number of attributes

of attributes  $m$ , the corresponding boxplot represents the range of values of the accuracy rate obtained by the classifier in the 100 data sets. The three upper plots refer to boundary complexity  $b = 0.3$  and the lower ones to boundary  $b = 0.4$ . Note that the three classifiers present similar behavior within each boundary complexity. The interquartile range of the boxplots, which contain 50% of the results, gives accuracy rates ranging in the interval  $[0.6, 0.8]$  for boundary length  $b = 0.3$  and  $[0.62, 0.7]$  for  $b = 0.4$ . The median of accuracy rates gives a value close to  $1 - b$ . Taking into account the whole spread of values in the boxplot, there is a fairly high dispersion of accuracy rates with respect to a given boundary length  $b$ . This indicates that there are other complexity issues affecting classifier error, which cannot be accounted only by the boundary length. Nevertheless, it seems that this complexity is not due to the increasing number of attributes, since in all cases, the median and spread of the accuracy rates are very similar. These results indicate that the number of attributes is not a complexity dimension influencing classifier error by itself. That is to say, the number of attributes can be a factor of complexity that can be modeled by the boundary metric. An increasing number of attributes does not necessarily incur in higher errors unless this alters the measure of boundary length. This is an interesting hypothesis that should be further analyzed but it supports the fact that classifier error does not correlate necessarily with the number of attributes as shown in figure 1(a).

In the second part of the analysis, we studied the variability of the accuracy rate with respect to the number of instances. As before, we generated 1000 artificial problems for  $b = 0.3$  and  $b = 0.4$ . We fixed the number of attributes  $m = 2$ , and built data sets with the



**Figure 4.** Accuracy rate of different classifiers with a fixed boundary length and increasing number of instances

following number of instances  $n = \{200, 500, 1000, 2500, 5000\}$ . Thus, for each boundary length  $b$  and instances  $n$ , we generated 1000 data sets. The ratio instances/features ranged from  $200/2$  to  $5000/2$ .

Figure 4 depicts the classification accuracy obtained by the classifiers. The x-axis refers to the number of instances and the y-axis the accuracy rate. For a given number of instances  $n$ , a boxplot represents the range of values of the accuracy rate obtained by the classifier in the 100 data sets. The three upper plots refer to boundary complexity  $b = 0.3$  and the lower ones to boundary  $b = 0.4$ . Note that an increasing number of instances allows for slightly better accuracy rates in general. This happens for all the three classifiers and is more notorious with boundary length  $b = 0.3$ . However, this increase of accuracy rate is still within the spread of the boxplots with the fewest number of instances. The most relevant observation is that the spread of accuracy rates decreases with increasing number of instances. A higher number of instances allows for further redundancy, which probably means that the classifiers can generalize better. Again, observe that the three classifiers have a very similar behavior. Even though they represent three different learning schemes, their accuracy rates have the same tendency. There are no significant differences among the classifiers. Our interpretation of this fact is that the complexity of the data set is more influential to the classifier’s behavior than the bias of the classifier itself. This is true for the current set of synthetic problems and may not be extrapolated to real-world problems. However, this reminds us that much caution should be taken when we extract conclusions from the comparisons of several classifiers on a small set of real world problems. Under a few number of data sets, the results of a particular classifier

could be easily biased by the particular selection. We should investigate whether the general tendency of the classifier is the same as other classifiers on a larger set of problems or on the contrary there are significant differences. We could also enrich the study by the introduction of complexity measures such as the length of the class boundary to fully understand the behavior of the classifiers. If significant differences are identified, the use of complexity characterization can also be useful to detect when these differences occur.

#### **4. Conclusions**

We analyzed learner behavior through a data characterization based on three dimensions: the number of attributes, the number of instances, and the length of the class boundary. Empirically, we showed that the number of attributes and the number of instances are not correlated to the classifiers' accuracy. Although this lack of correlation was already known, researchers still use these measures to characterize the data sets. We found that the boundary length can be considered a significant factor to assess the complexity and estimate classifier accuracy. In this study, we realized by means of synthetic data sets that the dimensionality does not affect classifier accuracy and the information on complexity provided by the number of attributes and instances can be embedded in the measure of length of the class boundary. Nevertheless, the variability observed in the results indicates that other complexities may be involved in data characterization. As a future work we aim to extend this analysis to other complexity measures.

This paper highlights the benefits of using synthetic data sets along the experiments because they allow us to vary these three dimensions independently and work under a controlled scenario. However, the generated synthetic data sets do not contain a real structure concerning the distribution of points in the feature space since the points follow a uniform distribution. Because of this distribution, we are modeling an upper bound of complexity. We expect real-world problems to have more structure due to some underlying physical process and thus, points can be grouped in more easily identifiable patterns or clusters. The synthetic data sets could be also designed to be closer to such real-world problems.

Another interesting observation from the study is that the classifiers behaved similarly. In any case, the dependence between the problem's structure and the classifier's behavior could mean that the classifier's error is mainly due to the difficulty of the problem rather than the classifiers' own constraints. In fact, we have already attained a mature development of the classifiers and we must go one step further and tackle data complexity analysis. If we can identify the real structure of the problem and have some control over the sampling processes, we could design algorithms that reduce the complexity of the problem.

#### **Acknowledgements**

The authors would like to thank *Enginyeria i Arquitectura La Salle, Universitat Ramon Llull*, the *Ministerio de Educación y Ciencia* for its support under project TIN2005-08386-C05-04, *Generalitat de Catalunya* for its support under grants 2005FI-00252 and 2005SGR-00302, and the *Govern d'Andorra* for its research grant.

## References

- [1] D. Aha, D. Kibler, and M. Albert. Instance-based learning algorithms. *Machine Learning*, 6(1):37–66, 1991.
- [2] A. Asuncion and D. Newman. UCI machine learning repository. University of California, Irvine, School of Information and Computer Sciences, 2007. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- [3] M. Basu and T. K. Ho. *Data Complexity in Pattern Recognition*. Springer-Verlag New York, 2006.
- [4] E. Bernadó-Mansilla and T. K. Ho. On classifier domains of competence. In *17th International Conference on Pattern Recognition*, volume 1, pages 136–139. IEEE Computer Society, 2004.
- [5] E. Bernadó-Mansilla, T. K. Ho, and A. Orriols-Puig. Data complexity and evolutionary learning. In *Data Complexity in Pattern Recognition*, pages 115–134. Springer, 2006.
- [6] T. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998.
- [7] T. K. Ho. Data complexity analysis for classifier combination. In *MCS '01: Proceedings of the Second International Workshop on Multiple Classifier Systems*, pages 53–67, London, UK, 2001. Springer-Verlag.
- [8] T. K. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289–300, 2002.
- [9] J. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, pages 185–208, Cambridge, MA, USA, 1998. MIT Press.
- [10] J. Quinlan. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Mateo, California, 1993.
- [11] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers, 2nd edition, 2005.