

Preliminary Approach on Synthetic Data Sets Generation based on Class Separability Measure

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

Usually, performance of classifiers is evaluated on real-world problems that mainly belong to public repositories. However, we ignore the inherent properties of these data and how they affect classifier behavior. Also, the high cost or the difficulty of experiments hinder the data collection, leading to complex data sets characterized by few instances, missing values, and imprecise data. The generation of synthetic data sets solves both issues and allows us to build problems with a minor cost and whose characteristics are predefined. This is useful to test system limitations in a controlled framework. This paper proposes to generate synthetic data sets based on data complexity. We rely on the length of the class boundary to build the data sets, obtaining a preliminary set of benchmarks to assess classifier accuracy. The study can be further matured to identify regions of competence for classifiers.

1 Introduction

Algorithms for supervised classification have practical application in a wide variety of domains such as medicine, business, and learning. At the current state of research, many algorithms do exist. Research over the last decades has lead to very competitive learners, not only in terms of accuracy, but also in computational cost and improved interpretability. With such a variety of algorithms, practitioners require estimations of classifier quality as well as guidelines for classifier selection. The usual approach is to assess learner performance in different sets of data, most of them from public repositories. Also, comparative advantages of a particular method of interest against a set of representative methods are analyzed under this approach. Nonetheless, the limitations of such an approach are well known. Assessing classifier performance on real-world data sets may

provide misleading estimates, because the estimate includes both the limitations of the algorithm and the intrinsic complexity of the data set. For example, if data are not representative enough, poor accuracy cannot be fully attributed to the algorithm. Comparing different learners with the same data sets may seem to overcome this issue, since the data set complexity is common to all approaches. However, this is also misleading because there are data set complexities that affect learners differently. For this reason, there is not any *global* learner that outperforms in all the problems, but *local* winners outperforming in certain types of problems. Research is not mature enough to identify these local winners and their associated domains of competence.

A recent approach is to try to relate classifier performance to the intrinsic complexities of the data sets. Several studies have investigated the data set characterization and have defined a set of geometrical descriptors [7]. Characterizing data sets with these descriptors has been found useful in order to understand classifier performance and identify preliminary domains of competence [2].

Another approach to investigate learner capability is to use synthetic data sets. This allows us to analyze the algorithm under a controlled scenario. The data sets can be generated according to particular features under which the algorithm is analyzed. This is the approach taken in [8] to investigate algorithm dependence on different degrees of class imbalance. Moreover, other issues such as data sparsity, noise, missing values or dimensionality, among others, can be introduced in a controlled way. Some of these features are often present in real-world problems but cannot be easily identified nor taken separately. The use of synthetic data sets offers better understanding of algorithm behavior since the complexity of the problem under study is known. In the literature, other studies resort to using synthetic

data sets due to the difficulty of obtaining real-world problems such as in [5], where privacy policies are a barrier to experimenting with real data. In other cases, there are few examples available due to the economical cost of collecting data or the difficulty of performing the experiments. Therein, synthetic data generation is used to enlarge the data set. For example, in a handwriting recognition task [9], new training examples are artificially obtained by geometrical transformation of the original sample to achieve increased learner performance.

In this paper, we propose the use of synthetic data sets as a way of analyzing the behavior of algorithms. We aim to solve the aforementioned difficulties associated with real-world problems. We acknowledge the use of geometrical descriptors to characterize the complexity of the data set [7]. Therefore, we rely on such descriptors to build the artificial data sets. The remainder of this paper develops the approach and investigates its suitability to analyze algorithm behavior.

2 Data Complexity

Ho and Basu [7] worked on the data set complexity and how this affected the performance of classifiers. They defined a set of descriptors that characterized 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 [4], as well as to investigate the classifier domains of competence [3], and the use of classifier combination [6], among other applications. Measures of class separability were deemed critical to classifier performance. In particular, the *length of class boundary* achieved high correlations with several algorithms' performance and thus, was used as an estimator of the algorithms' accuracy.

The length of the class boundary 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. The measure considers the number of points next to the class boundary; highly interleaved or randomly labeled data will provide high boundary lengths. Otherwise, if classes are well separated, the length of boundary will be small. We rely on this metric to build our synthetic data sets.

3 Synthetic Data Sets Generation based on Data Complexity

The procedure to generate data sets based on the boundary complexity is as follows (we first restrict to binary class problems). We characterize a synthetic data set by: the number of features m , the number of instances n , and the desired complexity, defined by the length of the class boundary, b . Once these parameters are set, we generate n points with the values of the attributes distributed uniformly in the m -dimensional space. Then, we label the class of each point according to the specified boundary length b . To do that, we build the MST, and then label the class of the points to achieve such complexity. We observe that 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 connecting different classes, i.e., $p = b \cdot (n - 1)$ and $b \in [0, 1]$. By varying the values of m , n , and b we obtain a set of data sets with different levels of complexity. The next section analyzes the approach.

4. Experiments and Results

The synthetic data sets would be suitable as benchmarks if they were able to set varying levels of complexity to which classifier performance can be related. For this purpose, we first analyzed classifier response to different levels of the length of the class boundary. Moreover, since for a given data set and boundary length, there are several labelings available, we also studied classifier variability under the same boundary length. In this second part of the experiments, we analyzed the variability with respect to different data sets (i.e., different MST) under the same complexity. As the boundary length is based on distance computations, we selected three classifiers (Naïve Bayes, PART and Random Tree, run with Weka [10]), that do not use distances in the learning process to avoid tied results. A 10-fold cross-validation procedure was used to estimate the accuracy rate of each classifier.

First, we generated a set of 1150 artificial two-class problems as follows. We fixed the number of instances $n = 26$ and number of attributes $m = 5$, whose values were uniformly distributed in the range $[0, 1]$. For each complexity level, $b = [1/25, 1]$, we generated 50 different data sets and analyzed the classifiers' accuracy. Figures 1(a), 1(b), and 1(c) summarize the classification accuracy obtained by Naïve Bayes, PART, and Random Tree respectively. The x-axis depicts the length of the boundary and the y-axis the accuracy rate. Apparently, the three classifiers have the same trend with respect to the complexity of data. The higher the complexity, the lower the accuracy. This suggests that the boundary is

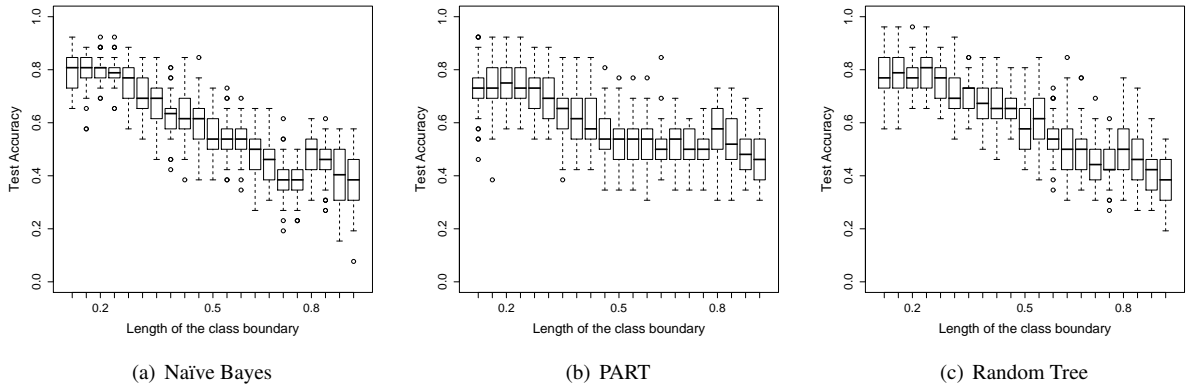


Figure 1. Accuracy of classifiers with respect to increasing levels of *boundary length* with the same MST.

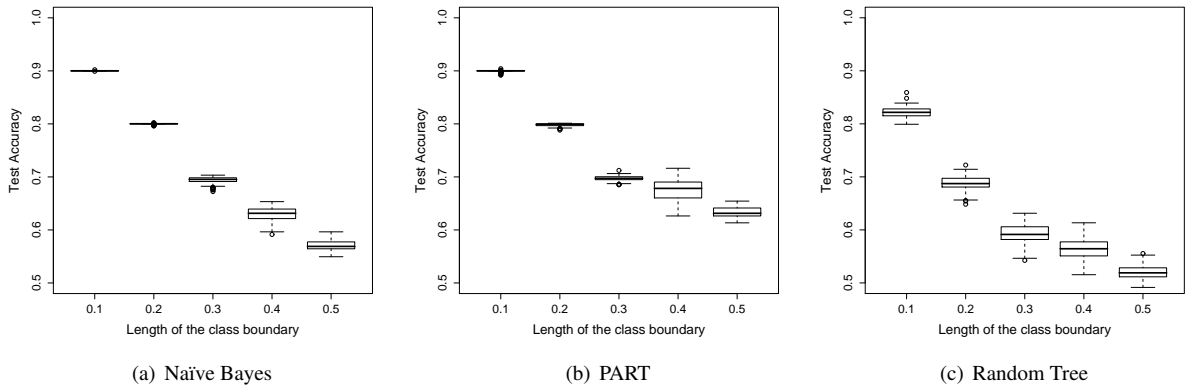


Figure 2. Accuracy of classifiers with respect to increasing levels of *boundary length* with different MST.

a reliable indicator of the complexity level. However, we should note two deviations from this general trend. First, we note that the behavior is somehow irregular for complexities greater than 0.8. Second, for a given boundary length, the boxplot shows that there is a certain variability in the classifiers' accuracy with respect to the 50 data sets obtained using different labelings. This points out that although the boundary seems to be the dominant factor of complexity, there are other types of complexities that also influence classifier behavior and should be studied in detail.

To further investigate classifier behavior, we generated a set of 500 artificial two-class problems, with $n = 1001$ instances and $m = 10$ attributes. There were 100 data sets for each complexity level $b = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. This time, each data set had a different MST, i.e., each data set was generated with different feature values. Once the MST was obtained, we labeled the class of each point according to the required boundary. Figure 2 shows the classifiers' accuracy with increasing values of boundary length. See that the correlation of the classifiers' accuracy with the boundary length is similar to the previous experiment. Surpris-

ingly, there is less variability in the classifiers' accuracy for the same boundary length than in Fig. 1. This is due to the way in which labeling is performed. We start by labeling the points at the extremes of the MST, so that all the MST obtained share a similar structure. This means that the labeling of points in the MST is more influential on classifier performance than the exact distribution of the points in the feature space.

To validate if our synthetic data sets were comparable to real-world problems, we evaluated the complexity of two data sets from the UCI Repository [1]: *ionosphere* and *pima*. These data sets were normalized, and their boundary complexity and the classifiers' accuracy were measured. Next, we generated 100 data sets for each problem with the same characteristics (number of instances, number of features, and complexity) as the UCI problems. Such data sets were run with the selected classifiers and their accuracies plotted in Fig. 3. The boxplots show the results of the classifiers in the 100 data sets, while the result in the original data set is plotted by a point. See that the UCI data sets are located either in the range of variability described by the boxplot or above the boxplot. This means that our syn-

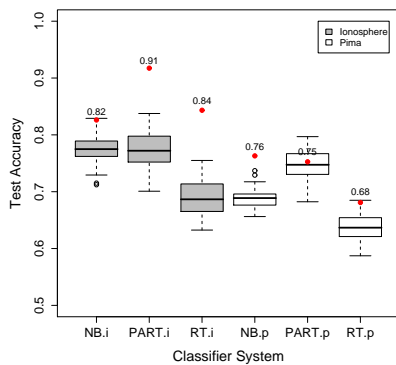


Figure 3. Artificial data sets and real-world data sets compared.

thetic data sets define a minimum accuracy bound, i.e., the complexity of the artificial problems is higher than that of the real-world problems. This may be caused by the fact that the synthetic data sets are initialized under a uniform distribution thus providing a pessimistic estimate of accuracy with respect to boundary.

5. Discussion

Although there is a general relationship between classifier accuracy and boundary length, we observed some variability for a fixed boundary length. In fact, this result was expected and supports previous studies that identified other factors responsible for classifier accuracy. The identification of these factors, the degree in which they influence accuracy, and how they are related is still under study. Furthermore, we should investigate how these factors can be introduced in the generation of data sets to provide explanations for the variability and see whether we can obtain an even much more diverse set of benchmarks.

Our synthetic data sets provided a lower expectation of the classifier's accuracy than those obtained with real-world problems, even though both were compared under the same boundary length and dimensionality. This indicates the presence of more structure in the real-world problems that must be characterized with the use of other complexity measures.

The dimensionality may also influence classifier accuracy. The generation of synthetic data sets allows us to vary either the number of attributes or the number of instances keeping the same boundary length. A known limitation of previous studies relying on real-world data sets is the apparent estimation of data complexity, which is always restricted to the available sample and not to the underlying problem. This effect is overcome with synthetic data sets. Sparsity can be introduced in a controlled way as well as other features such as noise or class imbalance.

6 Conclusions

We proposed the generation of synthetic data sets based on data complexity. Particularly, we focused on the length of the class boundary and found that it is a dominant factor in assessing the complexity of the data set. The synthetic data sets generated in this way offered a minimum expectation of the classifiers' accuracy. The inclusion of other complexity measures in the data set generation may benefit the current study and help characterize the variability of classifier accuracy. We believe this could provide a practical set of benchmarks to assess the quality of classifiers and identify types of problems where each classifier is best suited. The study of synthetic data sets could also be helpful to characterize the complexity of real-world problems.

Acknowledgements

We 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] A. Asuncion and D. Newman. UCI machine learning repository. University of California, 2007. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.
- [2] M. Basu and T. K. Ho. *Data Complexity in Pattern Recognition*. Springer-Verlag, 2006.
- [3] E. Bernadó-Mansilla and T. K. Ho. On classifier domains of competence. In *17th Int. Conf. on Pattern Recognition*, volume 1, pages 136–139, 2004.
- [4] 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-Verlag, 2006.
- [5] D. R. J. et al. Generation of synthetic data sets for evaluating the accuracy of knowledge discovery systems. In *11th Int. Conf. on Knowledge Discovery in Data mining*, pages 756–762, 2005.
- [6] T. K. Ho. Data complexity analysis for classifier combination. In *Proc. of the 2nd Int. Work. on Multiple Classifier Systems*, pages 53–67, 2001.
- [7] T. K. Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 24(3):289–300, 2002.
- [8] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligence Data Analysis*, 6(5):429–449, 2002.
- [9] T. Varga and H. Bunke. Comparing natural and synthetic training data for off-line cursive handwriting recognition. In *9th Int. Workshop on Frontiers in Handwriting Recognition*, pages 221–225, 2004.
- [10] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.