Sensible characterization of datasets : A dissimilarity approach

William Raynaut
IRIT UMR 5505, UT1, UT3
Universite de Toulouse
william.raynaut@irit.fr

Chantal Soule-Dupuy
IRIT UMR 5505, UT1, UT3
Universite de Toulouse
chantal.soule-dupuy@irit.fr

Nathalie Valles-Parlangeau
IRIT UMR 5505, UT1, UT3
Universite de Toulouse
nathalie.valles-parlangeau@irit.fr

ABSTRACT

Characterizing datasets has long been an important issue for algorithm selection and meta-level learning. Most approaches share a potential weakness when aggregating informations about individual features of the datasets. We propose a dissimilarity based approach avoiding this particular issue, and show the benefits it can yield in characterizing the appropriateness of classification algorithms.

Keywords

Dataset characterization, Dissimilarity, Meta-attributes, Meta-learning, Algorithm appropriateness

1. INTRODUCTION

In the traditionnal meta-learning framework, the dataset characterization problem consists in the definition of a subset of dataset properties (meta-level features of the dataset) that should allow a fine grain characterisation of datasets, while still complying to the requirements of the meta-level learner employed. However, to fit most learners requirements, dataset properties have to be aggregated into fixed-length feature vectors, which results into an important loss in information [1]. Relating in a way to "anti-essentialist" approaches, we investigate the possibility that limitations in the classical representations of datasets are among the main obstacles to well performing algorithm selection. We are thus focusing our efforts toward the definition of a representation that would allow the use of all available information to characterize the datasets.

2. MOTIVATION

The dataset characterization problem has been addressed along two main directions. In the first one, the dataset is described through a set of statistical or information theoretic measures as in the STATLOG project [2], and in most studies afterwards [5]. The second direction of approach to dataset characterization focuses, not on computed properties of the dataset, but on the performance of simple learners over the dataset. It was introduced as landmarking in [4], where the accuracies of a set of simple learners are used as meta-features to feed a more complex meta-level learner and further developments introduced more complex measures over the models generated by the simple learners, such as structural properties of decision trees [3].

The dataset characterization problem has thus already received quite some attention in previous meta-learning studies, but the aggregation of meta-features into fixed-length vectors processable through the meta-level learner has been a constant source of information loss.

3. APPROACH

Let us consider two datasets, A and B depicted in Figure 1. A describes 12 features of 100 individuals, and B, 10 features of 200 individuals. Let us say we want to compare the results of a set of 5 statistical or information theoretic measures over each individual feature, like mean, variance, standard deviation, entropy, and kurtosis (as illustrated over the second feature of A in Figure 1).

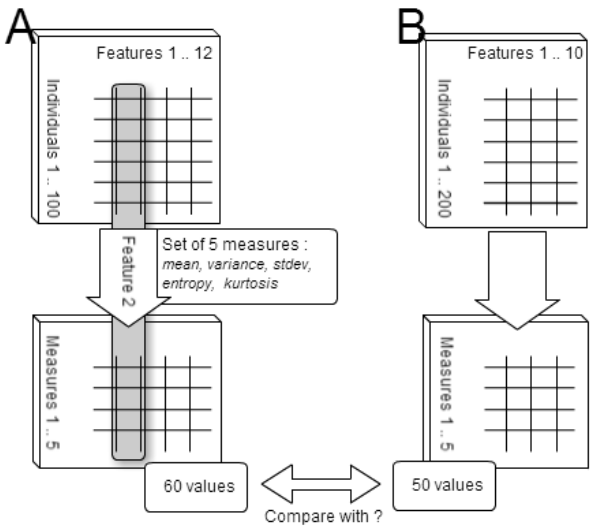

Figure 1: Measures over individual features

The complete information we want to compare is then a 60-values vector for A, and a 50-values vector for B. The standard approach would have been to average the measures over the different features, thus losing the characterization of the individual features (Figure 2)).

Our stance on the matter is to compare those features by most similar pairs, while comparing A's two extra features with empty features (features with no value at all). The assumption taken here is that a feature with absolutely no value is equivalent to no feature at all. To get back to our example, we end up comparing the 5 measures taken on the two closest (according to these very measures) features in A and B, then of the second closest, and so on, to finish on comparing the measures taken over the two extra features of A with measures taken over an artificial empty feature.

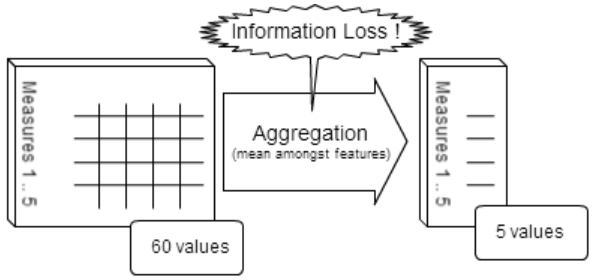

Figure 2: Averaging over individual features

These different comparisons sum up to an accurate description of how different A and B are, according to our set of measures. These pairwise comparisons would allow to ignore the presentation order of the features (which holds no meaningful information), focusing on the actual topology of the datasets.

4. VALIDATION

In [6], Wang & al. propose an intuitive definition of the goodness of dissimilarity functions in the context of learning. They define a dissimilarity function $d(x, x')$ to be *strongly* (ϵ, γ)*-good* for a given binary learning problem, if at least $1 - \epsilon$ probability mass of examples $z = (x, y)$ satisfy :

$$P(d(x, x') < d(x, x'') \mid y' = y,\ y'' = -y) \geq \tfrac{1}{2} + \tfrac{\gamma}{2}$$

In other words, the higher the chance the dissimilarity has to put examples of the same class closer together than those of different class, the greater the margin it will provide for separating the classes. This interpretation leads us to the definition of a binary problem that the proposed dissimilarity should be able to address.

Consider a set D of classification datasets, and a set A of classifiers. We execute every classifier of A on every dataset of D and measure a performance criterion c of the resulting model. Next, for each dataset x, we define the set A_x of the algorithms that are appropriate on this dataset along our performance criterion as those at most one standard deviation away from the best :

$$A_x = \{a \in A \text{ such that } |\max_{a' \in A}(c(a', x)) - c(a, x)| \leq \sigma_x\}$$

We can then consider, for each algorithm $a \in A$, the binary classification problem where instances are the datasets $x \in D$, and their class label stating whether a is appropriate on them. These problems thus characterize the appropriateness of the different algorithms on the datasets, which is an intuitive goal of the proposed dissimilarity. We can therefore compute for each algorithm $a \in A$ and dataset $x \in D$, the probability from which directly flows the (ϵ, γ)-goodness of d_ω^{ubr} :

$$P(d_\omega^{ubr}(x, x') < d_\omega^{ubr}(x, x'') \mid a \in A_x,\ a \in A_{x'},\ a \notin A_{x''})$$

The next result in [6] states that if d is a strongly (ϵ, γ)-good dissimilarity function, then there exists a simple classifier based on d that will, with probability at least $1 - \delta$ over the choice of $n = \frac{4}{\gamma^2} ln \frac{1}{\delta}$ pairs of examples of opposite class, have an error rate of no more than $\epsilon + \delta$. This result provides an easily understandable assessment of the dissimilarity adequateness to the problem.

We realised these measures over sets of datasets and classifiers from the OpenML meta-database, using in turn the full proposed dissimilarity, and the classic euclidean and Manhattan distances on the datasets meta-attributes. The γ parameter was brought as high as possible while keeping $\epsilon \leq 0.05$. Table 1 presents the δ and bound of error rate achievable for different numbers of examples and dissimilarity function.

	1000 examples		5000 examples	
	δ	error bound	δ	error bound
Proposed	0,871	0,921	0,501	0,551
Euclidean	0,945	0,995	0,755	0,805
Manhattan	0,952	1,002	0,783	0,833

	10000 examples		50000 examples	
	δ	error bound	δ	error bound
Proposed	0,251	0,301	0,001	0,051
Euclidean	0,570	0,620	0,060	0,110
Manhattan	0,613	0,663	0,086	0,136

Table 1: Error bound achievable with probability $1 - \delta$ by dissimilarity based classifiers for different numbers of examples

As we can see, the proposed dissimilarity seems to provide an improvement in characterizing the appropriateness of the different algorithms studied, giving good error bounds with much fewer examples. Yet this result is highly dependant on the choice of datasets and algorithms used to construct the appropriateness problems, and no assumption can be made toward its generalisability. What does stand, is that for *certain* algorithms, the use of the proposed dissimilarity will yield a significant improvement over classic distances in characterizing their appropriateness. Among the algorithms where the proposed dissimilarity most outperforms the other ones, we can note a majority of tree based classifiers. One can then postulate that the proposed dissimilarity characterizes well the appropriateness of tree-based classifiers, and thus that this appropriateness depends in a good part on the feature-specific meta-attributes it makes use of.

5. REFERENCES

[1] A. Kalousis and M. Hilario. Model selection via meta-learning: a comparative study. *International Journal on Artificial Intelligence Tools*, 10(04):525–554, 2001.

[2] D. Michie, D. J. Spiegelhalter, and C. C. Taylor. *Machine Learning, Neural and Statistical Classification*. Ellis Horwood, Upper Saddle River, NJ, USA, 1994.

[3] Y. Peng, P. A. Flach, P. Brazdil, and C. Soares. Decision tree-based data characterization for meta-learning. *IDDM-2002*, page 111, 2002.

[4] B. Pfahringer, H. Bensusan, and C. Giraud-Carrier. Tell me who can learn you and i can tell you who you are: Landmarking various learning algorithms. In *Proceedings of the 17th international conference on machine learning*, pages 743–750, 2000.

[5] R. Vilalta and Y. Drissi. A perspective view and survey of meta-learning. *Artif. Intell. Rev.*, 18(2):77–95, 2002.

[6] L. Wang, M. Sugiyama, C. Yang, K. Hatano, and J. Feng. Theory and algorithm for learning with dissimilarity functions. *Neural computation*, 21(5):1459–1484, 2009.