A Visualization-Based Exploratory Technique for Classifier Comparison with Respect to Multiple Metrics and Multiple Domains*

Rocío Alaiz-Rodríguez[1], Nathalie Japkowicz[2], and Peter Tischer[3]

[1] Dpto. de Ingeniería Eléctrica y de Sistemas, Universidad de León, Spain
[2] School of Information Technology and Engineering, University of Ottawa, Canada
[3] Clayton School of Information Technology, Monash University, Australia

1 Introduction

Classifier performance evaluation typically gives rise to a multitude of results that are difficult to interpret. On the one hand, a variety of different performance metrics can be applied, each adding a little bit more information about the classifiers than the others; and on the other hand, evaluation must be conducted on multiple domains to get a clear view of the classifier's general behaviour.

Caruana et al. [1] studied the issue of selecting appropriate metrics through a visualization method. In their work, the evaluation metrics are categorized into three groups and the relationship between the three groups is visualized. They propose a new metric, SAR, that combines the properties of the three groups.

Japkowicz et al. [2] studied the issue of aggregating the results obtained by different classifiers on several domains. They too use a visualization approach to implement a component-wise aggregation method that allows for a more precise combination of results than the usual averaging or win/loss/tie approaches.

In this demo, we present a visualization tool based on the combination of the above two techniques that allows the study of different classifiers with respect to both a variety of metrics and domains. We, thus, take the view that classifier evaluation should be done on an exploratory basis and provide a technique for doing so. This work is part of a research line that focuses on general issues regarding visualization and its potential benefits to the classifier evaluation process. Our aim is to adapt existing methods to suit our purpose and in this context, this paper extends a work based on MCST (Minimum Cost Spanning Tree) projection [2].

In particular, we assume that classifier evaluation requires two stages. In the first stage, the researcher should compute the results obtained by the various classifiers with respect to several representative metrics on several domains, in order to make the comparison as general as possible. This, of course, will create a considerable amount of data, which, in turn will need to be analyzed, in a second stage, in order to draw valid and useful conclusions about the algorithms

* Supported by the Natural Science and Engineering Council of Canada and the Spanish MEC projects TEC2005-06766-C03-01 and DPI2006-02550.

W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 660–665, 2008.

under study. We can say that this second stage is a data mining process in and of itself. The tool we are proposing is a visual data mining system for enabling this analysis. It is demonstrated on a study of 15 domains over three representative metrics as per Caruana et al. [1]. In particular, we demonstrate how our tool may allow us to combine information in a way that is more informative than the SAR metric [1].

2 Typical Study

Nine classifiers were evaluated by 10-fold cross-validation in the WEKA environment [3] with parameters set as default. Tables 1, 2 and 3 show the Error rate, RMSE and AUC, respectively for the 15 UCI domains assessed here (Sonar, Heart-v, Heart-c, Breast-y, Voting, Breast-w, Credits-g, Heart-s, Sick, Hepatitis, Credits-a, Horse-colic, Heart-h, Labor and Krkp).

Table 1. Error rate for different classifiers on several domains

	ERROR RATE														
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
Ideal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ib1	0.1342	0.2957	0.2378	0.2757	0.0986	0.0486	0.2800	0.2481	0.0381	0.1937	0.1884	0.1873	0.2317	0.1733	0.0372
Ib10	0.2402	0.2160	0.1753	0.2699	0.1077	0.0357	0.2600	0.1851	0.0384	0.1737	0.1405	0.1686	0.1660	0.0833	0.0494
NB	0.3211	0.2360	0.1652	0.2830	0.1284	0.0400	0.2460	0.1629	0.0739	0.1554	0.2231	0.2200	0.1629	0.1000	0.1210
C4.5	0.2883	0.2663	0.2248	0.2445	0.0917	0.0544	0.2950	0.2333	0.0119	0.1620	0.1391	0.1470	0.1893	0.2633	0.0056
Bagging	0.2545	0.2513	0.2080	0.2656	0.0895	0.0415	0.2600	0.2000	0.0127	0.1683	0.1463	0.1442	0.2105	0.1533	0.0056
Boosting	0.2219	0.2965	0.1786	0.3035	0.1010	0.0429	0.3040	0.1963	0.0082	0.1420	0.1579	0.1659	0.2142	0.1000	0.0050
RF	0.1926	0.2460	0.1850	0.3144	0.0965	0.0372	0.2730	0.2185	0.0188	0.2008	0.1492	0.1524	0.2177	0.1200	0.0122
SVM	0.2404	0.2463	0.1588	0.3036	0.0827	0.0300	0.2490	0.1592	0.0615	0.1483	0.1507	0.1740	0.1726	0.1033	0.0456
JRip	0.2692	0.2660	0.1854	0.2905	0.0986	0.0457	0.2830	0.2111	0.0177	0.2200	0.1420	0.1306	0.2104	0.2300	0.0081

Table 2. RMSE for different classifiers on several domains

	RMSE														
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
Ideal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ib1	0.3512	0.5342	0.3045	0.5042	0.2956	0.1860	0.5278	0.4848	0.1936	0.4252	0.4295	0.4261	0.2950	0.3197	0.1936
Ib10	0.3931	0.4277	0.2179	0.4305	0.2649	0.1519	0.4193	0.3700	0.1699	0.3406	0.3298	0.3587	0.2192	0.3213	0.2458
NB	0.5263	0.4164	0.2256	0.4480	0.3310	0.1945	0.4186	0.3542	0.2285	0.3409	0.4346	0.4179	0.2238	0.1997	0.3018
C4.5	0.5172	0.4531	0.2689	0.4311	0.2760	0.2105	0.4790	0.4526	0.1035	0.3565	0.3290	0.3521	0.2461	0.4209	0.0638
Bagging	0.3926	0.4177	0.2359	0.4335	0.2564	0.1769	0.4201	0.3768	0.0902	0.3388	0.3186	0.3440	0.2290	0.3412	0.0634
Boosting	0.4366	0.4700	0.2497	0.5105	0.2875	0.1864	0.5054	0.4294	0.0757	0.3507	0.3671	0.3690	0.2579	0.2281	0.0603
RF	0.3530	0.4166	0.2295	0.4686	0.2607	0.1615	0.4223	0.3912	0.1156	0.3512	0.3323	0.3376	0.2405	0.2962	0.1116
SVM	0.4837	0.4942	0.2872	0.5470	0.2667	0.1520	0.4979	0.3934	0.2479	0.3606	0.3837	0.4105	0.2885	0.2249	0.2110
JRip	0.4647	0.4360	0.2385	0.4475	0.2828	0.1932	0.44637	0.40846	0.1189	0.4075	0.3419	0.336	0.2574	0.3776	0.0782

Typical questions we would like to answer after the classifier performance analysis is performed are related to similarities/dissimilarities between classifiers: (a) Which classifiers perform similarly so that they can be considered equivalent? (b) Which classifiers could be worth combining? (c) Does the relative performance of the classifiers change as a function of data dimensionality? (d) Does it change for different domain complexities?

A first attempt at answering these questions could be to analyze directly the data gathered in the three tables. However, it does not seem straightforward given the quantity of results recorded (and there could be worse instances of this).

Table 3. AUC* (1-AUC) for different classifiers on several domains

	AUC* (1-AUC)														
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15
Ideal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Ib1	0.1361	0.4635	0.2403	0.3687	0.0622	0.0256	0.3400	0.2500	0.1912	0.3362	0.1917	0.2035	0.2512	0.1750	0.0105
Ib10	0.1373	0.4102	0.0920	0.3201	0.0325	0.0759	0.2553	0.1244	0.0672	0.1890	0.0911	0.1366	0.1138	0.0500	0.0094
NB	0.2000	0.2826	0.0955	0.2845	0.0483	0.0120	0.2122	0.0994	0.0747	0.1408	0.1040	0.1501	0.1009	0.0125	0.0479
C4.5	0.2653	0.3983	0.2032	0.3719	0.0629	0.0515	0.3534	0.2450	0.0505	0.3034	0.1064	0.1507	0.2341	0.2666	0.0012
Bagging	0.1478	0.2869	0.1296	0.3518	0.0362	0.0105	0.2469	0.1291	0.0050	0.1769	0.0771	0.1237	0.1178	0.1583	0.0007
Boosting	0.0938	0.3055	0.1187	0.3569	0.0370	0.0176	0.2770	0.1166	0.0123	0.2003	0.0945	0.1118	0.1389	0.0625	0.0007
RF	0.0889	0.2914	0.1215	0.3537	0.0376	0.0137	0.2499	0.1386	0.0072	0.1599	0.0886	0.1023	0.1444	0.0916	0.0012
SVM	0.2418	0.4335	0.1639	0.4072	0.0869	0.0316	0.3292	0.1633	0.5001	0.2487	0.1434	0.1912	0.2033	0.1250	0.0457
JRip	0.2631	0.4366	0.1591	0.3877	0.0839	0.0368	0.3871	0.2041	0.0579	0.3960	0.1285	0.1562	0.2427	0.2416	0.0055

As an alternative, metrics like SAR try to summarize all the gathered information with a point estimation. Thus, SAR carries out the projection $SAR^* = (1 - SAR) = RMSE + Error + AUC^*$ where $AUC^* = (1 - AUC)$. The closer to zero the SAR values (and all its components) are, the better the classifier performs. Table 4 shows the classifiers' performance values and ranking according to the SAR metric. We consider, however, that combining metrics uniformly may be dangerous. Instead we argue that we should select the information that is relevant to our purpose and concentrate on it to conduct the performance analysis.

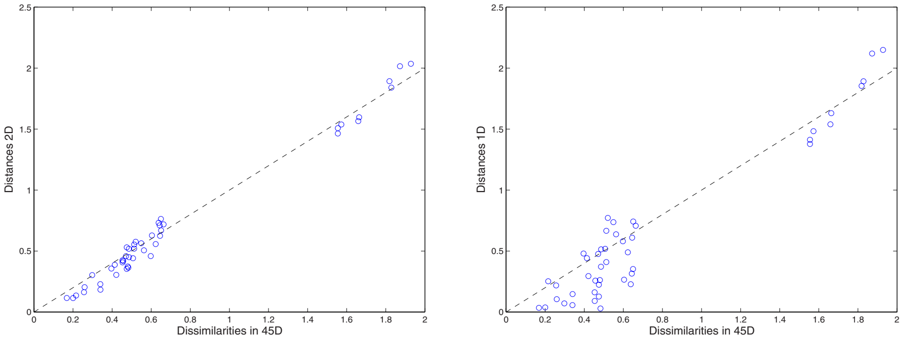
Table 4. Classifier Ranking according to SAR

Ideal	RF	Bagging	Ib10	Boosting	NB	JRip	C4.5	SVM	Ib1
0	.1958	.1965	.2001	.2037	.2126	.2362	.2365	.2420	.2530

Visual data mining allows to easily discover data patterns, a task that may be difficult by simply looking at the results organized in tables and inaccurate when summarized by a SAR-like measure. In this work, we demonstrate the use of MDS (MultiDimensional Scaling) to visualize the classifiers in a graph, so that interpoint distances in the high dimensional (metric/domain) space are preserved as much as possible in the 2D space. The technique we propose to conduct the classifier performance analysis has been implemented under Matlab. Performance data, however, can be loaded in standard file formats.

Let us now study what information may be extracted from a graphic where the information provided in Tables 1, 2 and 3 is not simply averaged (over domains and over different metrics) but is projected using MDS. The distance between two points is calculated as the Euclidean distance and the stress criterion (see below) is normalized by the sum of squares of the interpoint distances.

Before starting to explore the graphical representation, it is interesting to assess the stress criterion. It is important to know how much of the original data structure is preserved after projecting the data to two dimensions. We can also get an idea of the information gained when moving from a one dimensional



(a) (b)

Fig. 1. Shepard plot for the metric MDS projection: (a) from 45 to 2 dimensions. (b) from 45 to 1 dimension.

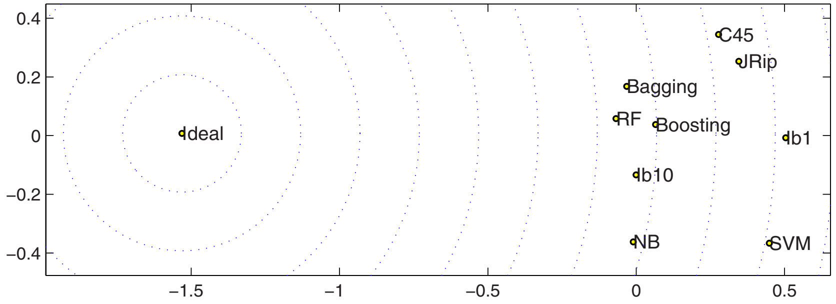


Fig. 2. metric MDS projection from 45 dimensions to 2 dimensions based on the RMSE, AUC* and Error rate gathered over 15 domains

representation to a two dimensional one. In our example the stress becomes 0.08 for two dimensions (not much loss of information), but it increases to 0.22 when considering only one dimension. This is supported by the Shepard plot in Fig. 1 that shows the reproduced distances in the new projected space (y axis) versus the dissimilarities in the original space (x axis). It can be seen that a projection to 2D leads to a narrow scatter around the ideal fit, while the scatter with a projection to 1D becomes larger and indicates a higher loss of information.

Now we focus on the whole information (Error rate, RMSE and AUC*) reflected in Fig. 2. In this particular case, we analyze nine classifiers described in the original high dimensional space by 45 dimensions (3 metrics X 15 domains) and then, projected to a 2-dimensional space. The ideal classifier is also introduced, to allow us to compare classifiers by their projected distance to the ideal classifier as well as to their relative position with respect to the other classifiers. Note that this second type of information is lost when a one-dimensional projection is used. Indeed, scalar performance measures, can only aim to convey one kind of information, usually the distance to ideal.[1]

There are cases where the projection may show no additional information. For instance, from Fig. 2 we can draw the conclusion that C4.5 and JRip perform similarly (which we can also see from Table 4). However, looking at the SAR metric in Table 4, we may also reach the conclusion that the C4.5 and SVM performance are very similar. In this case, though, Fig. 2 suggests that they behave differently. While their difference from the ideal classifier seems to be approximately equal, the distance to one another show that they behave very differently. This is confirmed by Tables 1, 2 and 3.

Fig. 1 also suggests that our tool may also be useful for model selection. Note, for example, the difference that appears, in Fig. 1, between 1-Nearest Neighbor (Ib1) and 10-Nearest Neighbor (Ib10).

3 Additional Functionality

We would like to point out that our technique is a general framework that can incorporate all other approaches. For example, we can use statistical approaches to discard some information and retain only the most relevant. This relevant information can then be visualized. Moreover, our system allows us to study a number of other questions that cannot normally be answered with traditional evaluation tools. These include questions for which data is analyzed either from a classifier point of view or from a domain point of view.

In the first case, we consider each classifier as an object described by the metrics recorded in the domains assessed (domain dimensions are reduced during the MDS projection as shown in Fig. 2).

In the second case, we can regard each domain as an object with attributes which are a measure of how several classifiers have performed on that domain. Note that the attributes are classifier performance measures and the classifier dimensions are the ones reduced. For example, let us assume that we concentrate on the posterior probability capabilities measured by the RMSE metric. Fig. 3 shows the similarities/dissimilarities among domains in terms of the difficulty for the classifiers to estimate posterior probabilities. The ideal domain D0, for which the estimation is perfect, is included for reference purposes. It is now feasible to identify groups of domains (e.g., {D3, D13, D5} or {D2, D7, D4}) for which the task of estimating posterior probabilities has similar complexity and conduct a further analysis within them.

Some questions our technique allows to address include the following:

- **Classifier-Centric Questions:**
  - Can the classifiers be organized into equivalence classes that perform similarly on a variety of domains?

[1] This is not the only type of information that gets lost, by the way, since, once in two dimensions, a lot more flexibility is possible, especially if we consider colours, motion pictures, and potentially more.

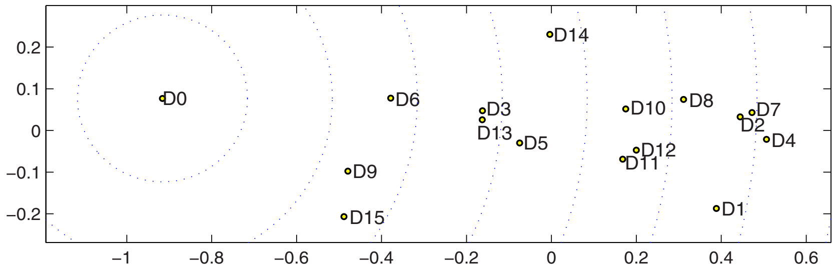


Fig. 3. metric MDS projection from 15 dimensions to 2 dimensions based on RMSE metric gathered for nine classifiers

- In what way are the classifiers similar or different from one another?
- Which classifiers would it be beneficial to combine? Which combinations would not improve the results?

– **Domain-Centric Questions:**
- Can domains be organized into equivalence classes within which various classes of classifiers behave predictably?
- What domain characteristics influence the behaviour of different domains (e.g., domain complexity, dimensionality, etc.)?

References

1. Caruana, R., Niculescu-Mizil, A.: Data mining in metric space: An empirical analysis of suppervised learning performance criteria. In: Proceedings of the Tenth International Conference on Knowledge Discovery and Data Mining (KDD 2004) (2004)
2. Japkowicz, N., Sanghi, T.P.: A projection-based framework for classifier performance evaluation. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS, vol. 5211. Springer, Heidelberg (2008)
3. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, San Francisco (1999)