Classifier Utility Visualization by Distance-Preserving Projection of High Dimensional Performance Data

Nathalie Japkowicz* and **Pritika Sanghi** and **Peter Tischer**
Clayton School of Information Technology
Monash University
Clayton, Victoria, Australia

Abstract

In this paper, we propose to view the problem of classifier evaluation in terms of a projection from a high-dimensional space to a visualizable two-dimensional space. Rather than collapsing confusion matrices into a single measure the way traditional evaluation methods do, we consider the vector composed of the entries of the confusion matrix (or the confusion matrices in case several domains are considered simultaneously) as the evaluation vector and project it into a two dimensional space using a recently proposed distance-preserving projection method. This approach is shown to be particularly useful in the case of comparison of several classifiers on many domains as well as in the case of multiclass classification.

Introduction

Evaluation in data mining has traditionally been performed by considering the confusion matrices obtained from test runs of several classifiers on various domains, collapsing each matrix into a value or pair of values (e.g., accuracy, precision/recall), and comparing these values to each other. Additionally, statistical tests are often applied (e.g., the t-test), in order to establish the statistical significance of the observed differences (Witten & Frank 2006).

More recently the research community acknowledged that basing important decisions on a single or a pair of values may be inappropriate. ***(selfcite), (Caruana & Niculescu-Mizil 2006).*** This line of thought is particularly prevalent in the subcommunity concerned with cost-sensitive learning and the class imbalance problem. From these concerns, emerged new evaluation methods that took more information into consideration simultaneously. This gave rise, in particular, to the use of methods previously unknown to the field, e.g., ROC-Analysis (Fawcett 2003), or to the creation of new approaches, e.g., Cost-Curves (Drummond & Holte 2006). Since more information is included in these methods (information derived from cost or imbalance considerations in the two examples above), they are necessarily of a visual nature.

In this paper, we propose to evaluate classifiers by reporting on more information than the traditional evaluation approaches; and as a result, a visualization is proposed. Where our method departs from traditional approaches such as ROC Analysis and Cost-Curves is in the fact that we do not add information beyond that contained in confusion matrices. Instead, we only consider single confusion matrices, but, rather than combining the data from the confusion matrices into a value or pair of values, and then, comparing these values to one another, we consider the performance of each classifier as a data point in high-dimensional space and propose to use a variation of an existing distance- preserving projection method in order to visualize this performance.[1] Two different kinds of classifier performance are indicated, by means of Euclidean distances: their distance relative to the ideal classification; and their distance relative to the two other classifiers.

The vectors representing each classifier can take different formats. They can, simply, be 4-dimensional vectors containing all the entries of the confusion matrix on a single binary domain, 9 dimensional vectors containing all the entries of the confusion matrix on a single 3-class domain, and so on. As well, they can be formed by the confusion matrices obtained by a single classifier on several domains, be they multi-class or binary domains.

The advantages of our approach are multiple. First, it allows us to visualize the results, rather than compile them into a table. This makes it easier for the researcher to interpret their meaning. Second, the method allows us to consider complex sets of results simultaneously in a way that does not summarize them to the same extent as the extent to which traditional measures such as accuracy or precision/recall do. Third, the mode of summarization used by our approach is pair-wise, as opposed to traditional measures that aggregate various, not necessarily compatible, values together. Finally, our approach not only compares the performance of classifiers to the ideal performance, like other evaluation methods, but in addition, by finding a classifier closest in performance and indicating their relative performance, it proceeds in as

*Nathalie Japkowicz is currently on sabbatical leave at Monash University. Her permanent address is: SITE, University of Ottawa, Ottawa, Ontario, Canada, K1N 6N5

[1] Note that what we propose is different from what is done in other techniques. For example, ROC Analysis continues to summarize confusion matrices into two numbers, Recall and FP Rate. What it does that is different is that it considers several instances of these two values, simultaneously.

way that can be quite useful for understanding the inner-workings of relative sets of classifiers or to identify classifiers that do not behave similarly to other systems.

The remainder of the paper is organized as follows: Section 2 describes the proposed projection method. Section 3 shows how this method fares on the simplest types of domains, binary ones, when considered on each domain separately. Section 4 extends our study to the case where classifiers are compared on several domains simultaneously. and Section 5 considers the case of multi-class problems. In both sections, we underly the particular advantages of our technique. Section 6 concludes the paper and points to possible future work.

Projection Approach

Our approach is a variation on an approach by (Lee, Slagle, & Blum 1977; Yang 2004). It is described as follows:

Let $d(x, y)$ represent the distance between x and y in the original higher dimensional space; let $P(x)$ and $P(y)$ be the projections of x and y onto the two-dimensional space; and let $d_2(P(x), P(y))$ represent the distance between the projected points in a two-dimensional space. In this case, we are projecting the performance of the classifiers, c_i where $i = 1, 2, \ldots n$. We introduce the ideal classifier as p_0. p_0 is mapped to the origin.

Find the classifier which is closest to ideal, p_1, and put this on the y-axis at $(0, d(p_0, p_1))$.

For the remaining classifiers, at each stage we find the classifier, p_i, which is nearest to the classifier which has just been plotted, p_{i-1}. When plotting p_i we want to preserve two constraints:

$$d_2(P(p_i), P(p_{i-1})) = d(p_i, p_{i-1}) \quad (1)$$

i.e. we want the projections of p_i and p_{i-1} to be the same distance apart as p_i and p_{i-1}.

We also want to satisfy the second constraint:

$$d_2(P(p_i), P(p_0)) = d(p_i, p_0) \quad (2)$$

i.e. we want the projection of the ideal classifier and the projection of p_i to be the same distance apart as the classifiers are. This means that in the projected space the distance from the origin is a measure of how close the classifier is to ideal. The better the classifier, the closer its projection will be to the origin.

Most times there will be two possible positions for $P(p_i)$ which satisfy both constraints. When there is a choice of solutions, the solution is chosen to satisfy a third constraint as closely as possible:

$$d_2(P(p_i), P(p_{i-2})) = d(p_i, p_{i-2}) \quad (3)$$

The distance measure currently used in this algorithm is the Euclidean distance.

Experiments on Single Binary Domains using Confusion Matrices

In this section, we compare the information provided by our measure to the information given by traditional evaluation measures on a classification domain taken from the UCI Repository of Machine Learning: Breast Cancer. Specifically, we study these performance measures when using 8 different classifiers: Naive Bayes (NB), C4.5 (J48), Nearest Neighbour (Ibk), Ripper (JRip), Support Vector Machines (SMO), Bagging (Bagging), Adaboost (Adaboost) and Random Forests (RandFor). All our experiments were conducted using Weka and these particular algorithms were chosen because they each represent simple and well-used prototypes of their particular categories.[2] We evaluated the algorithms using 10-fold stratified cross-validation.

The purpose of these preliminary experiments is to test our approach on relatively simple domains for which the existing traditional evaluation approaches are a good indication of the performance of the classifiers. In all the experiments of this section, the high dimensional space considered has four dimensions, the four entries of the confusion matrices. After we will have convinced ourselves that the approach is acceptable and understood how to interpret its results, we will study its outcome in more complex situations. The results obtained on the Breast Cancer domain when using the traditional evaluation measures (Accuracy (Acc), True Positive Rate (TP), False Positive Rate (FP), Precision (Prec), Recall (Rec), F-Measure (F) and the Area Under the ROC Curve (AUC) are displayed in Table 1. The graph obtained using our new measure is shown in Figure 1 and its companion table entitled "Breast Cancer Projection Legend". Table 2 compares the ranking of classifiers obtained by the traditional measures and our measure, respectively. We rank classifiers in terms of the distance between their classification matrix and the classification matrix of the ideal and neighbouring classifiers.

	Acc	TP	FP	Prec	Rec	F	AUC
NB	71.7	.44	.16	.53	.44	.48	.7
J48	75.5	.27	.04	.74	.27	.4	.59
Ibk	72.4	.32	.1	.56	.32	.41	.63
JRip	71	.37	.14	.52	.37	.43	.6
SMO	69.6	.33	.15	.48	.33	.39	.59
Bagging	67.8	.17	.1	.4	.17	.23	.63
Adaboost	70.3	.42	.18	.5	.42	.46	.7
RandFor	69.23	.33	.15	.48	.33	.39	.63

Table 1: Breast Cancer - Traditional measures

Table 2 shows that the results we obtained are believable since they have enough in common with AUC and the F-measure, two quite reliable evaluation metrics in the case of binary classification. Conversely, please note the contrast between the ranking obtained using accuracy, a less reliable performance measure, and the other three measures, including ours. In addition to what the other measures do, our measure indicates which classifiers are close to each other in performance. Figure 1 shows that the performance for Adaboost and NB is quite similar on this problem. Another

[2] As the purpose of the experiment was to interpret the results produced by our evaluation method and not to optimize performance, default settings of Weka were used.

Ranking	Accuracy	AUC	F	Distance CM	Distance All Data
1st	J48	NB (tie1)	NB	NB (tie1)	J48
2nd	Ibk	Adaboost (tie1)	Adaboost	Adaboost (tie1)	JRip (tie1)
3rd	NB	Ibk(tie1)	JRip	JRip	Adaboost (tie1)
4th	JRip	Bagging (tie2)	Ibk	SMO (tie2)	SMO (tie1)
5th	Adaboost	RandFor (tie2)	J48	RandFor (tie2)	NB (tie1)
6th	SMO	JRip	SMO (tie1)	Ibk	RandFor (tie1)
7th	RandFor	SMO (tie3)	RandFor (tie1)	Bagging	Ibk (tie1)
8th	Bagging	J48 (tie3)	Bagging	J48	Bagging

Table 2: Ranking by Accuracy, AUC, F-measure and Our Approach

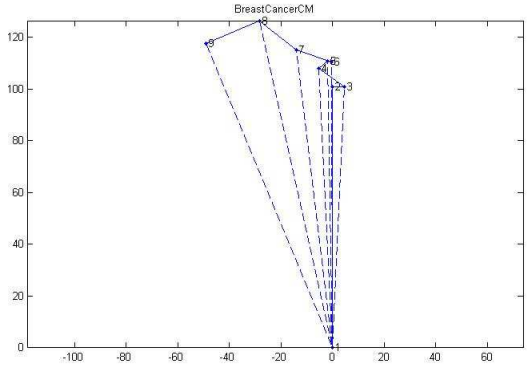

Figure 1: Projection of the results on a BinaryClass domain: Breast Cancer

Breast Cancer Projection Legend			
Classifier number	Classifier name	Distance from origin	Distance from previous classifier
1	Ideal		
2	NB	101	
3	Adaboost	101	5
4	JRip	108	12
5	SMO	111	5
6	RandFor	111	1
7	Ibk	116	14
8	Bagging	129	18
9	J48	127	22

cluster is formed by SMO, JRip and RandFor; Ibk stands by itself while Bagging and J48 are closely related. In this particular case, given that the relative distances within the clusters are not much higher than the distances between clusters, especially, relative to the distances to ideal, we do not believe that the clustering information is that meaningful. However, the next section will present cases where it is.

Experiments on Multiple Binary Domains

In this part of the paper, we experiment with the use of our approach on multiple domains. Like in the previous section, we use the Breast Cancer data set, to which we add two UCI binary data sets: Labour and Liver. This means that we are projecting vectors of size 12 (3 confusion matrices of 4 entries each) into a two dimensional domain. The results of our approach are presented in Figure 2 and its companion table entitled"Three Binary Domains Projection Legend".

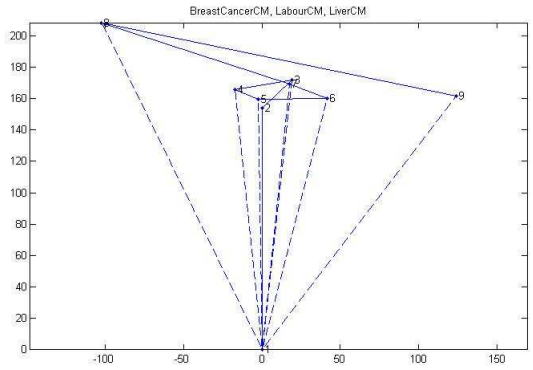

Figure 2: Projection of Three Binary Domains

Three Binary Domains Projection Legend			
Classifier number	Classifier name	Distance from origin	Distance from previous classifier
1	Ideal		
2	RandFor	154	
3	Ibk	173	26
4	JRip	167	37
5	Adaboost	160	16
6	Bagging	166	44
7	J48	170	26
8	SMO	232	126
9	NB	203	230

The results are quite interesting: they show that all the methods, except for SMO and NB, fall within the range of generally acceptable approaches. SMO and NB produce much worse results and are shown to behave very differently from one another as well, since they are not clustered together. To better understand the graph, we consider this result in view of the results obtained by the traditional measures of performance that are displayed in Table 3, for the three domains considered.

This comparison tells us something interesting: SMO fails quite miserably according to all three measures (Ac-

		Accuracy	F-Measure	AUC
NB	BC:	71.7	.48	.7
	La:	89.5	.92	.97
	Li:	55.4	.6	.64
J48	BC:	77.5	.4	.59
	La:	73.7	.79	.7
	Li:	68.7	.59	.67
Ibk	BC:	72.4	.41	.63
	La:	82.5	.86	.82
	Li:	62.9	.56	.63
JRip	BC:	71	.43	.6
	La:	77.2	.83	.78
	Li:	64.6	.53	.65
SMO	BC:	69.6	.39	.59
	La:	89.5	.92	.87
	Li:	58.3	.014	.5
Bagging	BC:	67.8	.23	.63
	La:	86	.9	.88
	Li:	71	.624	.73
Adaboost	BC:	70.3	.46	.7
	La:	87.7	.91	.87
	Li:	66.1	.534	.68
RandFor	BC:	69.23	.39	.63
	La:	87.7	.91	.9
	Li:	69	.64	.74

Table 3: Performance by Traditional Measures on the Breast Cancer (BC), Labour (La) and Liver (Li) domains.

curacy, F-measure and AUC) on the Liver data set. NB, on the other hand, only fails badly when accuracy is considered. The F-Measure and AUC do not pick up on the problem. This means that, unless accuracy was considered—a measure that is gradually becoming least trusted by data mining researchers—we could not have detected the problem encountered by NB on the Liver data set. In contrast, our method identified both the problems with NB and SMO and stated that they were of a different nature. Our method seems to warn us that these two classifiers are sometimes unreliable, whereas the other systems are more stable.

Please note that SMO's problem is something that would not have been picked up (except possibly if the F-measure had been considered) by an averaging of performance on all domains since SMO gets averages of: 72.46% in accuracy, .44 in F-measure and .65 in AUC versus 74.7% accuracy, .64 in F-measure and .75 in AUC, for Adaboost, quite a good classifier on these domains. Once its performance results averaged, NB would not have exhibited any problem whatsoever, no matter which traditional evaluation method were considered. Indeed, it produced averages of: 72.2% for accuracy, .67 for the F-measure, and .77 for the AUC. Once again, what is remarkable about our visualization approach is that the graph of Figure 2 tells us immediately that an abnormal situation has been detected with respect to SMO and NB and that this problem is of a different nature in each case. It does not tells us what the problem is, but it warns us of that problem in a quite effective way.

To further study the behaviour of our evaluation method, we also looked at results using five rather than three domains. To the three domains previously considered, we added the following two: Anneal and Contact Lenses, both from the UCI Irvine Repository for Machine Learning, as well. These two domains are slightly different from the other three since they are multi-class domains. In fact, one of them—Anneal—will be studied individually in the following section in order to illustrate the behaviour of our method on a multi-class domain. For this section, however, please note that the kind of aggregating that we are performing would be meaningless if we were to average the results of the five domains since accuracy in a binary domain has a different meaning from accuracy in a multiclass domain. Furthermore, neither the F-measure nor the AUC would be meaningful in a multiclass situation.

The results are presented in Figure 3:

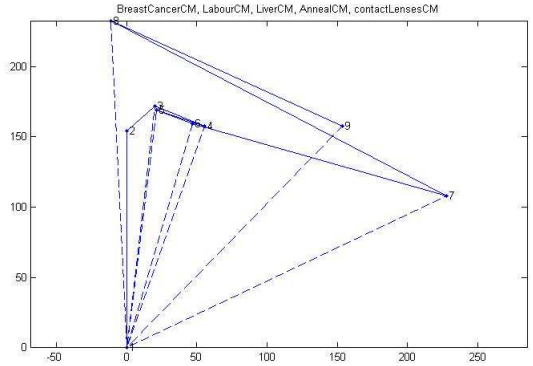

Figure 3: Projection of the results obtained on 5 domains

5 Domains Projection Legend			
Classifier number	Classifier name	Distance from origin	Distance from previous classifier
1	Ideal		
2	RandFor	154	
3	Ibk	173	27
4	JRip	167	38
5	J48	170	36
6	Bagging	166	27
7	SMO	233	130
8	Adaboost	220	181
9	NB	252	248

Here again, we are alerted of problems—lack of stability—with SMO, NB and Adaboost. This is quite interesting, actually, since these three classifiers are often considered quite strong. Our approach warns that although they may be strong on some domains, they can also be quite detrimental on others. This means that if someone is looking at selecting a single general classifier with acceptable classification performance on all domains, they should keep away from SMO, NB and, this time, Adaboost. If however, one is looking for the best classifier on a particular domain, it is possible that SMO, Adaboost or NB be the classifier of

choice (as it is known that they are on Text Classification Domains (Sebastiani 2002)).

Experiments on Single MultiClass Domains using Confusion Matrices

In this last section, we consider how our approach fares on multiclass domains. In particular, we consider the Anneal domain from UCI. Anneal is a 6-class domain (though one of the classes is represented by no data point). The data set is quite imbalanced since the classes contain 684, 99, 67, 40, 8 and 0 instances, respectively. The results obtained on this domain are displayed in Figure 4 along with the companion table entitled "Anneal Projection Legend". Once again, the graph encourages us to beware of NB and Adaboost, though it also shows us that Adaboost and NB's problems are not related. We compare the results of Figure 4 to the accuracy results obtained on this domain, displayed in Table 4.

While the accuracies (the only simple compact measure that can be used in multi-class domains) suggest that NB and Adaboost do not classify the data as well as the other domains, it does not alert us of the seriousness of the problem to the same extent that our approach does. Indeed, while it is true that NB's accuracy of 86.3% is comparatively much lower than SMO's accuracy of 97.4%, because in and of itself 86.3% is not a bad accuracy on a 6-class problem, it is conceivable that if a user had a specific interest in using NB rather than SMO or any other good method, s/he could decide that the tradeoff in accuracy is not worth a switch to a classifier other than NB since NB's accuracy is good enough for his/her particular application. This is quite different from the story painted in Figure 4 in which SMO and Adaboost are exaggeratedly far from the ideal in comparison to the other classifiers.

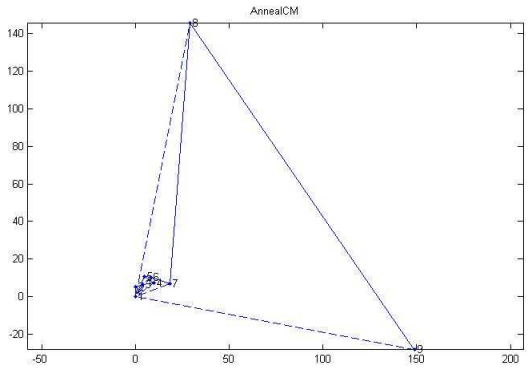

Figure 4: Projection of the results on a MultiClass domain: Anneal

In order to interpret the results, it is important to remember that the Anneal problem is severely imbalanced. The effects of this imbalance are clearly seen in the confusion matrices of Adaboost and NB in Figure 5.

As shown in Figure 5, Adaboost only gets the points from the largest class and the third largest class well-classified, ignoring all the other classes. NB classifies all the classes accurately, except for the two largest classes. We do not have space here to include the confusion matrices of the other methods, but we can report that they all did quite a good job on all classes. In effect this means that all the classifiers but NB and Adaboost are able to deal with the class imbalance problem, and that NB and Adaboost both behave badly on this domain, although they do so in different ways. This is exactly what the graph of Figure 4 tells us. The accuracy results do suggest that NB and Adaboost have problems, but they do not differentiate between the two kind of problems. Furthermore, as discussed earlier the overall accuracy of the NB and Adaboost classifiers is not drastically worse than other classifiers.

Anneal Projection Legend			
Classifier number	Classifier name	Distance from origin	Distance from previous classifier
1	Ideal		
2	RandFor	5	
3	Ibk	7	4
4	J48	12	6
5	JRip	12	6
6	Bagging	13	3
7	SMO	20	11
8	NB	148	139
9	Adaboost	151	211

Adaboost:

a	b	c	d	e	f	<- classified as
0	0	8	0	0	0	a = 1
0	0	99	0	0	0	b = 2
0	0	684	0	0	0	c = 3
0	0	0	0	0	0	d = 4
0	0	0	0	67	0	e = 5
0	0	40	0	0	0	f = U

NB:

a	b	c	d	e	f	<- classified as
7	0	1	0	0	0	a = 1
0	0	99	0	0	0	b = 2
3	38	564	0	0	79	c= 3
0	0	0	0	0	0	d = 4
0	0	0	0	67	0	e = 5
0	0	2	0	0	38	f = U

Figure 5: The confusion matrices for Adaboost and NB

Conclusion and Future Work

This paper presented a new evaluation method which, rather than aggregating entries of confusion matrices into single measures and averaging the results obtained on various domains by the same classifier, treats all the data pertaining to the performance of a classifier as a vector containing the

NB	J48	Ibk	JRip	SMO	Bag	Boost	RandFor
86.3	98.4	99.1	98.3	97.4	98.2	83.6	99.3

Table 4: Accuracies on the Anneal Data Set

confusion matrix entries obtained by that classifier on one or several domains. These vectors are then projected into a 2-dimensional space by a projection method that has the advantage of guaranteeing that the distances between some of the data points in high-dimensional space are preserved in the 2-dimensional space. This approach has several advantages, the main being that it offers a visualization method that allows us to spot immediately any irregularity in the behaviour of our classifiers. It also indicates whether the detected irregularities are similar to each other or not. This is quite informative for a simple visualization method. In many ways, it improves upon the traditional evaluation approaches.

A lot of work can be done in the future, some of which was already started. For example, we experimented with using vectors to representing the classifiers outcome on each point of the testing set instead of classification matrices. We believed that this would allow us to disaggregate the performance results further, by expanding the confusion matrix. Unfortunately, our experiments suggest that when doing so, the original space is too large, resulting in quite flat results in the two-dimensional space that may not indicate any information of interest. Other future work includes experimenting with different distance functions and testing our method more thoroughly.

References

Caruana, R., and Niculescu-Mizil, A. 2006. An empirical comparison of supervised learning algorithms. In *The Proceedings of the 23rd International Conference on Machine Learning (ICML2006)*, 161–168.

Drummond, C., and Holte, R. 2006. Cost curves: An improved method for visualizing classifier performance. *Machine Learning* 65(1):95–130.

Fawcett, T. 2003. ROC graphs: Notes and practical considerations for data mining re searchers. Technical Report HPL-2003-4, HP Laboratories, Palo Alto, CA.

Lee, R.; Slagle, J.; and Blum, H. 1977. A Triangulation Method for the Sequential Mapping of Points from N-Space to Two-Space. *IEEE Transactions on Computers* 26(3):288–292.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34(1):1–47.

Witten, I. H., and Frank, E. 2006. *Data Mining: Practical Machine Learning Tools and Techniques with J ava Implementations, Second Edition*. The Morgan Kaufmann Series in Data Management Systems. San Francisco, CA: Morgan Kaufmann Publishers.

Yang, L. 2004. Distance-preserving projection of high dimensional data. *Pattern Recognition Letters* 25(2):259–266.