Confusion Matrix Visualization

Robert Susmaga

Poznań University of Technology, Piotrowo 3a, 60-965 Poznań, Poland

Abstract. Using the technique of multidimensional scaling the paper demonstrates a method of visualizing a configuration of classes as it is perceived by a classifier. The methodology serves to assist the analysis of multi-class classification problems, where the final result of averaged accuracy or averaged error is not sufficient. The approach may be used to control and tune different classifiers applied to the same data set or a single classifier applied to different data sets. The results of such analyses may then be used for identifying the combinations of classes that proved to be worst recognized.

1 Introduction

When searching for a good classifier of a data set, and later, when tuning the selected classifier to achieve even better results, the data set is usually processed by applying cross-validation tests [16]. The most general result of such a test is the averaged number of misclassified objects, which, however, provides no information on the recognition/classification of particular classes. And such information may be interesting, since the classification error is only very rarely uniformly distributed for all classes. Instead, some pairs of classes are relatively well distinguished by the classifier, while others are mostly confused. Easy identification of all such situations is one of the applications of the presented approach.

Therefore it is useful to generate not only the averaged accuracy or averaged error of such test, but also what is called the averaged confusion matrix, i.e. the matrix that shows which classes have been confused with which during the test. Such matrix provides much more detailed information on the results of the test than the mere accuracy or error. It shows which classes were classified properly or almost properly and which where misclassified/confused with other classes and in what degree. Analysis of confusion matrices usually proves very useful. Unfortunately it becomes more and more difficult as the size of the matrix grows.

The paper deals with the visualizations of the confusion matrices. It introduces a method that allows transforming the confusion matrix into a matrix of inter-class distances. The distances are then visualized using the well-known technique of multidimensional scaling.

The rest of the paper is organized as follows. Section 2 briefly presents the methodology of multidimensional scaling. The consecutive subsections of Section 3 provide details on how to transform a confusion matrix into distances and show the method of visualizing these distances. They also discuss

M. A. Kłopotek et al. (eds.), *Intelligent Information Processing and Web Mining*

potential weaknesses of the presented approach. The final Section 4 contains conclusions and future research prospects.

2 Distance Visualization Methodology

Distances between objects in multidimensional space can be visualized using different methodologies, which include the Self-Organizing Maps (SOM) [10] or Multidimensional Scaling (MDS) [11,14,15]. This paper focuses on the latter technique.

MDS is a method allowing humans to comprehend distances in high-dimensional data sets. It is concerned with reducing the dimensionality of multidimensional data so that it can be presented in two dimensions. Unlike other dimension reducing techniques (e.g. PCA), MDS does not produce any explicit methods of converting the values of original attributes into any new ones. Instead, it simply suggests the x-y coordinates of the objects in new, two-dimensional space, so that the two-dimensional distances between particular pairs of objects are as close as possible to the distances between the objects in the original, multidimensional space.

To achieve this MDS minimizes a special function that expresses the difference between two matrices: the matrix $[d_{ij}]$ of distances in the original, multidimensional space and the matrix $[a_{ij}]$ of distances in the constructed, two-dimensional space. A common form of this function, called traditionally 'stress', is defined as follows [14]:

$$S = \frac{1}{F} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{(a_{ij}-d_{ij})^2}{d_{ij}}, \text{ where } F = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_{ij}.$$

When S is 0 then the reconstructed map accurately renders the multidimensional distances. Otherwise it may only be viewed as a helpful approximation of the multidimensional configuration of original objects. In particular, if two objects are located close to each other in the plane, then they are also relatively close to each other in the multidimensional space.

There exist also an analytic method of solving the MDS problem, in which spectral matrix decomposition is utilized. Out of the two, the stress-based approach seems more general, as the stress function might also be computed after the spectrum decomposition method has been employed. Naturally, neither the spectral nor the stress-based approach guarantees finding an ideal solution to the problem, as such a solution may not exist. Both methods merely try to provide best possible approximations of the multidimensional configuration of data.

In some sense the two-dimensional map does not bring in anything new to the original problem because, theoretically, the whole information on the distances is already included in the distance matrix. What is more, the map may be inaccurate (which is manifested by non-zero stress). It is, however, the possibility of viewing the relative positions of objects on a plane that makes the result so extremely useful. This is probably due to the fact that humans are much more skilled in comprehending graphical representations of things than in deciphering and understanding multiple rows of figures [11]. The graphical advantage may even override the reduced precision (non-zero stress!) of the presented solution.

Finally, it must be also remembered that the visualization is performed in the two-dimensional map, which in fact merely approximates the configuration of the classes of objects in the original, multidimensional space of attributes. In result, it would be unreasonable to expect full correlation between the ability of classifiers to discriminate the classes and the distances depicted in the two-dimensional map of objects.

3 Visualizing the Classification Results

The averaged accuracy of a reclassification test is a widely accepted way of assessing the capability of the classifier to discriminate objects from different classes. High values of the accuracy usually designate good classification properties of the classifier. But the classification accuracy is by far not the only measure of the classifier performance. Especially in systems where there exist one distinguished class of objects (the so-called set of 'positive' objects), i.e. objects that are to be discriminated from all the other objects (called 'negative' objects in this context), several measures have been put forward, most of which can be defined using the elements of the following table:

Table 1. An example of '*Objects from/Classfied as*' matrix

Numbers of objects	Classified as positive	Classified as negative
Positive	A	B
Negative	C	D

The table assumes that there were $A + B$ positive objects, but only A of them have been correctly recognized as positive, while B of them have been incorrectly recognized as negative. At the same time, out of $C + D$ originally negative objects, only D have been recognized correctly as negative, while C incorrectly recognized as positive.

The most popular measures based on the above table are: the recall, calculated as $A/(A + B)$, the precision, calculated as $A/(A + C)$, the F-measure, calculated as $2A/(2A + B + C)$ and the accuracy, which is calculated as $(A + C)/(A + B + C + D)$.

Especially well studied here are the recall and the precision, known for their widely discussed trade-off [2,6]. Still another frequently examined and discussed property of a classifier is its receiver operating characteristics (so-called ROC curve), often used in classifier performance visualization [1,4,7].

Please notice that in the above two classes of objects may in fact be observed: the class of some distinguished, positive objects and the class of negative objects, although in many applications only the positive objects are assumed to constitute a genuine 'class', the negative ones are simply not representatives of this class, they can be representatives of other classes, as well as the elements of some background, noise, etc.

In this paper it is assumed that the classifier has been used to discriminate more than two classes of objects. In fact, all further considerations become useful only in real multi-class problems, i.e. when the number of classes to be analysed is sufficiently high (as described e.g. in [12]).

In any case the reclassification test generates much more information than the mere classification accuracy. The above 2×2 table is then naturally generalized to an $N \times N$ table (or matrix) of classes. This matrix, called the confusion matrix (CFM), contains information on which classes were confused with which during the classification test. In result, one can deduce how the classifier perceives the spatial configuration of classes.

3.1 The Confusion Matrix

Formally, the confusion matrix is an integer-valued non-negative matrix $[c_{ij}]$, with i=1..N, j=1..N, where N is the number of classes. The element c_{ij} is the number of those objects from the class i that have actually been assigned to the class j during the test. Of course if $i = j$ then objects from class i were assigned to class i, which indicates proper functioning of the classifier. The sum of all the elements c_{ii} (the elements on the main diagonal of the matrix) is thus the number of all properly classified objects. All non-zero off-diagonal elements, on the other hand, indicate errors committed by the classifier.

Table 2. An exemplary confusion matrix

	1	2	3	4
1	15	0	0	0
2	0	5	20	0
3	0	15	10	5
4	0	1	3	1

According to the exemplary matrix from Table 1, class 1 is free from any misclassifications, which means that objects of this class are easily discernible from all other objects and the other way round. Then there are two classes, 2 and 3, as to which there is a lot of confusion, as most objects of class 2 is recognized as objects from class 3 and also many objects from class 3 were not properly discriminated from the objects of class 2. The last class 4 is also mostly confused with some of the other classes, namely with classes 2 and 3.

In this paper we present a classifier-based approach to the analysis of the configurations of classes. As we treat the classifier as the central tool in the machine learning applications, we apply it to a particular configuration of classes and treat two particular classes as distant from each other only when the classifier is correctly able to discriminate and (as a result of this) accurately classify objects from these two classes. On the other hand, classes that are misclassified by the classifier are treated as close to each other. Such classes are, in some sense, overlapping each other in the space implied by the classifier. In result, they cannot be properly discriminated by the classifier.

We thus observe the classes not in the space of the attributes describing the objects, but in the space implied by the classifier. Consequently, the problem discussed in this paper is an effective illustration and representation of the classes (and thus the CFM) as they are perceived by the classifier.

The information contained in CFM is easily interpretable as long as the CFM is small. As larger matrices are quickly becoming much harder to comprehend, there have been attempts to visualize such matrices. Descriptions of those approaches, which include different forms of so-called ‘mosaic plots’ may be found e.g. in [5,8,9].

3.2 Classifier-implied Relations between Classes

Our primary source of information on the classes is the CFM. It contains results of an attempted simultaneous discrimination of classes, which characterize them globally. At this moment, however, we shift our interest to a local characterization of the classes. In particular, we need a measure that would estimate (though maybe only approximately) the classifier-implied distance between each pair of classes.

The following extreme situations may occur as far as such relations between pairs of classes are concerned:

- The classifier perfectly discriminates the two classes (no incorrect classifications). The classes are distant from each other in the space implied by the classifier.
- The classifier completely confuses the two classes (no correct classifications). The classes seem completely indistinguishable from each other in the space implied by the classifier.

All other cases are mid-situations and indicate some partial overlapping of the classes. Below, we suggest a method of measuring this overlapping.

Strictly speaking, the misclassification between a class p and a class q is expressed in terms of two entries of the CFM: c_{pq} and c_{qp}. Treating these two classes as the only classes in the problem, we should describe the relation between these classes with just four entries of the CFM: c_{pp}, c_{qq}, c_{pq} and c_{qp}. This is, however, difficult, since the classifier operates in a multi-class context, and assuming such a detached point of view does not work well, also technically. To realize that imagine the following simple measure, which would describe (in terms of c_{pp}, c_{qq}, c_{pq} and c_{qp}) the percentage of misclassified objects between the two classes: $(c_{pq} + c_{qp})/(c_{pp} + c_{pq} + c_{qp} + c_{qq})$. In a properly posed two-class problem the sums $c_{pq} + c_{pp}$ and $c_{qp} + c_{qq}$ (and thus the sum $c_{pp} + c_{pq} + c_{qp} + c_{qq}$) will always be positive, making the computation feasible. In a multi-class problem, however, it may happen that all objects of both classes p and q can be (incorrectly) assigned to classes other than p and q, making the sum $c_{pp} + c_{pq} + c_{qp} + c_{qq}$ equal to zero. In result, relations between pairs of classes p and q must be described in terms of all entries c_{pi} and c_{qi}, for $i = 1..N$, as in a properly posed multi-class problem only the sums $\sum\limits_i c_{pi}$ are guaranteed to be different from zero.

3.3 MDS Visualization of the Implied Distances

Imagine two classes p and q. They are characterized, among others, by the entries c_{pq} and c_{qp} of the CFM. Because the discussed distance should in a sense measure the overlap of the classes, it will be combined of two respective components: the overlap of the class p with the class q and the overlap of the class q with the class p, $p \neq q$ (for p=q the distance measure will be assumed to be 0). If we count the objects from class p and assume that c$_{pq}$ of them have been (incorrectly) assigned by the classifier to class q, then the overlap of class p with class q is the ratio of c_{pq} to $\sum\limits_i c_{pi}$ (similar reasoning holds for the class q). The averaged overlap, AO_{pq}, may then be given by:

$$AO_{pq} = \frac{1}{2}\frac{c_{pq}}{\sum\limits_{i=1}^{N} c_{pi}} + \frac{1}{2}\frac{c_{qp}}{\sum\limits_{i=1}^{N} c_{qi}}.$$

Hence the distance D_{pq} between both classes may be expressed as:

$$D_{pq}\text{=}1\text{–}AO_{pq}.$$

Notice the equality of the weighing coefficients in the formula for AO_{pq} (both coefficient are equal to $\frac{1}{2}$). It implies that the two classes are treated equally, regardless of their actual cardinalities. Thus a poorly recognized class, even a very small one (as far as its cardinality is concerned), will strongly influence the ultimate distance. An opposite effect could be achieved by using as weights the ratios of class cardinalities to their combined cardinality. Such an approach would certainly give preferentiality to larger classes and, unfortunately, marginalize classes of small cardinalities. Out of the two we shall use the first option, although we do admit that proper choice of these coefficients is a complex issue that remains beyond the scope of this paper.

An important property of the introduced measure is that it renders the distance symmetric ($D_{pq} = D_{pq}$), which defies the fact that the original CFM is asymmetric. A non-symmetric distance would certainly make the result more consistent with the original information contained in the CFM, but its visualization would cause more problems, as the Euclidean distances, implied by the methodology of multidimensional scaling, are by definition symmetric. Again, a comprehensive treatment of this subject is far beyond the scope of this paper.

According to the above formulae, the distances produced from the CFM presented in Table 2 are given in Table 3 (notice the symmetry of this matrix).

Table 3. Distances implied by the confusion matrix from Table 2

	1	2	3	4
1	0.00	1.00	1.00	1.00
2	1.00	0.00	0.35	0.90
3	1.00	0.35	0.00	0.62
4	1.00	0.90	0.62	0.00

The interpretation of the distances closely follows that of the exemplary CFM. Class 1, which is absolutely free of any misclassification, turns out to be most distant from all the other classes (distance 1.0). Classes 2 and 3 are relatively close together (distance 0.35), as objects of these classes tend to be reciprocally misclassified. Finally, class 4 is also relatively distant from class 1, but less distant from classes 2 and 3.

The above conclusions become immediately noticeable after utilizing the procedure of multidimensional scaling (see Fig. 1), which graphically illustrates different classes as points in a two-dimensional space. It promptly reveals the proximity of classes 2 and 3 and clearly underlines the distinctiveness of class 1, which is characterized by the farthest distance from all other classes.

Although the above result is principally meant to convey information on the relative distances between the classes, it might be also augmented with details characterizing the classes themselves. E.g. different class cardinalities may be rendered as different sizes, shapes or colours of the symbols that represent the classes. This would compensate the information on the cardinalities of the classes, which is lost in the process of transforming the confusion matrix into the class distance matrix.

3.4 Some Disadvantages of the Approach

Despite its assets, the presented approach has some disadvantages. First of all, it might be criticized for being strongly dependent on the choice of the

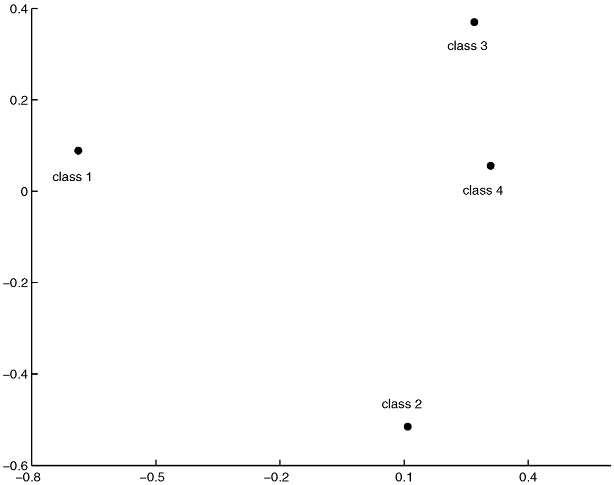

Fig. 1. The MDS map of distances from Table 2

classifier. In fact it depends not only on the classifier but also on its parameters, on the choice of the attribute space (the input to the classifier) and on the type of reclassification test used to produce the CFM. In this sense it seems more arbitrary than approaches that simply attempt to find a good representation of classes in terms of condition attributes. It must be noted, however, that if one of the purposes of the research is proper discrimination of the classes (which is often the case) then the mere choice of attribute space usually provides no information on the final performance of classifiers. And this performance may only be reliably measured after a particular classification tool is applied. In any case the classifier turns out to be an indispensable part of the data analysis process.

Secondly, the methodology provides no information on the classes in terms of the original attributes. It shows neither the relative location of classes in the attribute space, nor their spatial size, density or shape. Fortunately, for those interested in accurate class discrimination, learning the actual shape, size and the relative location of classes is important only as long as it is helpful in preventing misclassification. In opposition to such understanding of class representation and characterization, the introduced measure of distance between the classes is a much-desired 'net' result, which comprehensively aggregates information on all those aspects of classes that influence their discrimination. To properly comprehend this measure it is enough to remember that easily discriminated classes are thought of as distant, while the confused classes – as near ones.

Finally, the introduced distance measure has the upper limit of 1.0, which precludes differentiating between completely separable classes. It means that when the distance matrix to be visualized has all non-diagonal values equal to 1.0 (which happens, predominantly, in trivial or nearly trivial situations, i.e. when the only non zero elements of the processed CFM are situated on its main diagonal) the MDS will not provide a reliable map. This is because it is actually impossible to find a planar configuration of more than 3 points that are equidistant from one another.

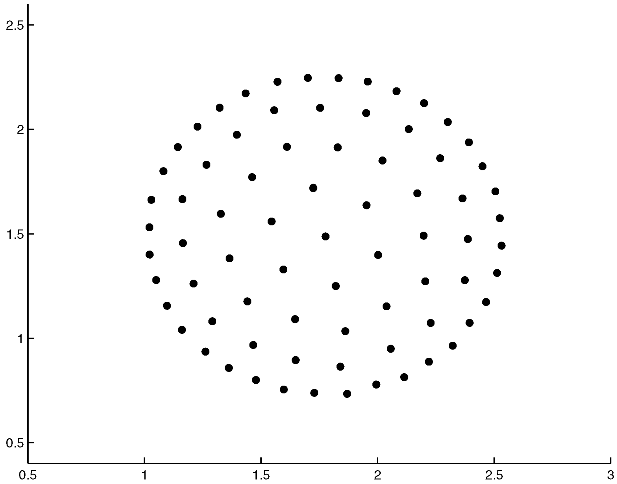

Fig. 2. Exemplary MDS map of 75 equidistant points

Fortunately, such ideal situation is simple enough not to require any particular visualization. As such it can (and should) be detected before the MDS is applied. But even after applying the MDS the equal distances between the classes are easily identifiable, as the MDS is known to generate very regular configurations of points in reply to matrices that contain many equal values. Fig. 2 illustrates the result of applying MDS to a 75×75 matrix that contains 0.0 on the main diagonal and 1.0 everywhere else.

4 Conclusions and Future Prospects

The paper deals with the problem of visualizing configurations of classes, as perceived by a particular classifier in a particular validation test. Thanks to the introduced measure the entries of the confusion matrix are transformed into distances, and these are subsequently visualized using the MDS technique. This visualization proves especially useful with data sets containing objects from multiple decision classes.

The future development of the methodology should be devoted to integrating the visualization of distances between the classes as well as the sizes and possibly shapes of the classes. This may be done by augmenting the current visualization paradigm with new graphical elements and/or metaphors or by switching to a completely new paradigm. It would be also interesting and desirable to extend the visualized distances so as demonstrate the asymmetry of the processed confusion matrices.

5 Acknowledgement

The author of this paper is grateful to the two anonymous reviewers of the paper for their useful comments. The paper has been supported by the Polish Committee for the Scientific Research, grant no. 4T11F–002–22.

References

1. Bradley A. P (1997) ‘The use of the area under the ROC curve in the evaluation of machine learning algorithms’. *Pattern Recognition*, **30** (7), 1145–1159.
2. Buckland M. K., Gey F. (1994) ‘The relationship between Recall and Precision’. *Journal of the American Society for Information Science*, 45 (1), 12–19.
3. Diettrich T.G., Bakiri G. (1995) ‘Solving multiclass learning problems via error-correcting output codes’. *Journal of Artificial Intelligence Research*, **2**, 263–286.
4. Egan J. P. (1975) *Signal Detection Theory and ROC Analysis.* Series in Cognitition and Perception. Academic Press, New York.
5. Friendly M. (1994) ‘Mosaic displays for multi-way contingency tables’. *Journal of the American Statistical Association*, **89**, 190–200.
6. Gordon M. D., Kochen M. (1989) ‘Recall-precision trade-off: a derivation’. *Journal of the American Society for Information Science*, **40**, 145–151.
7. Hanley J. A., McNeil B. J. (1982) ‘The meaning and use of the area under a receiver operating characteristic (ROC) curve’. *Radiology*, **143**, 29–36.
8. Hartigan J.A., Kleiner B. (1981) ‘Mosaics for contingency tables’. *Computer Science and Statistics: Proc. of the 13th Symposium on the Interface*, 268–273.
9. Hofmann H. (2000) ‘Exploring categorical data: interactive mosaic plots’. *Metrika*, **51** (1), 11–26.
10. Kohonen T. (1995) *Self-Organizing Maps.* Springer-Heidelberg Verlag.
11. Kruskal J. B. (1964) ‘Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis’. *Psychometrika*, **29**, 1–27.
12. Lukasik E., Susmaga R. (2003) Phoneme, gender and speaker variability visualization in voiceless stop consonants. *Proc. of the IEEE Signal Processing Workshop* (*Circuits and Systems*), Poznań, Poland.
13. Provost F., Fawcett T. (1997) ‘Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions’. *Proc. of the Third International Conference on Knowledge Discovery and Data Mining*, Menlo Park, CA, AAAI Press, 43–48.
14. Sammon J. W. Jr. (1969) ‘A nonlinear mapping for data structure analysis’. *IEEE Transactions on Computers*, **18**, 401–409.
15. Schiffman S. (1981) *Introduction to Multidimensional Scaling: Theory, Methods, and Applications.* San Diego and London Academic Press.
16. Stone M. (1974) ‘Cross-validatory choice and assessment of statistical predictions’. *Journal of the Royal Statistical Society*, **36**, 111–147.