Methods for comparing two observed confusion matrices

José Rodríguez-Avi
University of Jaén
Campus Universitario
23071 Jaén, Spain
jravi@ujaen.es

Francisco Javier Ariza-López
University of Jaén
Campus Universitario
23071 Jaén, Spain
fjariza@ujaen.es

Mª Virtudes Alba-Fernández
University of Jaén
Campus Universitario
23071 Jaén, Spain
mvalba@ujaen.es

Abstract

This work focuses on the comparison of two observed confusion matrices. This situation arises when is necessary to compare two classification procedures (products, works or essays). In this case there is not a fixed hypothesis but two observed cases. The paper presents a systematization of methods to compare two matrices. Five different methods are presented. Each one is demanding a different level of specifications for the comparisons (e.g. overall, per class, per case), and is based on different assumptions.
Keywords: confusion matrix, hypothesis testing.

1 Introduction

A confusion matrix is a statistical tool for the analysis of paired observations and is a common tool for assessing the thematic accuracy of many remote sensed derived products (e.g. land cover classifications). For thematic quality assessment the values of an observed confusion matrix are compared with previous stablished product specifications that act as a fixed hypothesis. However, sometimes it is necessary to compare the result between two assignation procedures and as a result, there are not a fixed hypothesis but two observed cases, which means a different statistical approach in order to carry out a hypothesis testing analysis.

This work is focused on the comparison of two observed confusion matrices and makes two important contributions: i) the systematization of methods to compare two matrices, ii) the novel proposal of methods that can be applied in this application field.

The classical way to address this problem is by means of comparisons using overall indexes such as the overall agreement indexes OA (Story and Congalton, 1986) or the Kappa index (Rosenfield and Fitzpatrick-Lins, 1986) between the two matrices. Nevertheless, these approximations suffer from some drawbacks, for instance, they use partial information and need large sample sizes (based on approximation to standard normal distribution). We propose several procedures in order to decide about the equality or not for two confusion matrices, depending on the amount of specifications considered by the user.

2 Hypotesis and notation

Let $\boldsymbol{A_k}$ (Table 1) and $\boldsymbol{B_k}$ (Table 2) be two confusion matrices obtained under the same procedure. In these expressions, $C_1, \dots, C_k$ are the k categories analysed, and $n = \sum\sum n_{ij}$; $m = \sum\sum m_{ij}$ the total number of elements classified for each matrix. Moreover, diagonal elements n_{ii} and m_{ii} indicate the number of concordant elements which are the elements that have been classified in the same category, whereas n_{ij} and $m_{ij},\ i \neq j$ indicate the number of discordant elements which are the elements that Procedure 1 classify into category C_j while Procedure 2 classify into category C_i

Table 1 Confusion matrix $\boldsymbol{A_k}$

	Procedure 1		
Procedure 2	C_1	C_j	C_k
C_1	n_{11}	n_{1j}	n_{1k}
C_i	n_{i1}	n_{ij}	n_{ik}
C_k	n_{k1}	n_{kj}	n_{kk}
Totals	n_{+1}	n_{+j}	n_{+k}

Table 2 Confusion matrix $\boldsymbol{B_k}$

	Procedure 1		
Procedure 2	C_1	C_j	C_k
C_1	m_{11}	m_{1j}	m_{1k}
C_i	m_{i1}	m_{ij}	m_{ik}
C_k	m_{k1}	m_{kj}	m_{kk}
Totals	m_{+1}	m_{+j}	m_{+k}

Our goal is to propose a decision rule, in the sense of a hypothesis test, that allows to determine if both matrices, $\boldsymbol{A}$ and $\boldsymbol{B}$ are the same in terms of proportion of classified elements. In consequence, the null hypothesis is that the behaviour of both matrices is the same against the alternative, which stablish a difference between them.

3 Contrasts for the equality of two confusion matrices

Procedures are developed in terms of the avalilable information, which vary from a comparison between then global proportion of concordant elements to a cell-by-cell comparison.

3.1 A single binomial contrast

In this case, we are only interested in comparing the global proportion of concordant elements in both classifications. The hypothesis are:

$$\mathbb{H}_0: \pi_A = \pi_B \;\; versus \;\; \mathbb{H}_1: \pi_A \neq \pi_B \qquad (1)$$

where the estimators are $\hat{\pi}_A = \sum n_{ii}/n$; $\hat{\pi}_B = \sum m_{ii}/m$. This test is performed using the classical approach based on the comparison between two proportions.

3.2 Multiple binomials by rows or columns

Another proposal for comparing confusion matrices is to make an individual test by columns or by rows. Now we split the hypothesis (1) into k null hypothesis:

$$\mathbb{H}_0^j: \pi_A^j = \pi_B^j \;\; versus \;\; \mathbb{H}_1^j: \pi_A^j \neq \pi_B^j \qquad (2)$$

and estimators are $\hat{\pi}_A^j = n_{jj}/n_{+j}$; $\hat{\pi}_B^j = m_{jj}/m_{+j}$, $j = 1, \dots, k$. In each case, we calculate:

$$Z_j = (\hat{\pi}_A^j - \hat{\pi}_B^j) \Big/ \sqrt{\hat{\pi}_A^j(1-\hat{\pi}_A^j)/n_{+j} + \hat{\pi}_B^j(1-\hat{\pi}_B^j)/m_{+j}} \quad (3)$$

that follows a standard normal distribution and k p-values are obtained. The final decision is adopted according to Bonferroni's correction, to assure the Type I error level

3.3 An overall chi-square test

From (3), each Z_j^2 follows a χ_1^2 distribution. Assuming independence and that each $\mathbb{H}_0^j$ in (2) is true, the test statistic $\chi = \sum_{i=1}^k Z_i^2$ follows a chi square distribution with k degrees of freedom. Here the hypothesis is tested with a single test, and we will reject the hypothesis of the whole equality between the two matrices if $P[\chi_k^2 > \chi] < \alpha$. Additionally, when the global null hypothesis is rejected, we can analyze the rejection causes.

3.4 A Kolmogorov-Smirnov test

Tests described in subsections 3.1 to 3.3 refer to the equality among diagonal concordances either for the entire matrix or column by column. However, in some cases, we are interested on comparing through the complete information of each matrix. One method to make this comparison consists in turning both matrices in vectors, in the same order, and comparing them through a discrepancy measure between them. One test than can be applied in this case is the Kolmogorov-Smirnov for two discrete samples (Glesser, 1985; Arnold and Emerson 2011). In consequence, once rearranged the matrices as vectors, we calculate the empirical distribution function in each element, and the test statistic is the maximum, in absolute value, of the differences.

3.5 A multinomial distance bootstrap test

Another test based on distances is proposed. For this, we suppose that vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ obtained from confusion matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ follow a multinomial distribution, and, under the null hypothesis, with the same probabilities

To measure the nearness between P and Q, let us consider the following discrepancy measure between multinomial distributions

$$D(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{M} \left(\sqrt{a_i} - \sqrt{b_i}\right)^2, \qquad (4)$$

where $D(\boldsymbol{a}, \boldsymbol{b}) \geq 0$, $D(\boldsymbol{a}, \boldsymbol{b}) = 0 \; iff \; \boldsymbol{a} = \boldsymbol{b}$.

Now, the statistics is

$$T_{n,m} = 4(n+m)D(\boldsymbol{a}, \boldsymbol{b}) \; .$$

If $\mathbb{H}_0$ is true, $T_{n,m} \approx 0$, so, we reject the null hypothesis for "large" values of $T_{n,m}$. To obtain the p-value we proceed with a bootstraping procedure that is carried out generating a large number of samples under the null hypothesis and calculating the value of the test statistic $T_{n,m}$ from them. In this way, we are able to approximate the probability distribution of $T_{n,m}$. Finally, we will reject the null hypothesis of equality when the bootstrap p-value < α.

4 Application examples

These contrasts are applied to decide if two confusion matrices can be considered equal. Some examples are showed, considering the case of comparing directly the classification of two methods over the same data

5 Acknowledgements

Research in this paper has been partially funded by grant CTM2015-68276-R of the Spanish Ministry on Science and Innovation (European Regional Development Funds).

6 References

Arnold, T, & Emerson, J (2011), 'Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions', *R Journal*, 3, 2, pp. 34-39

Leon Jay Gleser, a (1985), 'Exact Power of Goodness-of-Fit Tests of Kolmogorov Type for Discontinuous Distributions', *Journal Of The American Statistical Association*, 392, p. 954,

Story, M. and Congalton, R.G. (1986) Accuracy Assessment: A User's Perspective. *Photogrammetric Engineering and Remote Sensing*, 52, 397-399.

Rosenfield, G. H., and K. Fitzpatrick-Lins (1986). A Coefficient of Agreement as a Measure of Thematic Classification Accuracy. *Photogrammetric Engineering and Remote Sensing*, 52(2):223-227, 1986