Fourteenth International Conference Zaragoza-Pau on Mathematics and its Applications
Jaca, September 12–15th 2016

Confusion matrix control using penalized divergences

M.V. Alba-Fernández[1], **J.L. García-Balboa**[2], **P. Jodrá**[3]

SUMMARY

Geographic Information (GI) supports decision making in several fields as climate change, crop forecasting, forest fires, national defense, civil protection or spatial planning. The quality of the GI is essential to ensure that decisions based on it are technically the best. There are different components to describe this quality. One of them is the thematic quality, as is stablished by the international standard ISO 19157. This thematic quality is usually quantitatively assesed by means of the so called confusion matrix or error matrix (i.e. when classification correctness has to be assessed). The confusion matrix is a contingency table stablished by the cross-reference of a set of categories. It contains a counting of the items that are well and bad classified. The control of a confusion matrix is usually carried out by using two widely adopted indices like the overall accuracy and the Kappa coefficient. However, some authors have criticized them because they only use the account of elements that are correctly classified, the marginal values, and the total number of elements in the matrix.

In this work, we propose to control a confusion matrix by considering all the elements in the contingency table which is modelled by a multinomial distribution. Specifically, we propose to analyze the evaluation of the thematic quality by means of a goodness-of-fit test for a fixed null hypothesis. Due to in this setting it is common to have some empty cells, classical statistic tests do not work properly. To overcome this drawback, we consider a class of tests for testing H_0 based on the power divergence family that include a penalty for the empty cells. We analyze by means of a simulation study the power of this family of goodness-of-fit tests for a particular class of alternatives which are sensitive to deviations of the null hypothesis indexed by two constants, one for those cells with fewer cases than predicted and the other for the opposite situation.

Keywords: Confusion matrix, multinomial, divergence

AMS Classification: 62F03, 62F40, 62P30

[1] Department of Statistics and O.R.
University of Jaén
email: mvalba@ujaen.es

[2] Department of Cartographic engineering, photogrammetry and geodesy
University of Jaén
email: jlbalboa@ujaen.es

[3] Department of Statistic methods
University of Zaragoza
email: pjodra@unizar.es