Knowledge-Based Systems 90 (2015) 257–262

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Corrigendum

Corrigendum to "Coverage-based resampling: Building robust consolidated decision trees" [Knowledge-Based Systems 79(2015)51–67]

Igor Ibarguren*, Jesús M. Pérez, Javier Muguerza, Ibai Gurrutxaga, Olatz Arbelaitz

Department of Computer Architecture and Technology, University of the Basque Country UPV/EHU, Manuel Lardizabal 1, Donostia 20018, Spain [1]

In the original article we published the results of the CTC algorithm using a coverage-based strategy and compared its results against a set of genetics based and classical algorithms for rule induction. Shortly before submission we changed the storage system for the results of our experiments and a recent review of the process has discovered a mistake in the ETL process that moved data from the old system to the new system.

The mistake resulted in the TPrate and TNrate to be swapped in our database and affects two subsections of the results section of the original article. Fortunately, the metrics (kappa, accuracy and GM) used to compare CTC to the genetics-based and classical algorithms are not affected by this. On the one hand, when the performance of the CTC algorithm is analyzed regarding changes in coverage value, the values of TPrate need to be swapped with TNrate values. Figs. 6 and 7 show the correct progressions for TPrate and TNrate. In consequence the values for the F1-Score in Figs. 8 and 9 also need to be corrected. The positive trend of the TPrate when the coverage is increased was used as one of the arguments to choose a high coverage value as representative for the CTC algorithm. When the mistake is corrected the TPrate shows a negative trend. However, most of the other given metrics (kappa, accuracy, TNrate, MCC and F1-Score) show a positive trend with an increase in coverage, so the original conclusion is still correct.

On the other hand, as CTC ranked first against genetics based and classical algorithms for rule-induction in the context of class imbalance, we additionally compared the results of CTC against a set of approaches tailored to tackle the class imbalance problem in a later section of the original article. These proposals were picked from the best ranking approaches found in the literature. The metric used for this comparison is a simplified definition of the area under the ROC curve (AUC) also affected by the TPrate and TNrate being swapped:

$$AUC = \frac{1 + TP_{Rate} - FP_{Rate}}{2}$$

The original article showed SMOTE+CTC ranking 1st while CTC ranked 7th. However, according to the Friedman Aligned-Ranks test performed on the corrected result data the best ranking algorithm is EUSCHC+C4.5 (C4.5 using evolutionary undersampling) as shown on Table 17. Fig. 14 offers a visual representation of the Friedman Aligned Ranks. CTC and SMOTE+CTC are unable to beat the selected proposals designed to tackle the class imbalance problem, only showing improvement over the approaches they are based on: C4.5 for CTC and SMOTE+C4.5 for SMOTE+CTC.

The authors would like to apologize for any inconvenience caused to the journal or the community.

DOI of original article: 10.1016/j.knosys.2014.12.023

* Corresponding author. Tel.: +34943018061.

E-mail addresses: igor.ibarguren@ehu.eus (I. Ibarguren), txus.perez@ehu.eus (J.M. Pérez), j.muguerza@ehu.eus (J. Muguerza), i.gurrutxaga@ehu.eus (I. Gurrutxaga), olatz.arbelaitz@ehu.eus (O. Arbelaitz).

[1] http://www.aldapa.eus

http://dx.doi.org/10.1016/j.knosys.2015.09.031

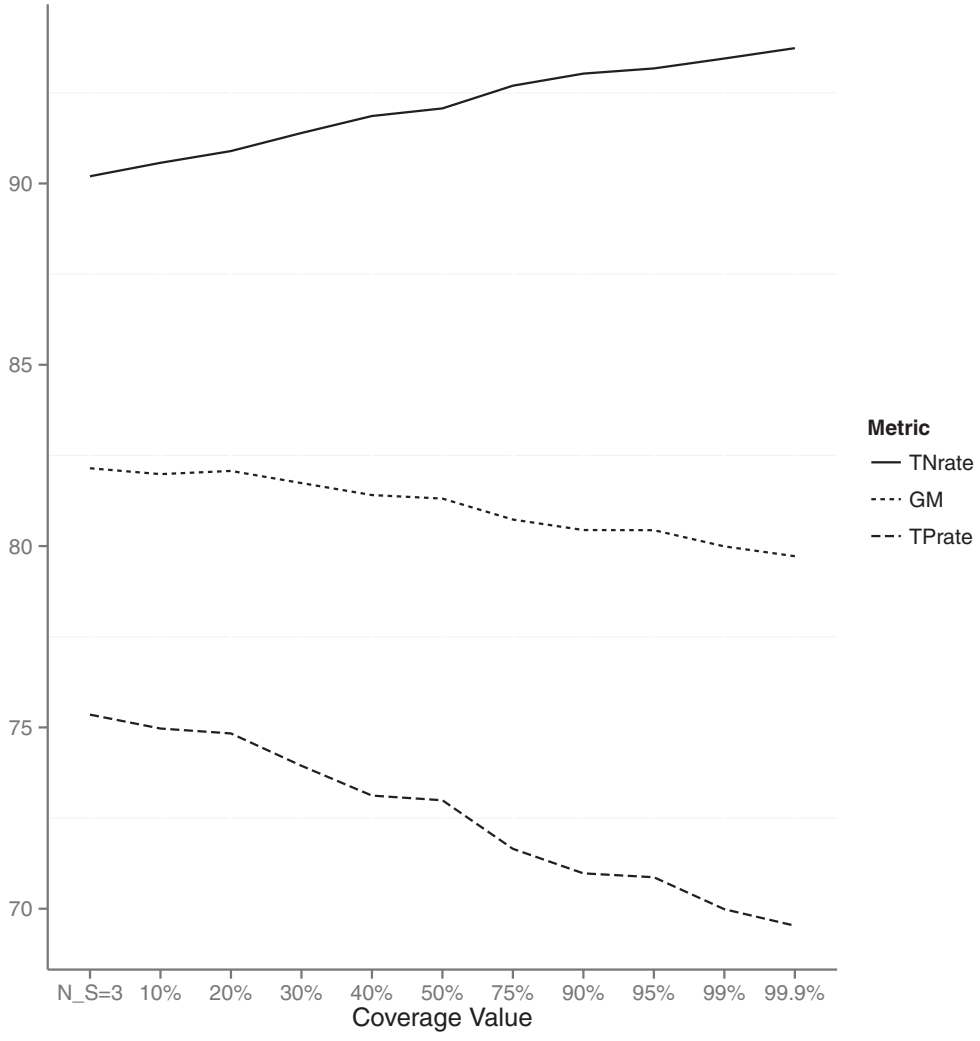

Fig. 6. Performance of CTC for a range of coverage values with maxSize subsamples for imbalanced data sets using GM, TP_{rate} and TN_{rate} as the performance measures.