Automated Text Classification in a Big-Data Context: Some Issues and Proposal Solutions

Corrado Lanera, Paola Berchialla, and Dario Gregori

University of Padova

Nowadays, it is a common thing to deal with analysis and management of big-data set of texts. In the current year we have worked on two projects concerning big-data analysis on the following topics. The first one was about an automatic categorization of 2.5M bibliographic references with abstract, and the other one was a supervised matching of Varicella cases in a 1.2M clinical free text records on 9853 distinct pediatric patients. In both projects we have encountered procedural and computational issues. Objective The objective of this work is to describe the possible issues arising from the analysis of a big data set of texts and set up a way to solve them inside the R environment.

All the analyses were conducted using R. Pre-processing of the data was required. This process was conducted according to the search of duplicate, elimination of noises (documents/words), and creation of a useful bag-of-words matrix. The latter was made by encoding strategic keywords, removing an ad-hoc set of common words, stemming all documents, considering n-Grams (up to $n = 2$), weighting the resulting set of features (words/n-Grams), and possibly eliminating sparse/lightweight ones. The creation of this matrix is the most computational demanding task because of the size of the objects involved at each step of the procedure. So, starting from the tm R package we have decomposed the procedures in order to construct the final matrix (based on a recently speed-up proposal for this task). For the first project, we have performed the categorization of the texts by creating co-Occurrence dictionaries for a number of identified labels. In this step the issue was the possibility to generalize the algorithm (i.e. mainly the labels and their quantity). This was solved by encoding the sets of labels for each documents with the product o prime numbers, each one corresponding to a single label.

For the classification of Varicella cases, we had to compare the results with a known gold standard. Among all patients we have identified some cases and somenon-cases we were sure about. In particular, the cases were identified by search and scan of key words (including the clinical codes) related with `varicel` (i.e. stemming of `varicalla` or `<274>` (i.e. varicella's code). The non-cases were identified according to the fact that none of that key words was present in any fields of any records related to the patient considered. Among the many algorithms we carried out for the classification and provided by RTextTools package (SVM, GLMNET, MAXENT, NNET, RF, TREE, BOOSTING, BAGGING and SLDA), only SVM,

GLMNET, MAXENT run successfully on entire dataset. In this case the matrix sparsity was the issue we had to sort out. In view of this, we have solved the problem removing the sparse terms.

We were able to perform the analysis for our projects using a 2.3 GHz quadcore CPUs server with 128 GiB of RAM. For the first project the task was accomplished and labels were attached on each document with a scalable algorithm in terms of numbers of labels. For the second project we finally did the analysis and we reach a 0.897 of accuracy, 0.825 on F_1 score and with 0.793, 0.859 in sensibility and specificity respectively.

Even if we performed our analysis in a quite big memory machine, following the standard procedure the system went in overflow of RAM easily and quickly. On the one hand, we have followed the general procedure described in the most recent literature. On the other hand we have realized that, in a such big-data context, the standard and useful functions provided by the considered packages can easily become unusable if not adapted and refined step by step.