pHealth 2020
B. Blobel et al. (Eds.)

doi:10.3233/SHTI200650

A Novel Method for Validating Multi-Classifiers. A Case Study for ICF-Based Health Status Classification

Federico STERNINI[a,1], Giuseppe FENZA [b], Domenico FURNO [c], Francesco Jr. ORCIUOLI [c], Alice RAVIZZA [a] and Federico CABITZA [d]
[a] *USE-ME-D srl, I3P Politecnico di Torino*
[b] *Dipartimento di Scienze Aziendali - Management & Innovation Systems, University of Salerno*
[c] *Riatlas, spin-off company of the University of Salerno*
[d] *Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi di Milano-Bicocca*

Abstract. In this paper, we propose a novel method for the validation of a multi-classification model according to the intended use and aim of a device for health status classification and the clinical needs of the practitioners involved.

Keywords. Accuracy, Multi-classifier, Health-status Classification, ICF

Introduction

All artificial intelligence (AI) applications that have a medical aim and that are aimed at providing benefit to the single patient, are considered software as a medical device. We call AI systems that satisfy these conditions Medical Artificial Intelligence (MAI) [1]. For this particular type of artificial intelligence, several requirements must be met to ensure that the device is safe, effective and developed with a constant high-quality level, as required by the European regulations. During preclinical verification, analysis of the performance of the device shall be completed [2]. In this paper, we present a proposal for the modification of already existing metrics for taxonomy classification and we present a case study of the application of this method.

1. Proposal

Classification can be divided into four main classes: binary, multi-class, multi-labelled, hierarchical. The case we want to present is a combination of multi-labelled and hierarchical classifications. Each element to be classified is associated with a set of labels, which are part of a taxonomical and hierarchical structure. The method is designed to evaluate the performance of multi-label classifiers, with labels organized with taxonomic

[1] Corresponding Author. Federico Sternini, USE-ME-D srl, I3P Politecnico di Torino, C.so Castelfidardo 30/a, 10129 Torino, TO, Italy; E-mail: federico.sternini@use-me-d.com.

structure. First of all, F1- score is defined as $F1 = \left(\frac{2}{recall^{-1}+precision^{-1}}\right)$, with $Precision = \frac{tp}{tp+fp}$ and $Recall = \frac{tp}{tp+fn}$, where *tp* is the number of true positives, *fp* is the number of false positives and *fn* is the number of false negatives. We propose a modification to the original definition of *tp*, *fp* and *fn* so that they could take into account the taxonomy of the labels.

Given $X = \{x_1, x_2, x_3, \dots, x_i, \dots, x_n\}$ a set of labels defined in the ground truth for the single item, $Y = \{y_1, y_2, y_3, \dots, y_j, \dots, y_m\}$ a set of predictions for the same item, $l(x)$ level of item x in the taxonomy, ranging from 0 (the apical and least specific one) to p (the most in-depth and most accurate), we defined the distance of the prediction from the ground truth following the definition of Sun et al. [3]. This distance is measured as the number of hypothetical steps needed to reach the desired item. A function to define the weight of a single contribution should be defined. It will be notated as $f(d)$, where $d(x_i, y_j)$ is the distance between the labels. The function *f*, which we denote as weight function, should be defined as monotonic decreasing, so that $f(0) = 1$ and $\lim_{x\to\infty} f(x) = 0$, thus ensuring the full value of the exact match of the label. Besides, another function $g(l(x_i), l(y_j))$ is defined, intended to limit the effect on the performance of predicted labels that are too generic and therefore not representative of any information. The function *g* is defined to allow contribution from exact matches only when the predicted or the reference label is in the apical levels of the tree that represents the taxonomy. Therefore, *g* is defined as

$$g = \max\left(\frac{\min(l(x_i), l(y_j)) - t}{|\min(l(x_i), l(y_j)) - t|}, 1 - d\left(x_i, y_j\right), 0\right) \quad (1)$$

where t is considered to be the deeper of the not enough informative levels of the tree. Finally the modified version of Precision and Recall are defined as follows:

$$Precision = \frac{\sum_i^n \max_j f(d(x_i, y_j))g}{m} \quad (2)$$

$$Recall = \frac{\sum_i^n \max_j f(d(x_i, y_j))g}{n} \quad (3)$$

2. Case study: ICF Classifier at Riatlas

We present the application of the proposed method for the MAI of the device Riatlas Healthcare, which is a Software as a Medical Device (SaMD) intended to facilitate the discharge of oncological patients after hospitalization. Riatlas Healthcare allows for the monitoring of various parameters, but also, it suggests, on the base of the data gathered during the first visit, the correct codes belonging to the International Classification of Functioning, Disability and Health (ICF). The algorithm performance verification is completed under the preclinical validation of the SaMD [2].

First, to complete the performance verification, two terms were set: the weight function *f* and the depth threshold *t*. In the context of the ICF code prediction, the parameters were chosen as follows:

- Weight function: the function *f(d)* is a stepwise function and is defined equal to 1 for an exact match, 0 for distances greater than 2, 80% for a distance equal to 1 and 50% for a distance equal to 2.
- Depth threshold *t*: *t* value is defined equal to 1, thus leading to no contribution of the partially correct prediction when the labels are within the first two layers of the taxonomic tree.

Then, performance is evaluated. The acceptability threshold is defined as the median value of F1 distribution. Given the low reliability of ICF scores and the room for personal interpretation of the codes, the acceptability threshold of the median was set to 70%, to ensure a conservative threshold without penalizing the algorithm performance.

The algorithm performed as described in Figure 1, where the trend of the performance over the tests is shown. The average F1-score over the tests was 81%, and the median was 82%.

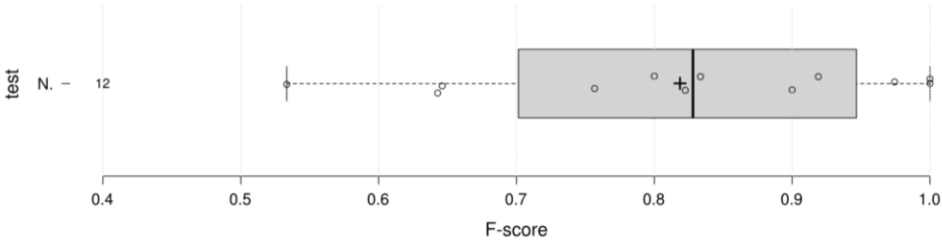

Figure 1. Performance of the algorithm.

3. Conclusions

In this poster, we show that the proposed method is capable of taking into account the partial correctness of predictions made in on taxonomically organized labels that are not entirely adherent to the ground truth, but that can provide meaningful information to the clinician. The application of the proposed method to the Riatlas Healthcare case is part of the verification of the device for regulatory purposes and evidenced the adequateness of device performance to the intended use.

References

[1] Cabitza F, Zeitoun J-D. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. Annals of Translational Medicine 2019;7:1. https://doi.org/10.21037/25300.

[2] Ravizza A, Sternini F, Giannini A, Molinari F. Methods for Preclinical Validation of Software as a Medical Device: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies, Valletta, Malta: SCITEPRESS - Science and Technology Publications; 2020, p. 648–55. https://doi.org/10.5220/0009155406480655.

[3] Sun A, Lim E-P, Ng W-K. Performance measurement framework for hierarchical text classification. Journal of the American Society for Information Science and Technology 2003;54:1014–28. https://doi.org/10.1002/asi.10298.