Classifying gene mutations in the scientific literature using neural network

Douglas Teodoro[1], Luc Mottin[2], Anaïs Mottaz[2], Paul Van Rijen[2], Emilie Pasche[2], Julien Gobeill[2], Patrick Ruch[2],

1 SIB Swiss Institute of Bioinformatics
2 HES-SO/HEG Geneva

Abstract

Discriminating between mutations that contribute to tumor growth and neutral mutations is essential for the success of precision medicine. Currently, the interpretation of genetic mutation is done by clinical pathologists via manual reviews of the scientific literature. In the context of the Classifying Clinically Actionable Genetic Mutations competition track, we investigate machine learning methods to automatically classify nine categories of genetic mutations present in text-based clinical literature. Given a scientific text article and a gene-variation pair, described in the article, our algorithm predicts the probability that the article provides evidence for the nine mutation classes. We use the paragraph2vec algorithm to embed the text in a vector space and use the vectors as features for the machine learning algorithm. The articles are divided into three parts: containing evidence to relevant gene-variation mutation pair, containing evidence to non-relevant gene-variation mutation pair, and not-containing evidence to gene-variation pair. To train and assess the methods, we use an expert-annotated dataset containing 3321 variant annotations provided by Memorial Sloan Kettering Cancer Center. We compare neural-based methods, such as Multi-Layer Perceptron (MLP) and Convolution Neural Networks, and tree-based methods, such as Random Forest and Extreme Gradient Boosting, against a Naïve Bayes baseline. Our best method (MLP) achieved an average precision of 0.7101 (0.9658 multi log-loss) compared to the 0.6220 average precision (1.1870 multi-log loss) of the baseline method. We are working to improve the classification errors by bringing further domain knowledge into the classifier. We expect that such methods could be useful for identifying relevant articles for manual curation.

Funding: ELIXIR-EXCELERATE/676559