World Academy of Science, Engineering and Technology
International Journal of Mathematical and Computational Sciences
Vol:14, No:12, 2020

Using the Smith-Waterman Algorithm to Extract Features in the Classification of Obesity Status

Authors : Rosa Figueroa, Christopher Flores

Abstract : Text categorization is the problem of assigning a new document to a set of predetermined categories, on the basis of a training set of free-text data that contains documents whose category membership is known. To train a classification model, it is necessary to extract characteristics in the form of tokens that facilitate the learning and classification process. In text categorization, the feature extraction process involves the use of word sequences also known as N-grams. In general, it is expected that documents belonging to the same category share similar features. The Smith-Waterman (SW) algorithm is a dynamic programming algorithm that performs a local sequence alignment in order to determine similar regions between two strings or protein sequences. This work explores the use of SW algorithm as an alternative to feature extraction in text categorization. The dataset used for this purpose, contains 2,610 annotated documents with the classes Obese/Non-Obese. This dataset was represented in a matrix form using the Bag of Word approach. The score selected to represent the occurrence of the tokens in each document was the term frequency-inverse document frequency (TF-IDF). In order to extract features for classification, four experiments were conducted: the first experiment used SW to extract features, the second one used unigrams (single word), the third one used bigrams (two word sequence) and the last experiment used a combination of unigrams and bigrams to extract features for classification. To test the effectiveness of the extracted feature set for the four experiments, a Support Vector Machine (SVM) classifier was tuned using 20% of the dataset. The remaining 80% of the dataset together with 5-Fold Cross Validation were used to evaluate and compare the performance of the four experiments of feature extraction. Results from the tuning process suggest that SW performs better than the N-gram based feature extraction. These results were confirmed by using the remaining 80% of the dataset, where SW performed the best (accuracy = 97.10%, weighted average F-measure = 97.07%). The second best was obtained by the combination of unigrams-bigrams (accuracy = 96.04, weighted average F-measure = 95.97) closely followed by the bigrams (accuracy = 94.56%, weighted average F-measure = 94.46%) and finally unigrams (accuracy = 92.96%, weighted average F-measure = 92.90%).

Keywords : comorbidities, machine learning, obesity, Smith-Waterman algorithm
Conference Title : ICSRD 2020 : International Conference on Scientific Research and Development
Conference Location : Chicago, United States
Conference Dates : December 12-13, 2020