Evolving fuzzy grammar for crime texts categorization

ABSTRACT

Text mining refers to the activity of identifying useful information from natural language text. This is one of the criteria practiced in automated text categorization. Machine learning (ML) based methods are the popular solution for this problem. However, the developed models typically provide low expressivity and lacking in human-understandable representation. In spite of being highly efficient, the ML based methods are established in train–test setting, and when the existing model is found insufficient, the whole processes need to be reinvented which implies train–test–retrain and is typically time consuming. Furthermore, retraining the model is not usually practical and feasible option whenever there is continuous change. This paper introduces the evolving fuzzy grammar (EFG) method for crime texts categorization. In this method, the learning model is built based on a set of selected text fragments which are then transformed into their underlying structure called fuzzy grammars. The fuzzy notion is used because the grammar matching, parsing and derivation involve uncertainty. Fuzzy union operator is also used to combine and transform individual text fragment grammars into more general representations of the learned text fragments. The set of learned fuzzy grammars is influenced by the evolution in the seen pattern; the learned model is slightly changed (incrementally) as adaptation, which does not require the conventional redevelopment. The performance of EFG in crime texts categorization is evaluated against expert-tagged real incidents summaries and compared against C4.5, support vector machines, naïve Bayes, boosting, and k-nearest neighbour methods. Results show that the EFG algorithm produces results that are close in performance with the other ML methods while being highly interpretable, easily integrated into a more comprehensive grammar system and with lower model retraining adaptability time.

Keyword: Evolving fuzzy grammar; Machine learning; Text categorization; CrimeSoft computing; Incremental learning