A comparative study of evolving fuzzy grammar and machine learning techniques for text categorization

ABSTRACT

Several methods have been studied in text categorization and mostly are inspired by the statistical distribution features in the texts, such as the implementation of Machine Learning (ML) methods. However, there is no work available that investigates the performance of ML-based methods against the text expression-based method, especially for incident and medical case categorization. Meanwhile, these two domains are becoming ever more popular, due to a growing interest of automation in security intelligence and health services. This paper presents a text expression-based method called Evolving Fuzzy Grammar (EFG) and evaluates its performance against the conventional ML methods of Naïve Bayes, support vector machine, k-nearest neighbour, adaptive booting, and decision tree. The incident dataset used is a real dataset that was taken from the World Incidents Tracking System, while Image CLEF 2009 was used as the source for radiology case reports. The results suggested variations of strength and weakness of each method in both categorization tasks, where a standard evaluation technique (i.e., recall, precision, and F-measure) was used. In both domains, the SMO and IBk methods were the best, while AdaBoost was the worst. It was also observed that the medical dataset was easier to categorize than the incident. Although EFG was ranked second lowest, it obtained the highest precision score in the bombing categorization, the highest score in armed attack recall, and was averagely ranked in the top three for the medical case categorization. It was also noted that the text expression-based method used in EFG was the most verbose and expressive, when compared to the ML methods. This indicates that EFG is a viable method in text categorization and may serve as an alternative approach to such a task.

Keyword: Text categorization; Text expression; Evolving fuzzy grammar; Machine learning Incidents; Medical