Social media mining: a genetic based multiobjective clustering approach to topic modelling

ABSTRACT

Social media mining is the process of collecting large datasets from user-generated content and extracting and analyzing social media interactions to recognize meaningful patterns in individual and social behavior. Everyday, more contents related to social media are generated by social media users (e.g., Facebook, Twitter). As the components of big data continue to expand, the task of extracting useful information becomes critical. Topic extraction refers to the process of extracting main topics from the pool of news feed and a typical method to perform topic extraction is through clustering. Clustering defines or organizes a group of patterns or objects into clusters, allows high-dimensional data to be presented in an apprehensive fashion to humans. Although effective, the performance of the k-means clustering algorithm depends heavily on the initial centroids and the number of clusters, k. Recently, several effective supervised and unsupervised machine learning methods have been developed in the domain of topics extraction. However, less works have been conducted in applying multiobjective based algorithm for topic extraction. Most of these algorithms are not optimized, even if they are, they are only optimized by using a single objective method and may underperform when solving real-world problems which are typically multi-objectives in nature. This paper investigates the effects of using a multiobjective genetic algorithm (MOGA) based clustering technique to cluster texts for topic extraction which is designed based on the structure and purity of the clusters in order to determine the optimal initial centroids and the number of clusters, k. Then, the mapping percentages between the predefined and produced clusters are used to assess the performance of the proposed algorithm. The best mapping percentage of 62.7 obtained using the proposed algorithm when k = 15 is obtained to outperform the performance of the generic k-means. The top five most representative words from each cluster are also extracted and validated by computing the number of tweets related to the predefined tags.