Chapter 62
Sentiment Analysis Using Cuckoo Search for Optimized Feature Selection on Kaggle Tweets

Akshi Kumar
Delhi Technological University, Delhi, India

Shikhar Garg
Delhi Technological University, Delhi, India

Arunima Jaiswal
Indira Gandhi Delhi Technical University for Women, Delhi, India

Shobhit Verma
Delhi Technological University, Delhi, India

Siddhant Kumar
Delhi Technological University, Delhi, India

ABSTRACT

Selecting the optimal set of features to determine sentiment in online textual content is imperative for superior classification results. Optimal feature selection is computationally hard task and fosters the need for devising novel techniques to improve the classifier performance. In this work, the binary adaptation of cuckoo search (nature inspired, meta-heuristic algorithm) known as the Binary Cuckoo Search is proposed for the optimum feature selection for a sentiment analysis of textual online content. The baseline supervised learning techniques such as SVM, etc., have been firstly implemented with the traditional tf-idf model and then with the novel feature optimization model. Benchmark Kaggle dataset, which includes a collection of tweets is considered to report the results. The results are assessed on the basis of performance accuracy. Empirical analysis validates that the proposed implementation of a binary cuckoo search for feature selection optimization in a sentiment analysis task outperforms the elementary supervised algorithms based on the conventional tf-idf score.

DOI: 10.4018/978-1-6684-6303-1.ch062

INTRODUCTION

The increasing traction of social media avenues to verbalize personal notions & beliefs has created a need to put in place a paradigm which can analyse the humongous amount of data involved, the task is typically referred to as sentiment analysis (Kumar & Sharma, 2016). Formally, Sentiment Analysis is defined as the study, and subsequent categorization, of an individual's feelings and opinions, communicated through text, with respect to a certain context (Kumar & Abraham, 2017; Kumar & Teeja, 2012). The categorization is carried out along the lines of polarities, such as positive and negative, etc. (Kumar & Sebastian, 2012; Kumar & Sharma, 2017).

Sentiment analysis, also known as opinion mining, is the means of recognizing and designating opinions communicated through a written piece to ascertain the author's connotation (positive, objective or negative) of that piece using a combination of statistical and computational techniques (Kumar & Jaiswal, 2017).

The core module of the Sentiment Analysis process employs feature extraction, a process used to convert input data, consisting of text indicating opinions, into an array of features, which can represent the input data very well (Kumar & Khorwal, 2017). Feature Selection is a technique used to select a subset of relevant features, discarding nonessential attributes (Kumar & Rani, 2016). Effective and efficient feature selection affects the quality of sentiments extracted and hence the classifier performance. But it has been observed that many features exist which don't contribute to accuracy, and thus can be removed without causing much loss. Fewer features reduce the complexity of the analysis, facilitating optimization.

Many researchers have adopted metaheuristic or stochastic methods for employing efficacious feature selection (Kumar, Khorwal, & Chaudhary, 2017). Metaheuristic methods exploit the trade-off which exists between a relatively robust solution and computational effort. Swarm intelligence-based stochastic methods are distinctly attractive for feature selection. Swarm Intelligence is the area of artificial intelligence that deals with systems composed of multiple entities called agents that correlate using self-organization and localized control. Agents are governed by simple rules and their behaviours are governed by their actual roles they play in their natural habitat. Movement of individual agents is decentralized, however, interaction between agents' results in a universal intelligent behaviour.

Cuckoo Search (CS) algorithm is a nature inspired, metaheuristic *optimization algorithm* which belongs to a group of swarm intelligence algorithms (Yang & Deb, 2009). The algorithm takes its inspiration from the cuckoo birds' parasitic practice of laying their eggs in the nests of hosts. The primary objective is to combine a set of binary coordinates for each solution, signifying if a particular feature belongs to the subsequent group of features or not. A classifier is trained with the selected features, encoded by the significance of the eggs. The solution's quality is then determined by evaluating each nest (Yang & Deb, 2009).

Recent literature has shown that CS algorithm has been surveyed as being more computationally efficient than PSO (Adnan & Razzaque, 2013).

Pereira et al. (2014) have developed a binary adaptation of CS algorithm, named Binary Cuckoo Search (Bcs). BCS is designed specifically to achieve optimum feature selection. It is the modified variant of the generic Cuckoo Search (CS) algorithm, which outputs the subset of features that are most efficient in classification.