Link based BPSO for feature selection in big data text clustering

Neetu Kushwaha, Millie Pant

PII: S0167-739X(17)32185-4
DOI: https://doi.org/10.1016/j.future.2017.12.005
Reference: FUTURE 3843

To appear in: *Future Generation Computer Systems*

Received date : 27 September 2017
Revised date : 24 November 2017
Accepted date : 3 December 2017

Link based BPSO for feature selection in Big data text clustering

Neetu Kushwaha,Millie Pant

Department of ASE, Indian Institute of Technology Roorkee, Roorkee 247001, India

neetumits@gmail.com, millidma@gmail.com

Abstract

Feature selection is a significant task in data mining and machine learning applications which eliminates irrelevant and redundant features and improves learning performance. This paper proposes a new feature selection method for unsupervised text clustering named link based particle swarm optimization(LBPSO). This method introduces a new neighbour selection strategy in BPSO to select prominent features. The performance of traditional particle swarm optimization(PSO)is limited by using global best updating mechanism for feature selection. Instead of using global best, LBPSO particles are updated based on neighbour best position to enhance the exploitation and exploration capability. These prominent features are then tested using *k*-means clustering algorithm to improve the performance and reduce the cost of computational time of the proposed algorithm. The performance of LBPSO are investigated on three published datasets, namely Reuter 21578, TDT2 and Tr11.Our results based on evaluation measures show that the performance of LBPSO is superior than other PSO based algorithms.

Keywords: Big Data, Text Clustering, Particle Swarm Optimization, Scale Free Network, *k*-means, Feature Selection

1. Introduction

In recent years, there has been a continuous growth of internet technology resulting in tremendous amount of text information. It is very difficult to process this text information manually and to extract valuable information from document corpus in time. Finding desired information in an age of 'big data' has been a challenge for standard information retrieval technology. Text analysis in the domain of text mining requires complex techniques to deal with numerous text documents. Text clustering (TC) is one of the most efficient techniques used in text mining domain, machine learning, and pattern recognition. With a good text clustering method, computers can automatically organize a document corpus into several hierarchies of semantic clusters[1]. It is a process of organizing documents into meaningful groups in such a way that documents of the same group are more similar to one another than documents belonging to different groups. Same topic is shared by the text documents that are in same cluster and different clusters represent different topics.

In order to apply text clustering algorithm, we need to transform these raw text documents into numerical format which consists of document's characteristics. To extract interesting patterns and insights from them, most fundamental and crucial step is document representation. In text clustering, documents are represented by their terms. Terms are either single term or the multi-word terms .Vector space model (VSM) is a very commonly used model for Document representation in text clustering[2]. In VSM, each term in the document is considered as feature /dimension.

Text documents contain high dimensional informative and uninformative (irrelevant, redundant, and noisy) features[3]. High dimensionality is always an ultimate challenge in the text document. Text clustering algorithms do not accomplish any type of feature selection(FS) method. Moreover, dimension reduction fails because of a vast number of text features and uninformative text features[4]. The efficiency of text clustering is affected by the dimensionality of text documents. The purpose of FS algorithms is to remove irrelevant or redundant features from the original set of features without sacrificing the prediction accuracy and computational time and to find a new subset of relevant features[5]. As we decrease the dimension of the text documents, the accuracy of the clustering algorithm increases. Feature selection techniques not only increase the clustering accuracy and efficiency but also reduces cost of computational complexity[6].

Feature selection methods can be broadly categorized into three types based on the different strategies of searching: filter, wrapper and Embedded methods. In filter method, it defines the relevant features without using any learning algorithm. On the other hand, wrapper method uses learning method to select informative feature. Generally, wrapper method outperforms filter method in terms of classification accuracy. Embedded methods integrate both feature selection methods to the learning model so as to achieve high accuracy or good performance with moderate computational cost (e.g. support vector machines and least square regression). Wrapper method can itself be broadly classified into two algorithms-sequential and heuristic algorithms. In sequential method, we start with empty set and add some features in it at every step until maximum objective function value is achieved. While, in heuristic search, evaluation is performed on different subsets of features to optimize the objective function value.

In the literature, several features selection methods have been introduced such as mutual information[7], sequential search algorithms[8] etc. Five features selection methods including χ^2 statistic document frequency, term strength, information gain (IG), and mutual information have been compared by Yang and Pedersen[9] . Principal component analysis (PCA) have been successfully used in text categorization[10], [11], Neural networks have also been widely applied by many researchers[11]–[14].