Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering

Q1 Kusum Kumari Bharti*, Pramod Kumar Singh

Computational Intelligence and Data Mining Research Lab, ABV-Indian Institute of Information Technology and Management Gwalior, Morena Link Road, Gwalior, MP, India

ARTICLE INFO

Article history:
Received 31 March 2014
Received in revised form 18 January 2016
Accepted 18 January 2016
Available online xxx

Keywords:
Text clustering
Binary particle swarm optimization
Mutation
Chaotic map
Opposition-based learning

ABSTRACT

Due to the ever increasing number of documents in the digital form, automated text clustering has become a promising method for the text analysis in last few decades. A major issue in the text clustering is high dimensionality of the feature space. Most of these features are irrelevant, redundant, and noisy that mislead the underlying algorithm. Therefore, feature selection is an essential step in the text clustering to reduce dimensionality of the feature space and to improve accuracy of the underlying clustering algorithm. In this paper, a hybrid intelligent algorithm, which combines the binary particle swarm optimization (BPSO) with opposition-based learning, chaotic map, fitness based dynamic inertia weight, and mutation, is proposed to solve feature selection problem in the text clustering. Here, fitness based dynamic inertia weight is integrated with the BPSO to control movement of the particles based on their current status, and the mutation and the chaotic strategy are applied to enhance the global search capability of the algorithm. Moreover, an opposition-based initialization is used to start with a set of promising and well-diversified solutions to achieve a better final solution. In addition, the opposition-based learning method is also used to generate opposite position of the gbest particle to get rid of the stagnation in the swarm. To prove effectiveness of the proposed method, experimental analysis is conducted on three different benchmark text datasets Reuters-21578, Classic4, and WebKB. The experimental results demonstrate that the proposed method selects more informative features set compared to the competitive methods as it attains higher clustering accuracy. Moreover, it also improves convergence speed of the BPSO.

1. Introduction

Q2

The volume of the text documents in digital format is gradually increasing and the text clustering has become a key paradigm to arrange digital text documents. The primary aim of a text clustering algorithm is to create clusters of the documents based on their intrinsic characteristics. Typical text clustering framework consists of a vector space model (VSM) representation that represents the documents in a common standard format and a clustering process that performs grouping of the text documents.

Majority of the text clustering paradigm employs bag-of-words approach [69], where each distinct term present in the documents' collection is considered as a feature for the document's representation. Hence, a document is represented by a multi-dimensional feature space where the cell value of each dimension corresponds to a weighted value, e.g., TF-IDF [68], of the concerning term within the document. Since the features are generated from distinct terms, even the moderate-sized documents in a text collection would result in hundreds of thousands of features. The clustering algorithms that do not perform feature reduction fail miserably not only because of large number of features but also because the non-informative (irrelevant, redundant, and noisy) features mislead and slow the algorithms. Therefore, feature selection, which removes non-informative features and selects a discriminative subset of features, has become a crucial preprocessing step in the text clustering to improve performance of the underlying algorithm and to speed up the computation. Feature selection is a combinatorial optimization problem with an aim to maximize accuracy of the underlying algorithm and minimize the number of features. As exhaustive search, a brute and force paradigm to solve combinatorial optimization problems, is not practically possible for larger problems owing to unreasonably high computational complexity, a number of feature selection methods have been proposed and their effectiveness have been studied to simplify the task.

* Corresponding author.
E-mail addresses: kkusum.bharti@gmail.com (K.K. Bharti), pksingh@iiitm.ac.in (P.K. Singh).

http://dx.doi.org/10.1016/j.asoc.2016.01.019
1568-4946/© 2016 Published by Elsevier B.V.

Please cite this article in press as: K.K. Bharti, P.K. Singh, Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering, Appl. Soft Comput. J. (2016), http://dx.doi.org/10.1016/j.asoc.2016.01.019

According to the search strategy to obtain an informative subset of the features, the existing feature selection methods are classified into three categories: filter, wrapper, and hybrid. Filter methods perform statistical analysis of the feature set to select a discriminative subset of the features without considering an interaction with the learning algorithm. Methods in this category include the information gain [63], document frequency [47], term strength [91], mean-median [27], mean absolute difference [27], mutual information [61], chi-square [46], and odd ratio [52]. Due to less computational complexity, these methods are widely used for feature selection, especially in the case of very high dimensional feature space, e.g., text. Wrapper methods employ search strategy to obtain subsets of the features and use learning algorithm to evaluate effectiveness of the obtained subsets and select the optimal feature subset. Methods in this category include sequential forward selection [62], sequential backward elimination [62], and plus-l-take-away-r-process [54,75]. Though these methods are computationally expensive in comparison to the filter methods, they attain comparatively higher accuracy. Another category of the feature selection is the hybrid method. Hybrid methods integrate different feature selection methods for informative feature subset selection. It takes the advantage of one method and lessen the drawback of the other for the feature subset selection. Recently, hybrid methods have received considerable attention for feature selection due to their feature selection characteristics [11,12,34,88,92,93].

Feature selection is a NP-hard combinatorial optimization problem [14,26,53,84]. An exhaustive search is the best way to obtain ideal solution(s) for the combinatorial optimization problems. However, performing an exhaustive search over the entire solution space is not practical as it involves unreasonably high computational complexity. In recent decades, many researchers have explored meta-heuristic algorithms to solve the combinatorial optimization problems [41,65,73,96] and these algorithms are gaining popularity day by day owing to their global search capability. Genetic algorithm [31], ant colony optimization (ACO) [20], and artificial bee colony (ABC) [38] are some well explored methods of this category and these algorithms have been widely used for feature selection problem [30,37,74,76,83].

Particle swarm optimization (PSO) is a swarm-intelligence based meta-heuristic search and optimization method. It simulates the social behavior of organisms such as bird flocking and fish schooling. Owing to its simplicity, information sharing capability, fast converge, and population based natures, the PSO-based algorithms are successfully explored in different domain such as power flow analysis [2], bioinformatics [1], image processing [66], financial decisions [50], data clustering [18]. In this paper, BPSO is used for feature selection problem. However, it has been used by other researchers to solve the feature selection problem in other domains. Recently, several of its variants such as the chaotic BPSO [18], the catfishBPSO [17], the correlation and taguchi chaotic BPSO [19], and the hamming distance BPSO [7] have been proposed for the feature selection. Chuang et al. [18] propose chaotic maps (logistic map and tent map) based methods to improve search ability of the BPSO. Chaotic map has an ergodicity, stochastic and regularity properties. These properties of a chaotic system provide a good exploration of the search space at the cost of exploitation. Chuang et al. [17] present an improved version of the BPSO for feature selection problem. The authors introduce a concept of catfish effect, which replaces 10% worst fit particles by the particles initialized at extreme points in the search space when the fitness of the global best particle (*gbest*) does not improve for a defined number of iterations. Banka and Dara [7] present a hamming distance based BPSO (HDBPSO) for feature selection problem. The authors use hamming distance as a proximity measure to update velocities of the particles in the BPSO. Experimental analysis is conducted on three different benchmark text datasets. The results and study favour their proposed method HDBPSO to the competitive methods.

Tan et al. [77] use genetic algorithm for feature selection problem. Akadi et al. [24] use GA for gene selection in a microarray data. Crossover and mutation are strong components of the GA. These components are responsible for exploration and exploitation, respectively in the algorithm. Some researchers integrate crossover and/or mutation to improve the search ability of the PSO. Ghamisi et al. [30] integrate GA with PSO for feature selection problem. Zhang et al. [95] use bare bone PSO for feature selection problem. In this approach, a reinforced memory strategy is used to update local leaders of the particles. Moreover, a uniform combination of crossover and mutation is proposed to balance the exploration and exploitation of the algorithm.

As discussed above, the filter methods are computationally more efficient than the wrapper methods, however, the wrapper methods attain better accuracy compared to the filter methods. Therefore, to integrate advantage of the one method and lessen drawback of the other, various researchers present integrated methods named hybrid or two stage methods for feature selection [24,83]. These methods select informative subset of features from the original feature space, thereby decrease computational complexity and increase performances of the underlying algorithms.

Chunag et al. [19] present correlation and taguchi chaotic binary particle swarm optimization for feature selection. In this approach, correlation approach is used to reduce dimensionality of the feature space and then taguchi chaotic binary particle swarm optimization is used to further refine the selected feature subspace. Zhang et al. [94] combine reliefF [42] and MRMR for gene selection problem. Akadi et al. [24] introduce a two-stage selection process for gene selection. They combine MRMR with GA to create an informative genes' subset. Initially, they use MRMR to filter out noisy and redundant genes from high dimensional space and then use GA to select a subset of relevant discriminative features. In their study, the authors use support vector machine (SVM) and naive bayes (NB) classifiers to estimate fitness of the selected features. Their experimental results show that their model is able to select the smallest gene subset that obtains the most classification accuracy in leave-one out cross-validation (LOOCV).

Initialization of the particles' positions, update of the particles, and parameters' values update play an important role to control search capability of the BPSO. Hence, in this paper, we aim to cover each possible aspect of the BPSO to improve its search capability. Population initialization plays a crucial role in the population based methods as it affects convergence speed and quality of the final solution. Different initialization strategies have been tested with the PSO to improve its performance [60,89]. In this paper, we use opposition-based strategy for population initialization. Moreover, it is also used to avoid stagnation of the swarm at the local optimal solution when the value of the gbest does not change for a defined number of iterations. The parameters of the BPSO play an important role to improve its search capability. Inertia weight is one of the key parameter to achieve this objective. Various adaptive/dynamic inertia weight criteria have been introduced by the researchers. A comprehensive review of different inertia weights is presented by [8,55]. In this paper, a new fitness based dynamic inertia weight is introduce to dynamically control the value of inertia weight of each particle based on its current status. It assigns high inertia weights to the low fit particles to explore distinct regions in the search space (exploration) and assigns low inertia weights to the high fit particles to exploit the region in their vicinity (exploitation). Instead of using uniform random number generation function, a chaotic map is used in velocity update equation for random number generation due to its ergodicity, stochastic and deterministic dynamic behavior. At the end, mutation and opposition-based strategy are applied

Please cite this article in press as: K.K. Bharti, P.K. Singh, Opposition chaotic fitness mutation based adaptive inertia weight BPSO for feature selection in text clustering, Appl. Soft Comput. J. (2016), http://dx.doi.org/10.1016/j.asoc.2016.01.019