Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 115 (2017) 258–265

Procedia Computer Science

www.elsevier.com/locate/procedia

7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India

Feature Selection using Gravitational Search Algorithm for Biomedical Data

Sushama Nagpal, Sanchit Arora, Sangeeta Dey*, Shreya

Netaji Subhas Institute of Technology,Sector-3, Dwarka, Delhi-110078, India

Abstract

Analysis of medical data for disease prediction requires efficient feature selection techniques, as the data contains a large number of features. Researchers have used evolutionary computation (EC) techniques like genetic algorithms, particle swarm optimization etc. for FS and have found them to be faster than traditional techniques. We have explored a relatively new EC technique called gravitational search algorithm (GSA) for feature selection in medical datasets. This wrapper based method, that we have employed, using GSA and k-nearest neighbors reduces the number of features by an average of 66% and considerably improves the accuracy of prediction.

Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications.

Keywords: Feature Selection;Gravitational Search Algorithm; Evolutionary Computation;Curse of Dimensionality; Biomedicine

1. Introduction

A majority of medical datasets use various features to describe the patient records. Certain features may be redundant, useless or have similar predictive power as that of other features. This voluminous data requires a large memory space, more computational time and effort and may also degrade the accuracy of the classifier. To effectively analyze these high dimensional, multivariate data, we need to look towards efficient feature selection(FS)

* Corresponding author. Tel.: +91 9868129869.
E-mail address: sangeetad.co@nsit.net.in

1877-0509
Peer-review under responsibility of the scientific committee of the 7th International Conference on Advances in Computing & Communications
10.1016/j.procs.2017.09.133

techniques. FS essentially refers to choosing a minimal subset of features out of the total features used to describe the data[3]. FS thus helps reduce redundancy and irrelevancy, and also helps in improving the performance of the classifier in several cases.

FS techniques are categorised into filter based, wrapper based and embedded methods[1]. Filter methods do not depend on the classifier for selecting features. They can be used to scale large datasets better[3][7]. Wrapper methods, however make use of a classifier for selection of features. The feedback from the classifier helps in choosing the best subset of features, thereby improving the overall performance[7]. Embedded methods can be addressed as a special class of wrapper based methods, in which the FS technique is suited to only a particular classifier. Relatively less research has been done on embedded methods[2]

Traditional methods for FS are based on searching the workspace to generate appropriate subsets. The search may be exhaustive search, wherein for n features, all the 2n subsets are generated and evaluated. This is practically infeasible, due to increased computation time, even when we are dealing with small data. Other search methods like random search, greedy search and heuristic search have also been adopted. However, many of these searching algorithms may stagnate in local optima, in addition to being computationally expensive[3].

In order to minimize the above problems, researchers are looking towards evolutionary computation (EC) techniques for FS. These are essentially a class of optimization algorithms which have an efficient global search technique[3]. Due to the advantages of EC over traditional methods, it has been used for FS in various domains including image analysis[15], face recognition[17], software fault[11] etc. It has also been used for biomedical datasets[19] and has shown significant reduction of number of features. To analyze these domains, researchers have used techniques like particle swarm optimization (PSO)[4], genetic algorithm(GA)[5], ant colony optimization (ACO)[17] among others.

Gravitational search algorithm (GSA), put forward by Rashedi et al [8], is a relatively new optimization technique, which is structurally similar to PSO. The latter uses memory, in the form of pbest (particle best) and gbest(global best), but GSA is memoryless which is an added advantage over PSO. The binary form of GSA, i.e BGSA[12], has been used for FS in the past and has shown promising results in various domains[12]. It has been relatively less explored in the field of medical diagnosis[10][16]. Therefore, in this paper we focus our work on GSA for FS in medical datasets.

Our first objective is to reduce the number of attributes used to describe the patient records whereas our second objective is to increase the accuracy with which diseases are predicted. For such dual objective problems, population based EC techniques give good results[3]. To accomplish our objectives, we have used GSA along with k-nearest neighbors(knn) classifier to implement a wrapper based method. The output of the classifier is used as feedback to GSA(used for FS) thereby helping in selecting those attributes/features which help increase the classifier's accuracy[7]. An optimal feature set is thus chosen. In this work, we have also compared our results with that of earlier researchers who have performed FS using GA[14] and PSO[19].

The rest of the paper is presented as described below: - Section 2 gives an outline of the related works in this field. Section 3 explains the technique of GSA and its use in FS. Section 4 depicts the experimental setup that has been used as well as the results of the experiment. Section 5 concludes our work and gives an insight into the scope of future work in this domain.

2. Related Works

FS has become an integral part of data classification problems, owing to the large number of attributes that the real world datasets contain. It has been explored using traditional methods for various classification