Computers in Biology and Medicine 90 (2017) 125–136

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease

T. Vivekanandan[a,b,*], N. Ch Sriman Narayana Iyengar[b,c]

[a] Department of Computer Science and Engineering, SITAMS, Chittoor, 517127, Andhra Pradesh, India
[b] School of Computing Science and Engineering, VIT University, Vellore, 632014, Tamil Nadu, India
[c] Department of Information Technology, Sreenidhi Institute of Science and Technology, Yamnapet, Ghatkesr, Hyderabad, 501301, Telengana, India

ARTICLE INFO

Keywords:
Optimization
e-Healthcare
Feature selection
Differential evolution
Big data
Fuzzy AHP
Feed-forward neural network

ABSTRACT

Enormous data growth in multiple domains has posed a great challenge for data processing and analysis techniques. In particular, the traditional record maintenance strategy has been replaced in the healthcare system. It is vital to develop a model that is able to handle the huge amount of e-healthcare data efficiently. In this paper, the challenging tasks of selecting critical features from the enormous set of available features and diagnosing heart disease are carried out. Feature selection is one of the most widely used pre-processing steps in classification problems. A modified differential evolution (DE) algorithm is used to perform feature selection for cardiovascular disease and optimization of selected features. Of the 10 available strategies for the traditional DE algorithm, the seventh strategy, which is represented by DE/rand/2/exp, is considered for comparative study. The performance analysis of the developed modified DE strategy is given in this paper. With the selected critical features, prediction of heart disease is carried out using fuzzy AHP and a feed-forward neural network. Various performance measures of integrating the modified differential evolution algorithm with fuzzy AHP and a feed-forward neural network in the prediction of heart disease are evaluated in this paper. The accuracy of the proposed hybrid model is 83%, which is higher than that of some other existing models. In addition, the prediction time of the proposed hybrid model is also evaluated and has shown promising results.

1. Introduction

Due to the tremendous growth in the volume of healthcare data, data analytics in a healthcare information system can extract valuable information. Recently, healthcare organizations have been moving towards digitization [10,18,20] of the massive volume of healthcare data to leverage data analytics in healthcare to realize extensive benefits. The potential benefits include detection of diseases at earlier stages, cost reduction, personalized treatment, improved patient experience, and superior care [21]. A medical record is a file that contains information on patient identity and examinations, treatments, therapies and services given to the patient [22].

The results of the digitized translation of medical records are called electronic medical records (EMRs). EMRs [19] are extremely complex and of huge size. Managing such a huge volume of data becomes difficult with traditional data processing systems.

The healthcare datasets available in the EMRs are generated from different medical sources such as diagnoses, procedures, medications, and lab results [23,35,37] and are maintained under different attributes or features. Feature selection or attribute selection [24,25] has become an important focus in many research applications, for which datasets with tens or hundreds or thousands of variables are available.

Feature selection [1,2,6,15] can greatly improve the accuracy of the resulting classifier model. Furthermore, it is important in finding the relevant subset of predictive features. For example, a physician might take a decision on the criticality of a particular disease based on a classification carried out using the selected features. The accuracy of the prediction is improved by optimizing the feature selection.

Optimization is the process of obtaining the best possible values of decision variables based on the selected objective function [34]. In recent times, evolutionary algorithms have been extensively used in various fields for finding near-optimal solutions.

In this paper, feature selection is carried out by selecting the critical features that form the root cause for the objective function of the problem under consideration, which is the prediction of cardio-vascular disease. The heart disease dataset in the UCI data repository [8] is used for

* Corresponding author. SITAMS, D.K. Audikesavalu Marg, Tripati-Banglore Bye pass Road, Chittoor, 517127, AP, India.
E-mail addresses: stvanand@gmail.com, stvanand@sitams.org (T. Vivekanandan), nchsniyr@vit.ac.in (N.C. Sriman Narayana Iyengar).

https://doi.org/10.1016/j.compbiomed.2017.09.011
Received 15 April 2017; Received in revised form 15 September 2017; Accepted 16 September 2017

experimentation. The differential evolution (DE) algorithm, designed by Rainer Storn and Kenneth Price [3], is used as a base for optimizing the feature selection.

The differential evolution (DE) algorithm, which is a kind of evolutionary algorithm, has been adopted as an effective global optimizer [4,5,7]. It is a powerful population-based stochastic technique that is efficient over continuous space [9,11,12,14,17]. DE has been successfully implemented in many fields of engineering, such as mechanical engineering, communication, water resource management, and pattern recognition [13].

Out of the 10 strategies given by Price and Storn in the traditional DE algorithm, the seventh strategy, represented by DE/rand/2/exp, is proven to be the best [3]. A modification has been made to this traditional DE strategy and a comparative study is carried out to analyse the performance of the modified DE strategy.

The output of the modified DE strategy (selected critical features) has been analysed using fuzzy AHP [40] and a feed-forward neural network, and heart disease prediction has been performed. The effectiveness of conjugating the modified DE strategy with fuzzy AHP [52,53] and a feed-forward neural network in the prediction of heart disease has been analysed.

2. Background study

2.1. Feature selection

The process of narrowing down the feature set into a reduced feature subset is known as feature selection. It is a process that is commonly used in machine learning and is an optimization problem [3]. Feature selection is carried out based on one of the management principles, known as the Pareto principle. According to the Pareto principle, 80% of effects come from 20% of causes [36]. Hence, it is also known as the 80-20 rule. Based on the above principle, the accuracy of heart disease prediction is comparatively improved by selecting a few critical attributes.

Feature selection is a search technique used to define the feature subset. Feature selection, along with an evaluation method for scoring different evolved feature subsets [1], is necessary to obtain an optimal output. Hence, feature selection works hand in hand with an optimization algorithm. This optimization algorithm minimizes the error caused by the selection of irrelevant attributes. The optimization algorithm thus aims at obtaining the optimal feature subset. In this work, the differential evolution (DE) algorithm is used to optimize the feature selection. The best proven traditional DE algorithm has been modified and the outputs are compared to evaluate the performance of the modified DE strategy.

2.2. Differential evolution

The differential evolution (DE) algorithm belongs to the field of evolutionary programming. In this paper, the DE algorithm is effectively used to reduce the number of features used in a dataset. This algorithm, which was designed by Rainer Storn and Kenneth Price [4], is used to optimize the selection of attributes from the given data.

DE has a wide range of applications due to its simple structure, flexibility, speed and robustness. DE is one of the best genetic-type algorithms for solving problems with real-valued variables.

In this paper, the DE algorithm is applied to a heart disease dataset for optimal feature selection. This medical dataset has wide range of values, as the number of attributes associated with each disease or abnormality is huge, and each attribute has its own normal range. Handling such wide ranges of values is a difficult task. Hence, these wide ranges of values are converted into real values based on Min-Max Normalization [16]. A value that lies between 0 and 1 is assigned to each attribute value.

Differential Evolution is carried out in three basic steps:

Step 1- Mutation: It is the search mechanism used in the production of a mutant vector.
Step 2- Selection: DE uses selection to direct the search towards the prospective region.
Step 3- Crossover: It is a mechanism of probabilistic and useful exchange of information among solutions to locate better solutions.

DE is one of the best genetic-type algorithms for solving problems with real-valued variables. Among the ten different strategies suggested by Price and Storn, one of the strategies is denoted as DE/rand/2/exp, where ‘DE’ stands for differential evolution and ‘rand’ indicates that the vector to be perturbed is randomly selected, i.e., the selection of the target vector is made at random. The value ‘2’ indicates that two vectors are chosen at random and the weighted difference between the two is added to a third randomly chosen vector for perturbation. Finally, ‘exp’ indicates the use of exponential crossover, i.e., the crossover is performed on the ‘D’ variables in one loop until it is within the CR bound. In this paper, the traditional DE algorithm has been replaced by the modified DE strategy represented by DE/rand/2-wt/exp.

In the traditional DE algorithm, three individual vectors are selected at random, and weighted difference between the first two vectors is added to the third vector to determine the mutant vector. The mutant vector is then perturbed with the randomly chosen target vector. However, in the case of the modified DE algorithm, four vectors are selected sequentially. The weighted differences between the first two selected vectors and the next two vectors are computed in parallel and the resultant two weighted difference vectors are added to generate the mutant vector. This mutant vector is perturbed with the randomly chosen target vector.

2.3. Fuzzy analytic hierarchy process (AHP)

Fuzzy AHP is one of the most prominent structured techniques used for decision making in fields such as industry, healthcare, government, and education. Due to its strong application in group decision making, the AHP technique is used around the world for a wide range of applications due to its high accuracy [53,55].

In this paper, prediction of heart disease is carried out using a feed-forward neural network integrated with fuzzy AHP in the following steps:

Step 1: The hierarchical structure showing the ultimate goal of the problem (here, the diagnosis of heart disease) and its alternatives is drafted.
Step 2: A pair-wise comparison of each attribute with the others is performed using a comparison matrix. The alternatives of each attribute are represented using a fuzzy triangular membership function prior to pair-wise comparison. Using this decomposition, the local weight and global weights are obtained, which are used to determine the criticality of the features.

2.4. Feed-forward neural network

A feed-forward neural network is a type of neural network in which the information moves in a single direction. The flow of information is from the input nodes to the output nodes, through the hidden nodes [59–62].

In this paper, the selected critical features from the modified DE algorithm are the input nodes. The alternatives of each attribute are then determined. The local weights of each alternative and the global weights of each attribute are calculated using the fuzzy AHP, as described in the previous section.

The steps involved in the feed-forward neural network model are as follows:

Step 1: Determine the hierarchical model as in the first step of the fuzzy AHP technique.
Step 2: Initialize all the local weights of the alternatives, global weights of the attributes and biases to random values, which are determined through fuzzy AHP.