Expert Systems with Applications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A highly accurate firefly based algorithm for heart disease prediction

Nguyen Cong Long [a,*], Phayung Meesad [a], Herwig Unger [b]

[a] Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Thailand
[b] Faculty of Mathematics and Computer Science, FernUniversitat in Hagen, Germany

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Rough sets
Attribute reduction
Feature selection
Chaos firefly algorithm
Type-2 fuzzy logic system

ABSTRACTS

This paper proposes a heart disease diagnosis system using rough sets based attribute reduction and interval type-2 fuzzy logic system (IT2FLS). The integration between rough sets based attribute reduction and IT2FLS aims to handle with high-dimensional dataset challenge and uncertainties. IT2FLS utilizes a hybrid learning process comprising fuzzy c-mean clustering algorithm and parameters tuning by chaos firefly and genetic hybrid algorithms. This learning process is computationally expensive, especially when employed with high-dimensional dataset. The rough sets based attribute reduction using chaos firefly algorithm is investigated to find optimal reduction which therefore reduces computational burden and enhances performance of IT2FLS. Experiment results demonstrate a significant dominance of the proposed system compared to other machine learning methods namely Naive Bayers, support vector machines, and artificial neural network. The proposed model is thus useful as a decision support system for heart disease diagnosis.

1. Introduction

Heart disease has significantly increased over the last decade and has become the leading cause of death for people in most countries around the world. There are many features of heart disease affecting the structure or function of the heart. These might be difficult for doctors to diagnose quickly and accurately. Therefore, it is necessary to employ computerized technologies in heart disease diagnosis to assist doctors to diagnose faster with higher accurately.

Currently, there are many heart disease diagnosis systems relying on soft computing techniques that have been proposed. In particular, integrating the use of several soft computing techniques to generate hybrid models have been investigated in order to perform better than a single technique. These models usually consisted of two states. In the first state, feature selection techniques are applied to select subset of features. The obtained subset of features is subsequently used as input for the classification techniques in the second state (Avci, 2009; Guan, Gray, & Leyffer, 2009; Khemphila & Boonjing, 2011; Nahar, Tasadduq, Kevin, & Yi-Ping, 2013; Sanz et al., 2013; Shilaskar & Ghatol, 2013).

Due to many features of heart datasets, which contain relevant as well as irrelevant and redundant features. Irrelevant features do not influence description of the target class. Redundant features do not contribute to anything but they make noise towards description of target class (Shilaskar & Ghatol, 2013). Those features not only affect the results of classification but also make the system run slowly. Therefore, removing those features before applying classifier techniques is necessary. For this purpose, attribute reduction or feature selection is needed in the heart disease diagnosis system. This reduces the risk of over fitting, improves generalization ability of the model, provides better predictability, and requires less computation causing smaller features (Shilaskar & Ghatol, 2013).

Many researchers have investigated feature selection techniques for heart disease diagnosis in past literature. Guan et al. (2009) proposed a feature selection based on support vector machine for medical datasets. For the SPECTF dataset, experiments shown when 12 out of 44 features were selected and accuracy of the proposed model was 76.5%. Moreover, the proposed model was compared to the standard support vector machine (SVM), recursive feature elimination (RFE) – SVM, L1-norm SVM, and two approximated L0-norm SVM methods. The results showed that their proposed model outperformed the other models in terms of accuracy.

Shilaskar and Ghatol (2013) proposed a heart disease diagnosis system consisting of feature selection and classification techniques. SVM classifier was combined with forward feature inclusion, back-elimination feature selection and forward feature selection for three datasets from the UCI datasets namely

* Corresponding author.
E-mail addresses: nguyenconglongvt@gmail.com (N.C. Long), pym@kmutnb.ac.th (P. Meesad), herwig.unger@fernuni-hagen.de (H. Unger).

http://dx.doi.org/10.1016/j.eswa.2015.06.024
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

Arrhythmia, SPECTF and Heart Disease datasets. Experimental results demonstrated that the feature selections improved accurately classification techniques and reduced the number of input variables. For the SPECTF dataset, accuracy of SVM increased 3% from 75% to 78% and the number of features reduced significantly from 44 to 19 features. For the heart disease dataset, the accuracy of SVM increased 4% from 81% to 85% and the number of feature reduced from 10 to 4 features.

Shao, Hou, and Chiu (2014) used logistic regression, multivariate adaptive regression splines and rough set techniques as feature selection to reduce the set of explanatory features for heart disease diagnosis. The remaining features were used as input for artificial neural network. The heart disease dataset form UCI datasets was used to train and test the models. The experiments showed that all of those techniques reduced the sets of features and improved classification performance. However, logistic regression only reduced one feature and increased approximately 2% of accuracy. For the rough set approach, this research used the greedy heuristics algorithm in the rough set exploration system (RESE). The result reduced two features and improved approximately 3% of accuracy. Multivariate adaptive regression splines combined with artificial neural network is the best system in this research, which achieved 82.14% accuracy with reduced 7 features.

Inbarani, Azar, and Jothi (2014) proposed a hybrid system which combined rough sets based feature selection using particle swarm optimization (PSO) and classification techniques for medical diagnosis. In this study, PSO based relative reduction and PSO based quick reduction were investigated to find optimal features. Then the results were used as input of Naïve Bayer, BayesNet and KStar techniques to classifier medical datasets. This proposed model was evaluated using Erythematic, Prognostic Breast Cancer, SPECTF datasets from UCI Repository for machine learning. The results illustrated that attribute reduction of both PSO based relative reduction and PSO based quick reduction achieved 15 out of 44 attributes and accuracy of the best classification technique namely Naive Bayer was approximate 88% for SPECTF dataset. However, this proposed approach did not select optimal number of attributes.

Nguyen, Abbas, Douglas, and Saeid (2015a) proposed a medical diagnosis system which were combined genetic fuzzy logic system with wavelet. The wavelet transformation was employed to extract discriminative features for high-dimensional datasets. Then fuzzy standard additive trained by genetic algorithm (GSAM) was applied to classifier medical dataset. This proposed model was evaluated using Wisconsin breast cancer and Cleveland heart disease datasets from UCI Repository for machine learning. The results shown that GSAM became highly capable when deployed with small number of wavelet features as its computational burden was reduced. However, this proposed approach had a shortcoming regarding selection of the optimal number of wavelet features and the accuracy of this proposed model was 78.78% for Cleveland heart disease datasets.

Nguyen, Abbas, Douglas, and Saeid (2015b) proposed an automated medical data classification using wavelet transformation (WT) and interval type-2 fuzzy logic system (IT2FLS). Fuzzy c-mean clustering algorithm was used to construct fuzzy rule based of the fuzzy system and genetic algorithm was applied to tune parameter of the fuzzy system. The WT was used to find a reduction of features therefore that reduces computational burden and enhances performance of IT2FLS. The proposed model was measured using Wisconsin breast cancer and Cleveland heart disease datasets from UCI Repository for machine learning. Results demonstrated a significant dominance of the wavelet–IT2FLS approach compared to other machine learning methods including probabilistic neural network, support vector machine, fuzzy ARTMAP, and adaptive neuron-fuzzy inference system. However, this proposed approach did not select optimal number of features and the accuracy of this proposed model was 81.01% for Cleveland heart disease datasets.

From the literature above, there are several limitations with existing models that can be resolved. Feature selection methods applied in the existing methods were heuristic feature selection. Those methods only considered heuristically selecting the subset of features, but redundant features may still remain and was not the best subset of features. Therefore, finding out the best minimal subset of features is needed. Recently, rough sets theory has been applied as a tool to discover data dependencies and reduce the number of attributes in inconsistent datasets (Pawlak, 1991). Rough sets are applied to attribute reduction to remove redundant attributes and select subsets of significant attributes, which lead to better prediction accuracy and speed than systems using original sets of attributes. Traditionally, greedy heuristics was applied to find attribute reductions based on rough sets (Hoa, 1996; Shi, Liu, & Zheng, 2004; Velayutham & Thangavel, 2011). Those approaches are quite fast. However, these may meet the problem of heuristic feature selection. Another alternative approach for attribute reduction based on rough sets is applied meta-heuristic algorithms to find the best minimal attribute reduction (Inbarani et al., 2014; Wang, Yang, Teng, Xia, & Jensen, 2007; Ye, Chen, & Ma, 2013). Firefly algorithm (FA) is one of the recent swarm intelligent techniques proposed by Yang (2008) and is a meta-heuristic algorithm that relies on flashing behavior of fireflies in nature to find global optimal solution in search space for special problems. FA has been successfully applied to a large number of difficult combinational optimization problems as well as NP-hard problems (Yang, 2008). Furthermore, particle swarm optimization (PSO) is just a special class of the firefly algorithms. In addition, the firefly algorithm is much more efficient in finding the global optima with higher success rates than PSO and genetic algorithms (GAs) (Yang, 2009). Therefore, firefly algorithm is applied in this research to combine rough sets for attribute reduction.

Hence, the main objective of this paper is to propose an efficiently heart disease diagnosis model to predict heart disease more accurately with reduced number of attributes. In the proposed model, the chaos firefly algorithm combined with rough sets is introduced to reduce the set of attributes. The remaining subsets of attributes are used as inputs for the type-2 fuzzy logic system. Furthermore, two comparisons are investigated in this research. Firstly, the rough sets based attribute reductions using the chaos firefly algorithm compared to the rough sets based attribute reductions using the binary particle swarm optimization. Secondly, type-2 fuzzy logic system compared to three well-known classification techniques namely Naive Bayes, support vector machine and artificial neural network.

The remainder of this paper is organized as follows: State-of-the-art systems are introduced in Section 2. Methodology is illustrated in Section 3. Experiment results and comparison of differential models are covered in Section 4. Finally, conclusions are summarized in Section 5.

2. State of the art

This section summarizes state of the art of heart disease diagnosis systems. Identified problems and suggested solutions for these problems are introduced in this section. Furthermore, the theoretical foundations are also given in this section.

2.1. Literature reviews

Recently, many soft computing techniques have been proposed for heart disease diagnosis. The heart disease diagnosis system