Applied Soft Computing 30 (2015) 812–822

Contents lists available at ScienceDirect

Applied Soft Computing

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

Medical data classification using interval type-2 fuzzy logic system and wavelets

Thanh Nguyen*, Abbas Khosravi, Douglas Creighton, Saeid Nahavandi

Centre for Intelligent Systems Research (CISR), Deakin University, Geelong Waurn Ponds Campus, Victoria 3216, Australia

ARTICLE INFO

Article history:
Received 2 July 2014
Received in revised form 10 January 2015
Accepted 5 February 2015
Available online 21 February 2015

Keywords:
Interval type-2 fuzzy logic system
Wavelet transformation
Genetic algorithm
Medical data classification
Breast cancer
Heart disease

ABSTRACT

This paper introduces an automated medical data classification method using wavelet transformation (WT) and interval type-2 fuzzy logic system (IT2FLS). Wavelet coefficients, which serve as inputs to the IT2FLS, are a compact form of original data but they exhibits highly discriminative features. The integration between WT and IT2FLS aims to cope with both high-dimensional data challenge and uncertainty. IT2FLS utilizes a hybrid learning process comprising unsupervised structure learning by the fuzzy c-means (FCM) clustering and supervised parameter tuning by genetic algorithm. This learning process is computationally expensive, especially when employed with high-dimensional data. The application of WT therefore reduces computational burden and enhances performance of IT2FLS. Experiments are implemented with two frequently used medical datasets from the UCI Repository for machine learning: the Wisconsin breast cancer and Cleveland heart disease. A number of important metrics are computed to measure the performance of the classification. They consist of accuracy, sensitivity, specificity and area under the receiver operating characteristic curve. Results demonstrate 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 neuro-fuzzy inference system. The proposed approach is thus useful as a decision support system for clinicians and practitioners in the medical practice.

1. Introduction

Healthcare data classification mechanism facilitates clinical practitioners in the diagnosis and treatment for medical diseases and conditions occurred in patients. It thus plays as a decision support system (DSS) substantially contributing to the healthcare quality improvement. The rapid development of data mining techniques enables the construction of clinical DSSs to be much more plausible than ever before.

Mazurowski et al. [1] investigated two methods of neural network training including classical backpropagation and particle swarm optimization, which were then applied for breast cancer diagnosis. A method called kernel F-score feature selection is presented in Polat and Güneş [2] for data pre-processing in the classification of medical datasets.

Peng et al. [3] proposed another technique for feature selection to deal with high-dimensional challenge of medical data classification. The approach integrates filter and wrapper methods into a sequential search procedure to improve the classification performance.

Alternatively, Nahar et al. [4] examined several computational intelligence techniques and identified the best algorithms for heart disease diagnosis. A process based on differential evolution for classifying items in medical databases is suggested in De Falco [5]. The tool automatically extracts explicit knowledge from the database under the form of if-then rules, which combine variables by the “and” operator.

Although many methods have been proposed, they can only provide nonintuitive classification results, which are not comprehensive and applicable to clinicians in the real practice, e.g. see [6,7] for applications of the black box artificial neural networks. It is therefore helpful if the behaviours of classification techniques are understood by humans via tools like linguistic rules. Moreover, the huge amount of data collected by healthcare practice is too complex and voluminous for processing and analysis.

Medical diagnosis and prognosis are decision making problems that commonly involve complexity and uncertainty. The use of fuzzy logic thus has been advocated to handle uncertainty and provide clinicians intuitive classification results via linguistic rules. A number of type-1 fuzzy approaches have been introduced for

* Corresponding author. Tel.: +61 3 52278281; fax: +61 3 52271046.
E-mail address: thanh.nguyen@deakin.edu.au (T. Nguyen).

http://dx.doi.org/10.1016/j.asoc.2015.02.016
1568-4946/

medical data classification, e.g. see [8–11]. Criticism was made concerning the fact that the membership function of type-1 fuzzy sets has no uncertainty associated with it. Type-2 fuzzy sets generalize type-1 fuzzy sets so that more uncertainty can be handled. A type-2 fuzzy set allows us to incorporate uncertainty about the membership function into fuzzy set theory [12].

Type-2 fuzzy logic systems (T2FLS) therefore have emerged as a more powerful tool to cope with uncertainty compared to traditional type-1 fuzzy systems. Among T2FLSs, the use of interval type-2 fuzzy logic system (IT2FLS) has demonstrated successfully in a number of applications, e.g. see [13–19]. Using fuzzy system in general or IT2FLS in particular often encounters the curse of dimensionality. The number of rules is exponentially proportional to the number of inputs of the fuzzy system. A large number of rules not only increases the computational complexity, but also limits the interpretability of the system. This problem becomes more severe as medical data are commonly assembled in high dimensions. Consequently, a dimension reduction or feature extraction tool should be implemented before the IT2FLS is executed.

Wavelet transformation (WT) and principal component analysis (PCA) are popular feature extraction methods that can be applied to high-dimensional datasets. With datasets having dimensions reduced, fuzzy systems would demonstrate more powerful ability in function approximation or classification.

The main caveat of PCA is that eigenvectors corresponding to the largest variance of the data are selected, but these directions do not necessarily provide the best separation of the classes being distinguished. In other words, there is a possibility that the information for separating the clusters is represented in principal components with low eigenvalues, which are often ignored. The use of WT is thus promoted in this research.

To deal with uncertain and high-dimensional medical data, this paper proposes a method using IT2FLS, fuzzy c-means clustering (FCM), genetic algorithm (GA) and wavelet features for healthcare data classification. To our best knowledge, it is the first application of IT2FLS method for medical diagnosis and also the first combination of wavelet features and IT2FLS for classification. Through this study, we examine and compare performance of the proposed wavelet–IT2FLS model with other classification methods frequently applied in literature. Experiments are conducted using two benchmark medical datasets to make sure conclusions driven out of this study are valid and general.

The rest of the paper is organized as follows. Section 2 presents the background of IT2FLS and its training process by FCM and GA. Wavelet transformation (WT) is described in Section 3. Section 4 is devoted for experimental results, which are followed by concluding remarks in Section 5.

2. Type-2 fuzzy logic systems

2.1. Background to fuzzy logic

A fuzzy logic system (FLS) is called a type-1 FLS (T1FLS) if it is described completely using type-1 fuzzy sets (T1 FSs) whilst a FLS that uses at least one type-2 fuzzy set (T2 FS) is called a T2FLS [20,21]. A T2FLS has more degrees of freedom than does a T1FLS because it comprises more parameters. It therefore suggests that T2FLS has the potential to outperform a T1FLS in handling uncertainties because of its larger number of design degrees of freedom. If uncertainties disappear, a T2FLS reduces to a T1FLS [22]. The structure of a general T2FLS is illustrated in Fig. 1.

T2FLS structure is similar to that of the T1FLS with the major differences being in the use of T2 FSs (rather than T1 FSs) in antecedent parts of fuzzy rules and the output processor. The output processor of a T1FLS transforms a T1 FS to a crisp number whilst a T2FLS has two components in the output processor. The first is a type reduction that transforms a T2 FS into a T1 FS and the second is the defuzzifier that transforms a T1 FS into a crisp number. A general T2FLS requires extensive computational cost and complicated implementation compared to a T1FLS. A special case of T2FLS, interval type-2 FLS (IT2FLS) has been widely used for reduced computational burden [23,24].

2.2. Interval type-2 fuzzy logic systems

T2FLSs [25] are constructed on the basis of T2 FSs. A general T2 FS is represented in three dimensions. The membership degree (MD) is not a crisp number but it is a FS. Third dimension is the value of the membership function (MF) at each point on footprint of uncertainty (FOU), which is the two-dimensional domain.

Interval type-2 FSs (IT2 FSs) represent the MD by an interval rather than a FS. The third dimension value in the IT2 FS is the same everywhere so that it is ignored and only the FOU is used to describe the IT2 FS.

Based on different fuzzy rule types, e.g. Mamdani, Takagi-Sugeno-Kang (TSK) or Tsukamoto, there are corresponding different IT2FLSs can be established. In this paper, the special case of TSK fuzzy rule is employed to develop an IT2FLS for medical data classification.

Some variants of IT2FLS are recognized depending on the MFs used in the antecedent and consequent parts are IT2 FSs and/or T1 FSs. We implement herein the IT2FLS where antecedents are IT2 FSs and consequents are interval T1 FSs. Assume the IT2FLS consists of K rules and p antecedents in each rule, the lth rule by R^l is denoted as follows:

$$R^l : \text{IF } x_1 \text{ is } \tilde{F}_1^l \text{ and} \ldots \text{and } x_p \text{ is } \tilde{F}_p^l, \quad \text{THEN } Y^l = C^l \tag{1}$$

where $l = 1, \ldots, K$. $\tilde{F}_i^l$ is the ith IT2 FS defined by a lower and upper bound MF:

$$\mu_{\tilde{F}_i^l}(x_i) = [\underline{\mu}_{\tilde{F}_i^l}(x_i), \bar{\mu}_{\tilde{F}_i^l}(x_i)] \tag{2}$$

and C^l is an interval T1 FS characterized by its centre and spread c^l and s^l respectively:

$$C^l = [c^l - s^l, c^l + s^l] \tag{3}$$

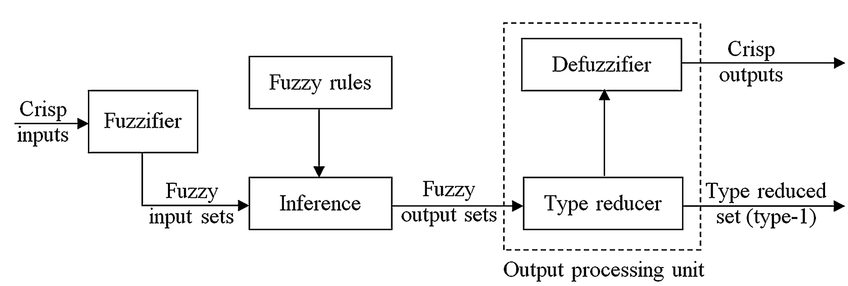

Fig. 1. Structure of type-2 FLS [22].