Expert Systems with Applications 42 (2015) 2184–2197

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Classification of healthcare data using genetic fuzzy logic system and wavelets

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

Centre for Intelligent Systems Research, Deakin University, Waurn Ponds, Victoria 3216, Australia

ARTICLE INFO

Article history:
Available online 23 October 2014

Keywords:
Fuzzy standard additive model
Genetic algorithm
Wavelet transformation
Healthcare data classification
Breast cancer
Heart disease

ABSTRACT

Healthcare plays an important role in promoting the general health and well-being of people around the world. The difficulty in healthcare data classification arises from the uncertainty and the high-dimensional nature of the medical data collected. This paper proposes an integration of fuzzy standard additive model (SAM) with genetic algorithm (GA), called GSAM, to deal with uncertainty and computational challenges. GSAM learning process comprises three continual steps: rule initialization by unsupervised learning using the adaptive vector quantization clustering, evolutionary rule optimization by GA and parameter tuning by the gradient descent supervised learning. Wavelet transformation is employed to extract discriminative features for high-dimensional datasets. GSAM becomes highly capable when deployed with small number of wavelet features as its computational burden is remarkably reduced. The proposed method is evaluated using two frequently-used medical datasets: the Wisconsin breast cancer and Cleveland heart disease from the UCI Repository for machine learning. Experiments are organized with a five-fold cross validation and performance of classification techniques are measured by a number of important metrics: accuracy, F-measure, mutual information and area under the receiver operating characteristic curve. Results demonstrate the superiority of the GSAM 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 helpful as a decision support system for medical practitioners in the healthcare practice.

1. Introduction

The automatic diagnosis of heart disease and breast cancer is an important, real-world medical problem. Heart disease affects health and working performance of patients, especially old people. The World Health Organization has estimated that 12 million deaths occur worldwide every year because of the heart diseases (Soni, Ansari, Sharma, & Soni, 2011). The heart disease actually can be detected early by performing a number of medical tests. However, these tests are often costly and confronted a certain difficulty. An inexpensive solution based on medical history of patients and some simple tests is commonly proposed by heart disease investigators.

Breast cancer is one of the largest causes of cancer deaths among women. At the same time, it is also among the most treatable cancer types if it can be diagnosed early. Early prediction of the characteristic of breast lumps (benign or malignant) occurring in patients thus helps to determine a suitable treatment for the cancer.

There have been a number of studies dealing with medical diagnosis in general or breast cancer or heart disease prediction in particular in the literature. Verma and Hassan (2011) proposed a hybrid ensemble approach using unsupervised learning strategies and parallel data fusion techniques for classification in medical databases. Marcano-Cedeño, Quintanilla-Domínguez, and Andina (2011) presented a novel improvement in neural network training for classifying the breast cancer lesions as benign or malignant. The method encompasses simulating the biological property of metaplasticity on multilayer perceptron with backpropagation.

On the other hand, Dangare and Apte (2012) suggested a system using medical terms such as sex, blood pressure, cholesterol, obesity and smoking attributes to predict the likelihood of patient getting a heart disease. Bhatla and Jyoti (2012) also investigated a number of machine learning methods for automated heart disease prediction systems. Likewise, Sundar, Latha, and Chandra (2012) introduced a prototype using naïve Bayes and weighted associative classifier for heart disease diagnosis.

* 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.eswa.2014.10.027
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

Similarly, Lakshmi, Krishna, and Kumar (2013) considered several data mining techniques and constructed a web based user friendly system for predicting heart disease survivability. Recently, Seera and Lim (2014) proposed a hybrid intelligent system that consists of the fuzzy min–max neural network, the classification and regression tree, and the random forest model for medical data classification.

Generally, the huge amounts of data collected in healthcare practice are too complex and voluminous for processing and analysis by traditional methods. Medical diagnosis and prognosis are decision making problems that commonly involve complexity and uncertainty. The use of fuzzy set theory thus has been advocated for medical diagnosis (Alvarez-Alvarez, Trivino, & Cordón, 2012; Mehrnejad & Shekofteh, 2012; Sanz et al., 2014). To deal with uncertain and high-dimensional medical data, this paper proposes a method using fuzzy SAM, genetic algorithm (GA) and wavelet features for healthcare data classification. To our best knowledge, it is the first application of fuzzy SAM method for medical diagnosis and also the first combination of wavelet features and fuzzy SAM in a classification system. Through this study, we examine and compare performance of fuzzy SAM with 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. The next section presents the background of fuzzy SAM and its combination with GA to formulate an integrated model called GSAM. The proposed approach that combines wavelet transformation (WT) and GSAM is described in Section 3. Other machine learning methods are briefly presented in Section 4 for comparisons. Section 5 is devoted for experimental results and discussions, which are followed by concluding remarks and future work in Section 6.

2. Fuzzy SAM with genetic algorithm

2.1. Fuzzy SAM

The fuzzy SAM system F: $R^n \rightarrow R^p$ stores m if–then rules and can uniformly approximate continuous and bounded measurable functions in the compact domain (Kosko, 1994, 1996). This approximation theorem allows any choice of if–part fuzzy sets $A_j \subset R^n$. It also allows any choice of the then–part fuzzy sets $B_j \subset R^p$ because the system uses only the centroid c_j and volume V_j of B_j to compute the output $F(x)$ from the vector input $x \in R^n$.

$$F(x) = Centroid\left(\sum_{j=1}^{m} w_j a_j(x) B_j\right) = \frac{\sum_{j=1}^{m} w_j a_j(x) V_j c_j}{\sum_{j=1}^{m} w_j a_j(x) V_j} = \sum_{j=1}^{m} p_j(x) c_j \quad (1)$$

The fuzzy system covers the graph of an approximand f with m fuzzy rule patches of the form $A_j \times B_j \subset R^n \times R^p$ or of the word form "*If* $X = A_j$ *then* $Y = B_j$". If-part set $A_j \subset R^n$ has joint set function a_j: $R^n \rightarrow [0,1]$ that factors: $a_j(x) = a_j^1(x_1)\ldots a_j^n(x_n)$. Then–part fuzzy set $B_j \subset R^p$ has set function b_j: $R^p \rightarrow [0,1]$ and volume (or area) V_j and centroid c_j. The convex weights:

$$p_j(x) = \frac{w_j a_j(x) V_j}{\sum_{k=1}^{m} w_j a_k(x) V_k} \quad (2)$$

give the SAM output $F(x)$ as a convex sum of then–part set centroids. We can initially ignore the rule weights w_j if we put $w_1 = \cdots = w_n > 0$. In this study, V_j is initialized randomly and then tuned by the supervised learning afterwards.

Fig. 1 shows the parallel structure of the additive systems and its state-space graph cover. The graph cover leads to an exponential rule explosion. A fuzzy system needs on the order of k^{n+p-1} rules to approximate a function f: $R^n \rightarrow R^p$ in a compact domain.

Learning is an important process of SAM with the aim of constructing a knowledge base that is a structure of if–then fuzzy rules. The SAM learning process conventionally includes two basic steps: (i) unsupervised learning for constructing if–then fuzzy rules and (ii) supervised learning for tuning rule parameters (Dickerson & Kosko, 1996).

The supervised learning often starts from a randomly initialized set of parameters and ends when it meets the determined stopping criteria. As training process costs much time and is often trapped in local minima, the initialization of parameters is thus a nontrivial issue. The unsupervised learning process, which is often completed by a clustering method, helps to initialize parameters of fuzzy rules more insightfully (Fig. 2).

In this paper, to enhance the efficiency of the SAM learning process, we propose the use of an evolutionary learning process, i.e. GA, to optimize the number of fuzzy rules before the supervised learning is performed. The evolutionary learning component is designed to alleviate the computational cost of the succeeding supervised learning. The entire integration between GA and fuzzy SAM is henceforth denoted GSAM as illustrated in Fig. 3.

2.2. Unsupervised learning by the AVQ clustering

We utilize the adaptive vector quantization (AVQ) clustering method (Kosko, 1991) to identify the centers of membership functions (MFs) in the antecedent part and the centroids in the consequent part. The well-separated distribution of the resulting clusters from the AVQ method is useful in identifying the allocation of fuzzy rules in the fuzzy SAM. The AVQ clustering method is briefly summarized below.

The clustering process uses K quantization vectors to search for fuzzy classes in the learning dataset that cover the unknown function f in the space XY. The K quantization vectors can be initialized randomly. For each data pattern at time t: $z(t) = [x(t)|y(t)]$, the algorithm searches for a fuzzy class that can contain $z(t)$ based on the closest q_j (competitive learning), which is selected based on the following conditions:

$$\|z(t) - q_j\| = \min_{i=\overline{1,k}} \|z(t) - q_i\| \quad (3)$$

where $\|z\|^2 = z_1^2 + z_2^2 + \cdots + z_n^2$.

Then q_j is updated to be closer to $z(t)$:

$$q_j(t+1) = q_j(t) + \mu_t[z(t) - q_j(t)] \quad (4)$$

Based on the competitive learning, q_j vectors are updated closer to the fuzzy classes covering the graph of the unknown function f. At the end of training, K quantization vectors q_j obtained reflect the distribution of fuzzy classes of the training data.

Denote the learning dataset as $\{z_t\}$, $t = 1, \ldots, N$, and the local conditional covariance matrix Q_j in pattern class D_j as $Q_j(t) = E[(z - \bar{z})(z - \bar{z})^T | D_j]$, the algorithm is presented as follows:

```
Step 1. Initialize $q_j$ randomly, $Q_j = 0, j = 1, \ldots, K$
Step 2. Consider the learning sample at time $t$: $z(t) = [x(t)|y(t)]$
Step 3. Search for $q_j$ at time $t$ based on Eq. (3).
Step 4. Update quantization vectors $q$ and matrix $Q$:
    If $i = j$: $q_i(t+1) = q_i(t) + \mu_t[z(t) - q_i(t)]$
```

$$Q_i(t+1) = Q_i(t) + \mu_t[(z(t) - q_i(t))(z(t) - q_i(t))^T - Q_i(t)] \quad (5)$$

```
    If $i \neq j$: $q_i(t+1) = q_i(t)$
```

$$Q_i(t+1) = Q_i(t) \quad (6)$$

```
Step 5. If $t < N$ then $t = t + 1$, back to step 2.
Step 6. End.
```