Expert Systems with Applications 40 (2013) 5402–5412

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Feature generation using genetic programming with comparative partner selection for diabetes classification

Muhammad Waqar Aslam [a], Zhechen Zhu [b], Asoke Kumar Nandi [b,c,*]

[a] Department of Electrical Engineering & Electronics, The University of Liverpool, Brownlow Hill, Liverpool L69 3GJ, UK
[b] Department of Electronic & Computer Engineering, Brunel University, Uxbridge, Middlesex UB8 3PH, UK
[c] Department of Mathematical Information Technology, University of Jyväskylä, P.O. Box 35, Jyväskylä FI-40014, Finland

ARTICLE INFO

Keywords:
Pima Indian diabetes
Genetic programming
Comparative partner selection

ABSTRACT

The ultimate aim of this research is to facilitate the diagnosis of diabetes, a rapidly increasing disease in the world. In this research a genetic programming (GP) based method has been used for diabetes classification. GP has been used to generate new features by making combinations of the existing diabetes features, without prior knowledge of the probability distribution. The proposed method has three stages: features selection is performed at the first stage using t-test, Kolmogorov–Smirnov test, Kullback–Leibler divergence test, F-score selection, and GP. The results of feature selection methods are used to prepare an ordered list of original features where features are arranged in decreasing order of importance. Different subsets of original features are prepared by adding features one by one in each subset using sequential forward selection method according to the ordered list. At the second stage, GP is used to generate new features from each subset of original diabetes features, by making non-linear combinations of the original features. A variation of GP called GP with comparative partner selection (GP-CPS), utilising the strengths and the weaknesses of GP generated features, has been used at the second stage. The performance of GP generated features for classification is tested using the k-nearest neighbor and support vector machine classifiers at the last stage. The results and their comparisons with other methods demonstrate that the proposed method exhibits superior performance over other recent methods.

1. Introduction

Diabetes is a condition in which the blood glucose level is higher than normal. Food containing specific carbohydrates is turned into glucose which is passed to the bloodstream where it is used by cells for growth and energy. Insulin is a hormone produced by pancreas for moving glucose from blood to cells. In diabetes either pancreas produces little insulin or the cells do not use the produced insulin properly. This results in an increase of glucose in the blood, which passes out of the body through urine and ultimately results in loss of fuel (glucose) for the body, even though it is present in large amount in the blood. Diabetes leads to many other diseases including heart disease, high blood pressure, nerve damage, numbness in hands or feet, diabetic retinopathy, and diabetic nephropathy. There are two main types of diabetes, type 1 and type 2. In type 1 the beta cells in pancreas, responsible for producing insulin are destroyed and as a result pancreas produces little or no insulin. Type 1 mostly occurs in children or young adults but can affect at any age. People suffering from this type have to take insulin injections regularly to stay alive. Type 2 is the most common type of diabetes, covering at least 90% of all the diabetes cases. In this type body becomes resistant to insulin and does not effectively use the insulin being produced. This type mostly occurs in the class of people who are more than forty years old but can also be found in younger classes. It can be treated by following a healthy diet plan, doing exercise regularly and/or taking tablets. In some extreme cases, insulin injections may also be required. However, diabetes still contributes to heart disease even if it is under control.

Diabetes has been increasing at a rapid rate and if it continues to increase at the current rate, there would be demand for a large number of physicians in future. In order to cope with this problem, the use of classifier systems in medical diagnosis has increased in recent times. The aim of this study is to make a system which can automatically figure out if a patient has diabetes, without the need of a physician. If the decisions made by physicians on previous patients having similar conditions are saved in a list along with patient conditions, a classifier system could be designed which makes use of the conditions and classifies that list according to the decisions made by physicians. No doubt, data taken from the patient and expert's opinion about the data are the most important

[*] Corresponding author at: Brunel University, Department of Electronic & Computer Engineering, Uxbridge, Middlesex UB8 3PH, United Kingdom. Tel: +44 1895 266119; fax: +44 1895 269782.

E-mail addresses: m.w.aslam@liverpool.ac.uk (M.W. Aslam), zhechen.zhu@brunel.ac.uk (Z. Zhu), asoke.nandi@brunel.ac.uk (A.K. Nandi).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.04.003

in diagnosis but a classifier system can also help physicians a great deal.

Pima Indian diabetes dataset (Frank & Asuncion, 2010) from UCI Repository of machine learning databases has been used in this study. In the past numerous methods have been used for classification of this diabetes dataset. Polat, Gunes, and Arslan (2008) proposed a two stage cascaded learning system using generalized discriminant analysis (GDA) and least square support vector machine (LS-SVM). They used GDA at the first stage to discriminate between healthy and patient data, and used LS-SVM at the second stage for classification. In another research (Polat & Gunes, 2007) used principal component analysis (PCA) for dimensionality reduction of diabetes data. Adaptive Neuro-fuzzy inference system (ANFIS) was used for the classification of this reduced dimensionality dataset. Temurtas, Yumusak, and Feyzullah (2009) used multilayer neural network (MLNN) trained by Levenberg–Marquardt (LM) method and a probabilistic neural network (PNN) for diabetes classification. Gadaras and Mikhailov (2009) used fuzzy rules based method for diabetes classification. Balakrishnan, Narayanaswamy, and Paramasivam (2011) used F-score selection and k-means clustering for the selection of optimal features and this selected feature subset was tested using SVM classifier. Kala, Vazirani, Khanwalkar, and Bhattacharya (2010) used radial basis function network (RBFN), while (Lekkas & Mikhailov, 2010) used fuzzy rules for classification of the diabetes data.

A genetic programming (GP) based method has been used for diabetes classification in this research, inspired by Aslam and Nandi (2010). The use of GP in classification problems is not new, it has been used quite a lot in the past for classification problems and the details can be found in a survey presented by Espejo, Ventura, and Herrera (2010). Zhang, Jack, and Nandi (2005), as well as Zhang and Nandi (2007) used GP for feature generation and K-nearest neighbor (KNN) for classification purpose. Guo, Jack, and Nandi (2005) used GP with Fisher criterion for the classification of roller bearing data. Eggermont, Eiben, and van Hemert (1999) presented a comparative analysis on different variations of GP for binary classification problems. Kishore, Patnaik, Mani, and Agrawal (2000) and Muni, Pal, and Das (2004) used GP for multi-class classification by dividing any *n*-class problem into *n* 2-class problems. Zhang, Ciesielski, and Andreae (2003) used multiple thresholds scheme for multi-class classification. Day and Nandi (2008) presented the idea of comparative partner selection (CPS) for exploring strengths and weaknesses of GP individuals, and the same idea has been used in this research for most of the experiments.

In this study selection of original diabetes features is performed at the first stage employing various methods. Different subsets of selected features are prepared using sequential forward selection method according to features' importance. At the next stage, new features are generated from each subset of selected features using GP. At the final stage, the new GP generated features are tested using KNN and SVM classifiers.

The paper is organized as follows: the proposed method is presented in Section 2. The diabetes dataset and selection of features is discussed in Section 3. The GP algorithm and the CPS variation introduced in GP is presented in Section 4. Experiments, results and comparison with other methods is presented in Section 5, while the conclusion is drawn in Section 6.

2. The proposed method

The proposed method can be divided into three stages. At the first stage various feature selection methods including Student's t-test, Kolmogorov–Smirnov test, Kullback–Leibler divergence test and F-score selection are used to evaluate the effectiveness of diabetes features for classification purpose. In addition, the effectiveness of GP as a feature selector is also investigated. Each method gives an ordering of features based on features' importance. The diabetes features are arranged in a decreasing order of importance using the results of all feature selection methods. Sequential forward selection method is used to prepare different subsets of the original diabetes features. Features are added one by one in each subsequent subset according to the order given by feature selection methods, e.g the first subset will contain the most important feature, the second subset will contain top two features and so on. The last subset will contain all the features. At the second stage GP is evolved (trained) to generate new features by making non-linear combinations of the existing features, for each subset of the original diabetes features. A variation of GP, called GP with CPS, is used at this stage. CPS allows GP to explore the search space more efficiently by getting insight into the strengths and weaknesses of GP generated features. At the last stage, the new features generated by GP (CPS) during training are tested using KNN and SVM classifiers. Different stages of the proposed method are shown in Table 1.

3. Selection of features

3.1. Pima Indian Diabetes dataset

This section explains the diabetes dataset used for all the experiments. The National Institute of Diabetes and Digestive and Kidney Diseases originally owned this data, and it was received by UC-Irvine Machine learning Repository in 1990 (Frank & Asuncion, 2010). The patients were females of Pima Indian heritage and at least 21 years old. There were total 768 cases, out of which 500 (65.1%) cases had no diabetes (class 0) and 268 cases (34.9%) had diabetes (class 1). Each of these cases had eight attributes and details of these attributes are given in Table 2. Since the mean of different attributes are quite far from each other, the data is normalized according to the following equation to make it zero mean with unit standard deviation.

$$X_{new,i} = \frac{(X_i - \overline{X})}{\sigma_X} \tag{1}$$

where $X_{new,i}$ is the new attribute value after normalization, X_i is the original attribute value, $\overline{X}$ is the mean of the attribute X and σ_X is the standard deviation of the attribute X.

3.2. Feature selection methods

Feature selection is an important part of a pattern recognition system. In any pattern recognition problem, all the features may not be always beneficial for classification and some of the features may not contribute meaningful information. These are irrelevant or redundant features which simply add noise to the dataset distribution and affect the classification accuracy. A high dimensional dataset also presents processing challenges as it has high computational cost. In order to minimize these effects, some form of dimension reduction is required as a preprocessing step.

Dash and Liu (1997) presented a survey on different feature selection methods. A feature selection method can evaluate the effectiveness of the features using one of the following criteria: distance, information, dependence, consistency and classifier error rate (Dash & Liu, 1997). Some of the feature selection methods employing these criteria are discussed in Nandi, Nandi, Rangayyan, and Scutt (2006). There are two main types of feature selection methods, wrapper type and filter type (Muni, Pal, & Das, 2006). In wrapper type methods the features selection is performed by the same classifier later used for classification. This method is computationally expensive and can result into over-fitting the model because it is biased to its underlying classification engine. In filter type methods feature selection is performed at the first stage,