Information Sciences 292 (2015) 1–14

Contents lists available at ScienceDirect

Information Sciences

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

An improved early detection method of type-2 diabetes mellitus using multiple classifier system

Jia Zhu [a,*], Qing Xie [b], Kai Zheng [c]

[a] School of Computer Science, South China Normal University, China
[b] Division of CEMSE, King Abdullah University of Science and Technology, Saudi Arabia
[c] School of Information Technology and Electrical Engineering, University of Queensland, Australia

ARTICLE INFO

Article history:
Received 5 March 2014
Received in revised form 9 August 2014
Accepted 29 August 2014
Available online 7 September 2014

Keywords:
Multiple classifier system
T2DM
Classification

ABSTRACT

The specific causes of complex diseases such as Type-2 Diabetes Mellitus (T2DM) have not yet been identified. Nevertheless, many medical science researchers believe that complex diseases are caused by a combination of genetic, environmental, and lifestyle factors. Detection of such diseases becomes an issue because it is not free from false presumptions and is accompanied by unpredictable effects. Given the greatly increased amount of data gathered in medical databases, data mining has been used widely in recent years to detect and improve the diagnosis of complex diseases. However, past research showed that no single classifier can be considered optimal for all problems. Therefore, in this paper, we focus on employing multiple classifier systems to improve the accuracy of detection for complex diseases, such as T2DM. We proposed a dynamic weighted voting scheme called multiple factors weighted combination for classifiers' decision combination. This method considers not only the local and global accuracy but also the diversity among classifiers and localized generalization error of each classifier. We evaluated our method on two real T2DM data sets and other medical data sets. The favorable results indicated that our proposed method significantly outperforms individual classifiers and other fusion methods.

1. Introduction

The specific causes of complex diseases such as Type-2 Diabetes Mellitus (T2DM) have not yet been identified; nevertheless, many medical science researchers suppose that complex diseases are caused by a combination of genetic, environmental, and lifestyle factors [18]. Early detection of such diseases can prevent and treat complex diseases when they do not have obvious clinical symptoms. Considering the greatly increased amount of data gathered in medical databases and the availability of historical data on complex diseases, such as patients' blood glucose, traditional manual analysis has become inadequate and naturally leads to the application of data mining techniques to discover interesting patterns so that early detection and successful recommendation for diagnosis becomes possible [15].

Recently, data mining techniques have been widely used to detect and improve the diagnosis of complex diseases [5,17,36], but past research showed that no single classifier can be considered optimal for all problems even by applying certain validation techniques, such as cross-validation [38]. Therefore, in this paper we focus on employing multiple classifier systems (MCS) to improve the accuracy of detection for T2DM.

* Corresponding author.
E-mail address: jzhu@m.scnu.edu.cn (J. Zhu).

http://dx.doi.org/10.1016/j.ins.2014.08.056
0020-0255/

MCS is a set of individual classifiers whose decisions are combined according to certain rules to produce the final output. MCS has many advantages, and studies show that the combination of homogeneous classifiers using heterogeneous features can improve the final result [9,24,29]. Moreover MCS can exploit the local behavior of each individual classifier to improve the overall classification performance [12].

However, MCS may perform worse than individual classifiers without proper design in certain circumstances. Many factors should be considered when designing an MCS, including MCS architecture, creation of ensembles, and classifier selection [30]. In this article, we focus only on decision combination or the so-called fusion strategy, which is pivotal for the performance of MCS [12].

Two general weighting approaches exist for combining base classifiers: static and dynamic. In static weighting approach, the weighting is determined completely in the training phase, and cannot be changed in the classification phase. The problem of static weighting is that the weight is assigned according to the training accuracy only, which makes the assumption that all classifiers have the same performance in the entire input space [8]. However, a base classifier may perform under average overall but may have good performance in a certain region of the input space. Fig. 1 is a typical example. Classifiers 1 and 2 are good at identifying rectangular objects but not circular objects, and classifier 3 is the opposite. If we use static weighting with majority voting method [22], the system classification accuracy will be only 50% as that of the accuracy of classifiers 1 and 2 if established on the provided training data only.

To overcome this problem, dynamic weighting is a better approach because the weights assigned to the output of each individual classifiers can change for each input vector in the testing phase [8]. In this article, we propose a dynamic weighted voting scheme called multiple factors weighted combination (MFWC) for multiple classifier decision combination. Compared with other dynamic weighting methods, our method not only considers the local accuracy factor for each classifier and uses a validation set to estimate the classification accuracy at the global level but also concerns the relationship between training and testing samples with generalization error because the generalization error of a classifier is a key function to measure the performance of a classifier generalized to unseen samples.

To prove that our method can work on medical data, we evaluated our method on real medical data sets including two T2DM data sets because diabetes is a typical complex disease and a major global health problem that affects hundreds of millions of people around the world and takes 11.6% of the total health care expenditure in the world in 2010 [36]. Our main contributions include the following:

1. We propose a dynamic weighted voting scheme called MFWC for classifiers' decision combination that can overcome the problem of static weighting approaches.
2. Compared with other dynamic weighting methods, we improve the calculation by considering not only the local and global accuracy but also the diversity among classifiers and localized generalization error of each classifier by using a weighted voting combination method so that performance can be guaranteed on unseen samples.
3. We perform extensive experiments on various medical corpora to demonstrate the effectiveness of our method.

The rest of this paper is organized as follows. In Section 2, we discuss related works in multiple classifier systems. In Section 3, we describe the technical details of our detection method for T2DM, particularly on the proposed dynamic weighted voting scheme. In Section 4, we present our experiments, evaluation metrics, and results. We conclude this study in Section 5.

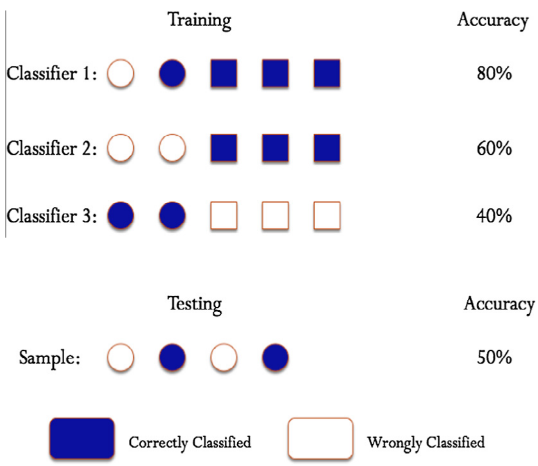

Fig. 1. An example of static weighting.