Expert Systems with Applications 42 (2015) 7601–7609

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Predicting hypertension without measurement: A non-invasive, questionnaire-based approach

Aiguo Wang [a], Ning An [a,*], Guilin Chen [b], Lian Li [a], Gil Alterovitz [c,d,e]

[a] Hefei University of Technology, School of Computer and Information, Hefei, China
[b] Chuzhu University, School of Computer and Information Engineering, Chuzhou, China
[c] Harvard University, Harvard Medical School, Center for Biomedical Informatics, Boston, USA
[d] Children's Hospital Informatics Program, Boston Children's Hospital, Harvard Medical School, Boston, USA
[e] Massachussetts Institute of Technology, Department of Electrical Engineering and Computer Science, Cambridge, USA

ARTICLE INFO

Article history:
Available online 11 June 2015

Keywords:
Hypertension prediction
Non-invasive
Questionnaire
Behavior Risk Factor Surveillance System
Artificial neural networks

ABSTRACT

Early detection of hypertension contributes to the prevention and reduction of the onset of cardiovascular diseases. Since lifestyle choices are linked to the occurrence and development of hypertension, determining hypertension risk factors and further establishing a predictive model with these factors will facilitate the early prevention and effective management of hypertension and improve individual health conditions. This study attempts to construct a prediction model based on the hybrid use of logistic regression and artificial neural networks (ANNs) for hypertension detection in a non-invasive, questionnaire-based way. First, the binary logistic regression model was used to select risk factors significant to hypertension. Second, after detailing the selection of ANNs architecture and the setting of relevant parameters, we constructed a multi-layer perception neural network model with back propagation learning algorithms to predict hypertension. Then, to mitigate the biased prediction results caused by a potentially unbalanced training set, we proposed an effective under-sampling technique and adopted it to balance the dataset prior to the training of the predictive model. To evaluate the performance of the proposed approach, we conducted extensive experiments on the questionnaires collected from Behavior Risk Factor Surveillance System. Experimental results show that ANN-based prediction model obtains over 72.0% accuracy and an area under the receiver-operator curve of 0.77 and achieves good stability in comparison with the logistic regression-based model. Further, the proposed approach obtains balanced prediction performance with the under-sampling technique. The results demonstrate the practicability of hypertension prediction with simple demographic data rather than with clinical tests and genomic data and of developing a hypertension surveillance system for a large scale of population in a non-invasive and economical way. Also, we actually provide a general framework for the simultaneous identification of risk factors and prediction of other chronic diseases.

1. Introduction

Hypertension is a long-lasting chronic health condition and affects a wide range of the population, particularly for adults over the age of fifty-five. Even worse, it is becoming prevalent among adolescents in both developing and developed countries (Ture, Kurt, Turhan Kurum, & Ozdamar, 2005; Midha et al., 2014). Hypertension is also a major risk factor for the occurrence and development of many cardiovascular diseases, such as stroke, heart failure and chronic kidney disease, and the poor management and treatment of hypertension leads to the increase in morbidity and mortality rates (Hsu et al., 2011; Jeppesen, Hein, Suadicani, & Gyntelberg, 2000; Vasan et al., 2001; Wong et al., 2003). Besides the fact that prevention and management of hypertension consumes a wealth of medical resources and healthcare services, it deteriorates the imbalanced distribution of medical resources and definitely puts on the society a considerable financial burden.

The main difficulty associated with hypertension prevention and management is the lack of clear clinical effects in the early stage of hypertension. As a consequence, individuals may easily disregard the occurrence of hypertension and develop potential serious complications (Vasan et al., 2001). Though hypertension

* Corresponding author at: Hefei Tunxi Road 193, Hefei University of Technology, Hefei 230009, China. Tel.: +86 180 1995 6086; fax: +86 551 6290 4642.
E-mail addresses: wangaiguo2546@163.com (A. Wang), ning.g.an@acm.org (N. An), glchen@chzu.edu.cn (G. Chen), llian@hfut.edu.cn (L. Li), gil_alterovitz@hms.harvard.edu (G. Alterovitz).

http://dx.doi.org/10.1016/j.eswa.2015.06.012
0957-4174/

is among the most common and costly health problems, it is also among the most preventable and can be effectively controlled through reasonable measures due to the fact that lifestyle choices are linked to the occurrence and development of hypertension (Chang, Wang, & Jiang, 2011; Hsu et al., 2011; Krawczyk & Wozniak, 2011; Sumathi & Santhakumaran, 2011; Wozniak, 2006). Therefore, investigating risk factors and identifying hypertension plays a crucial role in the effective prevention and reduction of the onset of cardiovascular diseases as well as better management and intervention of individual health conditions (Hsu et al., 2011). On the other hand, the investigation of hypertension risk factors is a crucial issue for preventive medicine and particularly drawing interests from public health researchers with the aim to bring down the onset of hypertension through early warning and prevention. In comparison with clinical test data, genomic data and anthropometric body surface scanning data (Chiu et al., 2007), lifestyle behavior information provides an alternative way for hypertension prediction, and they are easily collected and more meaningful in the prevention and management of hypertension. Furthermore, lifestyle risk factors could be indicators to remind or warn individuals to avoid or circumvent unhealthy behaviors in order to effectively prevent and better manage hypertension, and the prediction model can be used for locating those individuals who may be at high risk of hypertension and for the large-scale hypertension surveillance without the complex and expensive measurements.

Traditional approaches usually statistical techniques to determine the relationship between hypertension and the risk factors. Among these, artificial neural networks (ANNs) are data-driven methods and have the ability to adjust themselves to the data without positing any explicit specification of distribution form for the underlying model. This differs from traditional statistical procedures that are established on Bayesian statistical theory. As a nonlinear mapping model, ANNs are flexible and effective in modeling complex relationships between inputs and outputs and widely used for the medical diagnostics (Ziada et al., 2001). However, one of the main difficulties in constructing neural networks is the model selection problem. More precisely, one needs to select a suitable ANNs architecture and set its corresponding parameters due to the fact that ANNs are quite sensitive to these factors and inappropriate model selection can degrade their generalization ability.

With the aim to enable early identification of hypertension and risk factors and develop a practical screening tool, in this study, we proposed a questionnaire-based hypertension prediction approach that integrated logistic regression analysis and artificial neural networks with the aim of determining risk factors and predicting hypertension. After collecting and cleaning a publicly available dataset from Behavior Risk Factor Surveillance System (BRFSS) of Centers of Control and Prevention (CDC), we first applied the binary logistic regression model to select risk factors significantly relevant to hypertension and constructed the logistic regression-based prediction model. Then, we trained a multi-layer perception (MLP) neural network with back propagation algorithms using the selected risk factors as inputs to predict whether an individual suffers from hypertension. In the construction and training of ANNs, we detailed the selection of ANNs architecture and proposed to employ three rule-of-thumbs to narrow down the search space of feasible solutions towards a tradeoff between efficiency and accuracy. Additionally, considering that class imbalance problems are common in medical datasets and that the skewed class distribution makes many classification methods less effective and jeopardizes the accuracy of the minority class (Wang & Yao, 2012), we proposed an effective under-sampling technique to adjust the size of training sets prior to the training of ANNs.

The remainder of this paper is organized as follows. Section 2 reviews previous related research work and techniques. Experimental dataset, logistic regression analysis and artificial neural network models are illustrated in Section 3. In the experimental design and result analysis section, we detail the selection of neural network architectures and the setting of corresponding parameters, and describe an experiment to demonstrate their effectiveness for hypertension prediction in comparison with that of logistic regression based prediction model. The last section concludes our work with a brief summary and presents possible directions for the future studies.

2. Related work

A large number of researchers and medical experts have conducted considerable work in investigating hypertension risk factors and indicators and constructing effective prediction models with these factors. There are a variety of factors that can be used to predict hypertension, mainly including demographics, anthropometry body surface scanning data, clinical test data and even molecular-level data (e.g. genomic and proteomic data). To figure out the risk factors, Lee and Entzminger (2006) conducted a cross-sectional study in a Thai population of 1398 patients (382 men and 1016 women), and performed multiple linear regression to determine the relevance of several risk factors for hypertension. They found that old age, body mass index and low education attainment are significant risk factors. Akdag et al. (2006) applied the classification tree method to determine risk factors for hypertension among 1761 adults at the outpatient clinic in western Turkey between January 2002 and July 2004. They studied the effects of fourteen risk factors on hypertension, and their results revealed that body mass index, waist-to-hip ratio, sex, serum triglycerides, serum total cholesterol, hypertension in first-degree relatives, and saturated fat consumption are main risk factors. Accordingly, various machine learning and statistical analysis techniques with different metrics are utilized to find a mapping function between the factors and hypertension. Ture et al. (2005) compared a comparative study to evaluate the performance of nine commonly used classification methods for hypertension prediction among 694 subsets (452 hypertension patients and 242 controls). Their experimental dataset consisted of demographics, lifestyle information and clinical test results. Experimental results revealed that multi-layer perception (MLP) neural network and Radial Basis Function (RBF) neural network outperformed the other three decision tree and four statistical algorithms. Blinowska, Chatellier, Bernier, and Lavril (1991) proposed to apply Bayesian statistical methods that incorporated both prior knowledge and possible costs of wrong decisions for hypertension prediction using demographics and clinical test data, and the proposed method achieved satisfactory accuracy. However, since Bayesian method is built on statistical theory, difficulties in collecting a sufficient number of experimental cases and ensuring the integrity of each case hinder its wide applications in actual use (Blinowska, Chattellier, Wojtasik, & Bernier, 1993; Blinowska et al., 1991). Chang et al. (2011) proposed to use several data mining classifier techniques to determine the risk factors of hypertension in a vote-based scheme, and then build a predictive model using multivariate adaptive regression splines. Besides using clinical test data, researchers also explore the possibility of hypertension prediction using other types of data. For example, Hsu et al. (2011) focused on determining the relationship between hypertension and three-dimensional anthropometric scanning data (e.g. the circumferences of waist, wrist and gluteal), and they proposed to hybridize case-based reasoning and genetic algorithms for hypertension detection. Experimental results revealed the relationship between anthropometric data and hypertension and demonstrated the effectiveness of case matching techniques. In addition, to investigate the mechanism of hypertension at the