Neurocomputing 276 (2018) 2–22

Contents lists available at ScienceDirect

Neurocomputing

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

Computational intelligence approaches for classification of medical data: State-of-the-art, future challenges and research directions

Ali Kalantari [a], Amirrudin Kamsin [a], Shahaboddin Shamshirband [b,c,*], Abdullah Gani [a], Hamid Alinejad-Rokny [d], Anthony T. Chronopoulos [e,f]

[a] Faculty of Computer Science & Information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[b] Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[c] Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
[d] UNSW Australia, Australia
[e] Department of Computer Science, University of Texas, San Antonio, TX 78249, USA
[f] Department of Computer Science (Visiting Faculty), University of Patras, Greece

ARTICLE INFO

Article history:
Received 23 August 2016
Revised 29 December 2016
Accepted 4 January 2017
Available online 19 September 2017

Keywords:
Computational intelligence
Medical application
Big data
Detection
Ensemble algorithm

ABSTRACT

The explosive growth of data in volume, velocity and diversity that are produced by medical applications has contributed to abundance of big data. Current solutions for efficient data storage and management cannot fulfill the needs of heterogeneous data. Therefore, by applying computational intelligence (CI) approaches in medical data helps get better management, faster performance and higher level of accuracy in detection. This paper aims to investigate the state-of-the-art of computational intelligence approaches in medical data and to categorize the existing CI techniques, used in medical fields, as single and hybrid. In addition, the techniques and methodologies, their limitations and performances are presented in this study. The limitations are addressed as challenges to obtain a set of requirements for Computational Intelligence Medical Data (CIMD) in establishing an efficient CIMD architectural design. The results show that on the one hand Support Vector Machine (SVM) and Artificial Immune Recognition System (AIRS) as a single based computational intelligence approach were the best methods in medical applications. On the other hand, the hybridization of SVM with other methods such as SVM-Genetic Algorithm (SVM-GA), SVM-Artificial Immune System (SVM-AIS), SVM-AIRS and fuzzy support vector machine (FSVM) had great performances achieving better results in terms of accuracy, sensitivity and specificity.

1. Introduction

The era of big data has begun, due to large-volume, complex and growing number of data sets which are produced by various sources such as Internet of Things (IoT), government records, health records, multimedia, phone logs, social media and other digital traces [1–3]. Moreover, big data are being used to transform medical practice, inform business decision making, and modernize public policy [4]. Accordingly, the produced number of complex data from medical and healthcare increase rapidly with numerous essential information. Therefore, big data has infinite potential in efficiently storing, processing, querying, and analyzing medical data [5]. For instance, the Ayasdi organization provides information to the Mount Sinai Medical Center in the U.S. about the genetic sequencing of *Escherichia coli* (*E. coli*) bacteria. This data is used to investigate the resistance of antibiotics to bacterial strains. It utilizes topological data analysis which is a contemporary research methodology that can comprehend data characteristics [6]. In addition, there are various areas in healthcare industry including medical imaging [7–9], patient genomics [10,11], electronic health records (EHRs) [12,13], unstructured text data [14,15] and device, log and sensor data [16,17] that can have potential benefits from big data techniques and infrastructure [18].

However, there have been issues like security, privacy, the effectiveness of analysis and data quality which are very important in medical data and applications. Therefore, medical information shows valuable intellectual property and their usage is highly guarded which means the information management should not only be constrained on practices and established laws, but also by subjects expectation of privacy [19]. For example, a framework was proposed by Bertino et al. [20] to achieve the privacy and copyright protection for outsourced medical data by combining the techniques of binning and digital watermarking. Moreover, global disease network examination, using biological databases and pa-

* Corresponding author at: Department for Management of Science and Technology Development, Ton Duc Thang University, Ho Chi Minh City, Vietnam.
E-mail address: shahaboddin.shamshirband@tdt.edu.vn (S. Shamshirband).

https://doi.org/10.1016/j.neucom.2017.01.126
0925-2312/

tient data from electronic medical records (EMR), has emerged as a powerful modality for understanding the complexity of disease relationships [21,22]. In this study [23] the authors proposed the combination of different powerful approaches to compare disease network structure and connectivity between three different populations such as Hispanic/Latino, Caucasian and African American.

Additionally, there are other challenges of big data in healthcare that are addressed by the procurement of information from complex heterogonous patient sources. These challenges include obtaining clinical notes and comprehending them in the right context, organizing medical imaging data, perceiving data concerning biomarkers and understanding large amounts genomic data which can be useful in clinical settings when the patient is assessed. Also, data about the patient's psychological, behavioral, and social patterns can be accessed by various sensors [24]. A recent survey by Fang et al. [25] reviewed the current challenges, techniques and future directions for computational health informatics in big data. They summarized the challenges into four dimensions (Vs) of big data as volume, velocity, veracity and variety. In another recent investigation [5] the authors considered 6Vs (volume, velocity, veracity, variety, variability and value) possible challenges with complexity in big data.

However, the recent technologies like Artificial Intelligence (AI) can assist in solving different complex issues. AI in its widest sense would demonstrate the capability of a machine to perform tasks similar to the human thought. Thus, AI has been used for computer systems with ability of task execution which is more complex than simple programing [26].

Artificial intelligence is also increasing in three V's (volumes, velocities and variety of data) similar to big data. AI enables learning, delegation of difficult pattern recognition and additional responsibilities to computer-based methods under circumstances of large volumes of data. Furthermore, it helps velocity of data, by assisting fast computer-based decisions that make possible different choices. Besides, the variety issue is not solved only by parallelizing and distributing the problems. Variety is mitigated by capturing, structuring, and understanding unstructured data using AI and different analytics [27]. Therefore, AI and big data analytics could reshape the healthcare system with greater performance, provide healthcare insights, and improve the overall processes in two core elements for improving the productivity, efficiency and the quality of care [28]. With the intention of mitigating the performance of AI, the computational intelligence (CI) techniques adapt to medical data. CI typically refers to the ability of computer to learn a particular task from experimental observation or data, which facilitate the intelligent behavior in complex problems and changing environments [29].

CI techniques are classified based on single and hybrid methods, where single methods refer to those studies which use only one of the machine learning techniques (i.e. genetic algorithm (GA), particle swarm optimization (PSO), artificial immune system (AIS) and artificial neural network (ANN)) as a main method and the other classification refers to those studies that used hybridization of each two (or more than two) methods like Neuro-Fuzzy NF) and Fuzzy Support Vector Machine (FSVM). For instance, Latifoğlu et al. [30] used only artificial immune recognition system (AIRS) as the main technique for atherosclerosis diagnosis from carotid artery Doppler signals, which is considered as a single method classification. In another study by Gu et al. [31], for hybrid classification they used FSVM technique in medical datasets classification.

The aim of this study is to investigate the state-of-the-art of computational intelligence (CI) approaches in medical data and to classify the CI techniques (single and hybrid) in terms of chronological design. Also, the study's aim is to analyze the CI methods in terms of accuracy, sensitivity and specificity for medical and health informatics fields.

The rest of the paper is organized as follows. Section 2 provides the survey methodology. Section 3 presents the datasets used for CI approached applied to medical data. Section 4 defines the criteria for evaluation. Section 5 presents the state-of-the-art of single and hybrid CI approaches in medical data and presents the comparison based on the evaluation criteria. Lastly, Section 6 draws the overall conclusions.

2. Methodology

In this study, 71 articles related to CI techniques in medical data were reviewed and selected from highly cited publications and credible sources as: Science Direct, IEEE, Springer and Web of Science (WoS). This paper integrates different classification of CI techniques which are used for medical data.

Table 1 provides the list of literature works dealing with single and hybrid CI methods. The list of articles is given as a general overview of single and hybrid methods in terms of characteristics and current challenges of CI in medical data and healthcare. The table contains 2 horizontal sections (single and hybrid methods) and 4 vertical divisions which are single (or hybrid), classifier type, title of paper, and aim.

3. Datasets used for computational intelligence approaches in medical data

To make the evaluation of performance for each approach for detection, diagnosis, prediction, analyzing and classification, it is required that the related datasets define the level of accuracy, sensitivity and specificity. Table 2 indicates the classification of the datasets which were for the most part obtained from University of California at Irvine (UCI) Machine Learning Repository, which is considered as a public dataset.

Different types of medical datasets were used by researchers for coronary heart disease, ovarian, hepatitis, lung and breast cancer. For instance, the hepatitis disease dataset (including 19 attributes, and 155 samples based on two categories: 32 for die cases and 123 for live cases) was used to predict the disease of hepatitis [68]. In addition, medical images are one type of digital information data which are increasing. In these study [108,110] the authors used the public medical database of ImageCLEF 2007 in order to achieve a high classification rate.

4. Criteria used for evaluation

The effectiveness of CI techniques in medical data is evaluated on how capable the single and hybrid methods are in making correct diagnosis, detection, monitoring and prediction in terms of accuracy (i.e. the overall proportion of correct classification) and sensitivity (i.e. the proportion of the positives correctly recognized) and specificity (i.e. proportion of the negatives correctly recognized).

Table 3 indicates the analysis of three key aspects (accuracy, sensitivity and specificity) in medical data using computational intelligence approaches. The significance of the performance is emphasized, this is because medical data represents valuable intellectual property. As an example, in [67] the authors investigated computer aided medical diagnosis systems using AIRS method and they claim that, the proposed approach obtained accuracy of 100% and it was a very promising with respect to other classification application problems. In addition, Lahsasna et al. [75] designed a fuzzy based decision support system for CHD diagnosis that achieved 84.44% for accuracy, 79.2% for sensitivity and 88.7% for specificity.