Expert Systems with Applications 41 (2014) 4434–4463

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Review

Knowledge discovery in medicine: Current issue and future trend

Nura Esfandiari [a], Mohammad Reza Babavalian [a,*], Amir-Masoud Eftekhari Moghadam [a], Vahid Kashani Tabar [b]

[a] Faculty of Computer and Information Technology, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[b] Trauma Research Center, Kashan University of Medical Sciences, Kashan, Iran

ARTICLE INFO

Keywords:
Data mining application
Medical data mining
Medicine
Disease
Data mining algorithms

ABSTRACT

Data mining is a powerful method to extract knowledge from data. Raw data faces various challenges that make traditional method improper for knowledge extraction. Data mining is supposed to be able to handle various data types in all formats. Relevance of this paper is emphasized by the fact that data mining is an object of research in different areas. In this paper, we review previous works in the context of knowledge extraction from medical data. The main idea in this paper is to describe key papers and provide some guidelines to help medical practitioners. Medical data mining is a multidisciplinary field with contribution of medicine and data mining. Due to this fact, previous works should be classified to cover all users' requirements from various fields. Because of this, we have studied papers with the aim of extracting knowledge from structural medical data published between 1999 and 2013. We clarify medical data mining and its main goals. Therefore, each paper is studied based on the six medical tasks: screening, diagnosis, treatment, prognosis, monitoring and management. In each task, five data mining approaches are considered: classification, regression, clustering, association and hybrid. At the end of each task, a brief summarization and discussion are stated. A standard framework according to CRISP-DM is additionally adapted to manage all activities. As a discussion, current issue and future trend are mentioned. The amount of the works published in this scope is substantial and it is impossible to discuss all of them on a single work. We hope this paper will make it possible to explore previous works and identify interesting areas for future research.

1. Introduction

The growth of information storage technology has led to generate huge amount of raw data that considers two aspects. These aspects are algorithm development and rise of modern storage equipment. Valuable knowledge can be obtained by these raw data. In early 1990s, knowledge discovery from data (KDD) term was used with the aim of knowledge extraction from database (Piatetsky-Shapiro & Frawley, 1991): "*Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data*". Desirable features for extracted knowledge are reasonable time complexity, comprehensibility, accuracy and useful result. This kind of extracted knowledge can be used as a new knowledge. Data mining was originally considered as synonym of KDD (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Recent researches have shown that application of data mining in several fields is growing such as CRM (Ngai, Xiu, & Chau, 2009), education (Romero & Ventura, 2010; Romero, Ventura, & García, 2008), clinical medicine (Bellazzi & Zupan, 2008), financial fraud detection (Kirkos, Spathis, & Manolopoulos, 2007; Ngai, Hu, Wong, Chen, & Sun, 2011), intrusion detection (Pietraszek & Tanner, 2005) and genetic data analyzing (Jiang, Tang, & Zhang, 2004). We expect its application in other fields to increase in the same manner. In a one hand, medicine, plays a great rule on human life, and on the other hand, need of automation for knowledge extraction and impossibility of manual processing, application of data mining in medicine has become a great issue. As we will see in Section 3, research on medical data mining is growing fast. Recently, application of data mining in medicine and healthcare is most widely used by data mining developers and academic researchers compared to the other fields. The rapid growth of medical data mining in the recent years represents the kick-off medical data mining.

The goals of collecting diverse medical data from various resources are to assist physicians, improve public health and support patients. All activities in medicine can be divided into six tasks: screening, diagnosis, treatment, prognosis, monitoring and management. These tasks start from a patient with a hidden disease without any symptoms, and lead to better management of the resources to improve social health care services.

* Corresponding author. Tel.: +98 9131633805.
E-mail addresses: n.esfandiary@qiau.ac.ir (N. Esfandiari), m.babavalian@qiau.ac.ir (M.R. Babavalian), eftekhari@qiau.ac.ir (Amir-Masoud Eftekhari Moghadam), vahidkashanitabar@gmail.com (V.K. Tabar).

http://dx.doi.org/10.1016/j.eswa.2014.01.011
0957-4174/

Some of the medical data mining trend anticipated here. Application of data mining in some disease is more common than others. High epidemic and mortality rate, expensive test, time consuming and requirement of special experience are expressed as the main reasons. Even though the heart disease has most mortality, number of works published on cancer is significant. This is due to the diversity, increased patients and global concerns over the time. Screening, early and accurate diagnosis of risky disease such as cancer, heart disease and diabetes is more common. This issue is progressed by developing efficient algorithms and novel technologies such as microarray data. A decision system can be used to treat in several conditions such as emergency situation, shortage of physicians and decrease human errors. The successful application and proven reliability of data mining in other tasks such as diagnosis is effective in advancing this issue. In several fields such as ICU and post surgery, we need continuously care. Then monitoring is known as another tendency. At last, the management issues such as bed management and scheduling can contribute to the improvement of medical services. By a data mining view, decision tree, artificial neural network and support vector machine (SVM) algorithms are the most popular. That is because of the simplicity and comprehensibility of decision tree, popularity and ability of artificial neural network in general model extraction and efficiency of SVM. Several challenges such as huge data; integration of text, structural data, signal and image; web and image mining; and integration with the hospital workflow wait to be tackled.

The purpose of this paper is to review medical data mining applications. We will examine similar papers in the context of review and survey to identify and understand previously presented approaches. We try to present a review study by completing previous works and integrating medical and data mining issues at the same time. For this reason, 291 papers have studied that published between 1999 and 2013 in 90 different medical, data mining and biomedical engineering journals. These studies have been used to provide medical data mining definitions, challenges, goals and to make a review. Our review is based on the six medical tasks: screening, diagnosis, treatment, prognosis, monitoring and management. In each task, related papers according to the data mining algorithms are presented. Data mining algorithms are classified as classification, regression, clustering, association rule mining and hybrid approach.

The relevance of this paper is strengthened by the fact that medical data mining is also an object of research in different communities. These communities include data mining, medicine, pattern recognition, machine learning, statistics and management. It is even more important to make present comprehensive review that covers different approaches, even those with little work carried out on. Proposed review examines each previous paper by medicine and data mining point of view.

It is quite hard to distinguish between data mining and aforementioned fields. That is why we have chosen papers aiming knowledge extraction from medical data. The number of works published on medical data mining is substantial and it is impossible to discuss all of them on a single work. Each of electronic records, medical stored events, text data, results of tests, medical images and signals and web data, can be known as the medical data. All these data types should be taken into account when implementing an appropriate algorithm for extracting knowledge. In this work, only structural data has been considered. Therefore, other types of data such as text are not considered. We hope this work will make it possible to develop a deep understanding of various approaches.

The rest of this paper is organized as follows: The next section will consider motivation for this study. Data mining concepts and medical data mining definitions and goals, main challenges, research trend and adaptive standard framework are expressed in Section 3. Sections 4, described 291 papers based on six medical tasks. In Section 5, discussion of the reviewed papers is provided to shape the current status and highlighted important issues. We will conclude and recommend for future work in final section.

2. Motivation of this study

In this section, the method and materials of this study are described. The method is referring to the procedure of collecting and gathering related papers. In the material, two subjects are tailored: glance of the prior surveys in the field of knowledge extraction from medical data, and determine the scope of this study.

As a result, 291 papers associated with application of data mining in medicine were selected that have been published between 1999 and 2013. These papers will be used in Section 3 to provide basic concepts such as definitions and goals and in Section 4 to present review. Also, the gathered collection is examined to obtain highlight issues, some facts and figures and investigate finding.

2.1. Workflow applied for extraction of medical data mining papers

The workflow used to provide this survey contains four processes: *broad search*, *refine search*, *extract basic concepts* and *analyze*. An overview of the mentioned workflow is drawn in Fig. 1.

In *broad search* initial papers that associated with the application of data mining were extracted in medicine. We selected papers aiming knowledge extraction from medicine with "data mining" term appeared in their title, abstract and keywords. To select papers two ways were taken: search in four scientific databases ELS., IEEE*xplore*, Springer and Pubmed and, search in well-known journals in data mining, bioinformatics and medicine.

Raw papers that extracted from broad search process pass to the *refine search*. In this process, abstract and introduction of each paper was studied. Some of the papers tend completely to the medicine domain and the weight of data mining was little. Thus these papers ignored. Finally, 291 papers published in 81 journals between 1999 and 2013 were remained.

Obtained papers from previous process were used to provide medical data mining definition and goals and, made a review. *Extract definition and goals* process attempt to pluralize and unify the different definition and goals given in various papers. In *analysis based medical tasks* process each paper placed in one of the six medical tasks. Moreover, in each task analyzing based on data mining view is carried out. Finally, *current issue* is discussed and *future trend* is drawn.

2.2. Previous reviews in the context of knowledge extraction in medicine

In this section previous reviews associated with extraction knowledge in medicine are outlined. We can consider each of these review papers from two points of view: algorithm and medicine. In algorithm view, some of the papers focused on a certain algorithm such as neural network and others considered a discipline of AI such as machine learning. From medicine point of view, certain disease (such as urology) or group of disease (such as cardiovascular diseases) or part of medicine tasks (such as diagnosis) has been studied. The goal of this sub-section is to better understanding of medical knowledge extraction background. For this reason, twelve review papers are studied.

In Itchhaporia, Snow, Almassy, and Oetgen (1996) application of neural network cardiovascular disease is reviewed. For this purpose cardiovascular is divided into coronary artery disease, acute myocardial infarction, electrocardiography, arrhythmia identification, arrhythmia localization, cardiac image analysis and cardiac drug dosing. In the conclusion, three main disadvantages of neural net-