Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 32 (2007) 856–867
Expert Systems with Applications
www.elsevier.com/locate/eswa

# Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis

Mu-Jung Huang [a,*], Mu-Yen Chen [b], Show-Chin Lee [a]

[a] Department of Information Management, National Changhua University of Education, Changhua 50058, Taiwan, ROC
[b] Department of Accounting, National Changhua University of Education, Changhua 50058, Taiwan, ROC

**Abstract**

The threats to people's health from chronic diseases are always exist and increasing gradually. How to decrease these threats is an important issue in medical treatment. Thus, this paper suggests a model of a chronic diseases prognosis and diagnosis system integrating data mining (DM) and case-based reasoning (CBR). The main processes of the system include: (1) adopting data mining techniques to discover the implicit meaningful rules from health examination data, (2) using the extracted rules for the specific chronic diseases prognosis, (3) employing CBR to support the chronic diseases diagnosis and treatments, and (4) expanding these processes to work within a system for the convenience of chronic diseases knowledge creating, organizing, refining, and sharing. The experiment data are collected from a professional health examination center, MJ health screening center, and implemented through the system for analysis. The findings are considered as helpful references for doctors and patients in chronic diseases treatments.


*Keywords:* Chronic disease; Data mining; Case-based reasoning

## 1. Introduction

Medical treatment is facing a challenge of knowledge discovery from the growing volume of data. Nowadays enormous data are collected continuously through health examination and medical treatment (Tan, Yu, Heng, & Lee, 2003). How to effectively share experts' knowledge (or experience) is also another important challenge in medical treatments (Cios & Moore, 2002; Gunnlaugsdottir, 2003; Hendriks & Vriens, 1999). The threats from chronic diseases to people's health are always exist and increasing gradually. How to decrease these threats is an important issue in medical treatment.

Classification rules are typically usefully for medical problems that have been applied particularly in the area of medical diagnosis (Freitus, 2002). Additionally, numerous machine-learning (ML) techniques have been applied to the field of medical treatments over the past few decades, such as artificial neural networks, genetic algorithm, fuzzy sets, inductive logic programming, and so on (Becerra-Fernandez, 2000; Cios & Moore, 2002; Evans, 1999; Kukar, Kononenko, & Silvester, 1996; Sacha & Cios, 2000; Schmidt & Gierl, 2001; Setiono, 1996). Most of these applications are particular and involve individual ML technique only, such as using data mining (DM) in a medical diagnosis (Alonso, Caraca-Valente, Gonzalez, & Monte, 2002), extracting rules from pruned neural networks for breast cancer diagnosis (Setiono, 1996), and machine learning in prognosis of the femoral neck fracture recovery (Kukar et al., 1996). A weakness of these applications is lack of systematical integration that is a critical factor to enhance the performance of applying ML in medical treatments.

Therefore, this paper proposes a model of a chronic diseases prognosis and diagnosis (CDPD) system integrating DM and case-based reasoning (CBR). The main processes of the system include (1) adopting data mining techniques to discover the implicit meaningful rules from health examination data, (2) using the extracted rules for the specific

* Corresponding author.
*E-mail addresses:* mjhuang@cc.ncue.edu.tw (M.-J. Huang), chenmy@cc.ncue.edu.tw (M.-Y. Chen).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.01.038

chronic diseases prognosis, (3) employing CBR to support the chronic diseases diagnosis and treatments, and (4) expanding these processes to work within a system for the convenience of chronic diseases knowledge creating, organizing, refining, and sharing. The experiment data are collected from a professional health examination center, MJ health screening center, and implemented through the system for analysis.

The rest of this paper is organized as follows. Reviews of related work are first described in Section 2. Section 3 describes the CDPD system architecture. We then present the knowledge creating methodology in Section 4. Section 5 briefs on how the knowledge inferring methodology to predict the probabilities of specific chronic diseases. System implementation and verification results are reported in Section 6. Finally, we make our conclusion and discuss future work in Section 7.

## 2. Literature review

### 2.1. Data mining

Data mining, known as ''knowledge discovery in databases'', is the process of discovering interesting patterns in databases that are meaningful in decision-making and is also an application area that can provide significant competitive advantage to an organization (Bolloju, Khalifa, & Turban, 2002; Bose & Mahapatra, 2001; Daskalaki, Kopanas, Doudara, & Avouris, 2003). To design a framework for a knowledge-discovery process is critical. Researchers have described a series of steps that constitute knowledge discovery process, which range from few steps to more sophisticated models such as the nine-step model suggested by Fayyad, Piatetsky-Shapiro, and Smyth (1996) or the nine-step DM knowledge discovery process model by Cios and Moore (2002).

Data mining is an explorative and complex process involving multiple iterative steps. For giving us an overview of data mining process, Fig. 1 shows a general model.

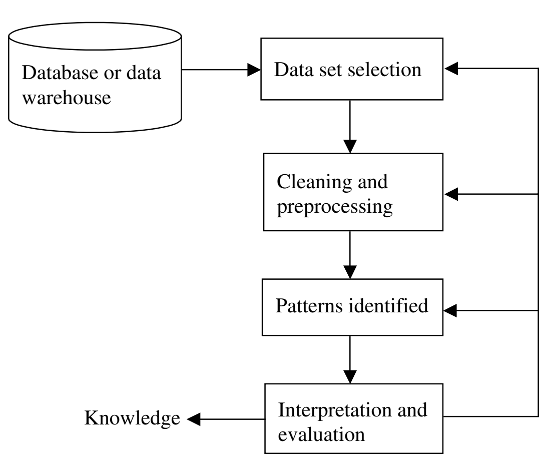


Fig. 1. Data mining phases.

According to Fig. 1, in the first step for data mining the main data sets that may be retrieved from operational databases or data warehouse are selected. The selected data set then undergoes cleaning and preprocessing for removing discrepancies and inconsistencies to improve its quality. Next, the data set is analyzed to identify patterns that represent relationships among data by applying algorithms, such as neural nets, decision trees, and so on. Then the patterns are validated with new data sets. It should be possible to transform the patterns into actionable plans that are likely to help the organization to achieve its goals. A pattern that satisfies these conditions becomes organizational knowledge. The steps in the mining process are iteratively until meaningful knowledge is extracted (Bose & Mahapatra, 2001).

DM has been successfully applied to several medical problem domains. Recent examples are as the followings. Kurgan and Cios adopted knowledge discovery approach to automated cardiac SPECT diagnosis (Kurgan, Cios, Tadeusiewicz, Ogiela, & Goodenday, 2001). Bellazzi, Larizza, Magni, and Bellazzi (2005) used temporal DM for the quality assessment of hemodialysis services. Kusiak, Dixon, and Shah (2005) used a DM approach for predicting survival time for kidney dialysis patients.

### 2.2. Case-based reasoning

CBR can mean adapting old solutions to meet new demands, using old cases to explain new situations, or reasoning from precedents to interpret a new situation (Kolodner, 1993). A planner of case-based reasoning must be a learning system because it must reuse its own experiences (Ellman, 1989). CBR requires a knowledge-based learning that makes the planner understand what should be learned and when it should be learned. This learning breaks down into three types: learning plans, learning expectations, and learning critics (Hammond, 1989).

In general, the reasoning procedure of the CBR system is as shown in Fig. 2. There are five steps in the case-based reasoning (Evans, 1999): (1)Assigning indexes. The indexes are critical features that characterize a case and determine how cases are stored in the case library. The major purpose of indexing is to allow a CBR system to retrieve one or more cases that are similar to the new problem. (2) Case retrieval. The goal of this step is to retrieve old cases stored in the case library. (3) Case adaptation. In case-based problem solving, the old solution case is used as an inspiration for solving new problems. Because a new case may not exactly match the old one, the old knowledge may often need to be fixed to fit the new one. (4) Case testing. The proposed solution will be tried out to see whether it really solves the new problem. (5) Case storage. Once the new problem is solved, it is stored in the knowledge base for future use.

In recent years, the major applications of CBR in medical are in supporting diagnosis. López and Plaza introduced a case-based system, BOLERO, which learns both