Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 33 (2007) 499–508
Expert Systems with Applications
www.elsevier.com/locate/eswa

Extracting drug utilization knowledge using self-organizing map and rough set theory

Hsin-Chuan Chou [a], Ching-Hsue Cheng [a], Jing-Rong Chang [a,b,*]

[a] *Department of Information Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Touliu, Yunlin 640, Taiwan*
[b] *Graduate School of Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Touliu, Yunlin 640, Taiwan*

Abstract

Cardiovascular disease is becoming the major cause of death in many industrialized countries. People who receive long-term treatments usually ignore the progress of the disease states. Therefore, it is critical and necessary to evaluate drug utilization and laboratory test in order to discover the knowledge that is beneath and can be extracted from those raw data. This paper utilizes techniques of self-organizing map (SOM) and rough set theory (RST) to discover the trend of individual patient's condition. With 10-fold cross-verification, the proposed SOM–SOM–RST process successfully and effectively detects patients whose diagnosis codes have been changed during the period of investigation and attains an accuracy of approximate 98%. This method can remind physicians to reevaluate the disease conditions of their patients.

Keywords: Self-organizing map (SOM); Rough set theory (RST); Cardiovascular disease; Drug utilization knowledge

1. Introduction

Pharmaceutical care is the most common treatment for ambulatory patients. When drugs are administered to a patient, the clinical goal is to achieve certain effectiveness. A modern clinician is expected to justify drug selection, hence is expected to know how potential beneficial effect are brought about (Hyfte, Maas, Tjandra-Maga, & Robbé, 2001). Although laboratory tests are tools helpful in evaluating the health status of individuals, many patients who receive long-term care do not reflect the disease state on laboratory tests significantly. Physicians, hence, might ignore progressive information revealed by these data. Upon this, it is crucial to find the trend, relationship, and knowledge among clinical data.

Cardiac disease has become the major cause of death in industrialized countries. Clinical medicine researchers have indulged themselves in the field of etiology that is based on the evidence of medicine (Department of Health, 2005; Kannel, 1997). Many researches have been done regarding the relationship between individual variables and disease states. Therefore, it is beneficial if the vascular disease is evaluated with the mapping of experimental data and treatment data (Department of Health, 2005). Drug utilization review concerns drugs interaction, drug-disease contraindication, therapeutic duplication, and early/late refill. It is rare to evaluate the laboratory values and treatment outcomes concurrently.

A physician would find it more appropriate to describe his or her knowledge by means of rules like "if fever is high and cough is moderate then disease is X" rather an "if fever is 38 °C and cough is 5–10 times a day then disease is X". Decision trees from symbolic rules are widely used to

* Corresponding author. Address: Department of Information Management, National Yunlin University of Science and Technology, 123, Section 3, University Road, Touliu, Yunlin 640, Taiwan. Tel.: +886 5 5342601x5312; fax: +886 5 5312077.
E-mail address: g9120806@yuntech.edu.tw (J.-R. Chang).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.05.020

separate and classify elements in a field of domain knowledge. The rule-based method such as ID3 algorithm (Han & Kamber, 2001) aims to minimize the complexity of the tree and thereby maximizing the information contained in each decision node. If the case is large, the decision trees tend to get quite complicated and difficult for human interpretation. Furthermore, classical IF–THEN rules can only separate the domain parallel according to a single component at a time. And decision tree would be little use where the separation rules are non-linear. Thus, it could be beneficial to utilize neural network methods in order to explore the relationship and its trend.

Techniques of utilizing soft computing in medical diagnostic support system (MDSS) are proven to be outperforming the physicians. Salchenberger, Venta, and Venta (1997) proved that the accuracy of neural network on detecting breast implant rupture than the radiologists. Baxt (1990) noted that the neural network correctly identifies 92% of patients with acute myocardial infarction and 96% of patients without infarction. Researches also stated that neural network can be more accurate than experienced cardiologist in diagnosing acute myocardial infarction (Josefson, 1997).

West and West (2000) demonstrated the SOM can be an effective tool for MDSS model selection that identify optimal target values for a breast cancer MDSS for two conditions, the lowest overall error rate and minimize the false negative error. Recently, SOM network also has been applied on medical images (Coppini, Diciotti, & Valli, 2004) and automatic ultrasound contouring for breast tumors (Huang & Chen, 2004).

Laboratory tests are the most common methods for physicians to investigate the disease state of patients. The result of laboratory test is also the evidence to claim health insurance reimbursement. For example, diabetes or cardiac disease patients, whose TG level over 200 mg/dl can receive treatment to achieve the goal of lowering than 160 mg/dl. In this paper, laboratory tests that relate to the disease under evaluation were extracted from clinical database.

This paper utilizes techniques of two-level self-organizing map (SOM) (Kiviluoto & Bergius, 1998) and rough set theory (RST) (Pawlak, 1982; Pawlak, 1991) to discover the trend of individual patient's condition. The process of SOM–SOM–RST is to perform SOM for laboratory data and medicine treatments in Taiwan, and then IF–then rules are extracted by RST.

Following Section 1, the related literatures are briefly reviewed in Section 2. Section 3 introduces proposed model in knowledge discovery. In Section 4, the experimental results are described, and conclusions are finally made in Section 5.

2. Preliminary

In this section, we briefly introduce the literatures about knowledge discovery in database, SOM, and RST.

2.1. Knowledge discovery in database (KDD)

Data occurring in medical applications are often characterized by at least one of the following properties that often cause difficulties (Villmann, 2002):

1. small data sets,
2. non-linear and high-dimensional data sets,
3. large noise without any clearly reproducible cause,
4. large data in image processing,
5. categorical and non-metric data (Ermini & Marchesi, 1998).

Upon the characteristics mentioned above, traditional methods such as conventional statistics and data analysis often fail. Artificial neural networks can help overcome some of the difficulties.

Artificial intelligence is the study of knowledge representations and their use in language, reasoning, learning, and problem solving (Rich & Knight, 1991). It attempts to encode knowledge digitally and electronically using computational power (Bui, 2004). On the contrary, knowledge can be discovered by artificial intelligent techniques. It has been interested in knowledge discovery in database (KDD). The finding of useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archaeology, and data pattern processing. KDD refers to the overall process of discovering useful knowledge from data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996; Han & Kamber, 2001). The overview of KDD process is illustrated in Fig. 1.

Though most work on KDD focus on data mining, however, the KDD process can involve significant iterations and can contain loops between any two steps. Furthermore, the KDD process involves numerous steps with many decisions made by user (Fayyad et al., 1996; Han & Kamber, 2001). To reduce user manipulation, this paper utilize self-organizing map, one of neural network techniques, to perform the discovery.

2.2. Self-organizing map (SOM)

Unsupervised learning of neural networks has been successfully applied in data mining and visualization (Ritter, Martinetz, & Schulten, 1992). There are many versions of SOMs: temporal Kohonen map (Coppini et al., 2004), recursive SOM (Voegtlin, 2002), tree structured SOM (Hagenbuchner, Sperduti, & Tsoi, 2003), and cell-splitting self-organizing neural networks (Chow & Wu, 2004). The basic SOM network has an input layer and an output layer. Also, the basic SOM is a topology-preserving map because there is a topological structure imposed on the nodes in the network. A topological map is simply a mapping that preserves neighborhood relations. SOM learns to classify input vectors according to how they are grouped. When training data sets are fed into the network, SOM will com-