Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 33 (2007) 794–808
Expert Systems with Applications
www.elsevier.com/locate/eswa

Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan

R.J. Kuo [a,*], S.Y. Lin [a], C.W. Shih [b]

[a] Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei 106, Taiwan, ROC
[b] Department of Industrial Engineering and Management, National Chiao-Tung University, Shin-Chu 300, Taiwan, ROC

Abstract

In addition to sharing and applying the knowledge in the community, knowledge discovery has become an important issue in the knowledge economic era. Data mining plays an important role of knowledge discovery. Therefore, this study intends to propose a novel framework of data mining which clusters the data first and then followed by association rules mining. The first stage employs the ant system-based clustering algorithm (ASCA) and ant K-means (AK) to cluster the database, while the ant colony system-based association rules mining algorithm is applied to discover the useful rules for each group. The medical database provided by the National Health Insurance Bureau of Taiwan Government is used to verify the proposed method. The evaluation results showed that the proposed method not only is able to extract the rules much faster, but also can discover more important rules.

Keywords: Data mining; Ant colony system; Cluster; Association rule

1. Introduction

In recent years, there are dramatic changes in the human life, especially the information technology. It has become the essential part of our daily life. Its convenience let us more easily to store any kind of the information regarding science, medicine, finance, population statistics, marketing and so on. However, if there is not a useful method to help us apply these data, then they are only the garbage instead of resources. Due to such demand, there are more and more researchers who pay more attention on how to use the data effectively as well as efficiently. And this is so called data mining.

Data mining includes many areas, in which there are databases techniques, artificial intelligence, machine learning, neural network, statistical techniques, pattern recognition, data visualization etc., is growing up very quickly. It is assigned an objective to find the hidden knowledge or information, which may be helpful to make decisions for business or policies, from large database automatically. Data mining can be classified into some topics, like classification, estimation, forecasting, clustering, association rule and sequential pattern (Peacock Peter, 1998). Among them, this study intends to propose a framework which integrates both the clustering analysis and association rules mining to discover the useful rules from the database through ant colony optimization system.

Therefore, the proposed method is consisted of two components: (1) clustering analysis and (2) association rules mining. The first stage employs the ant system-based clustering algorithm (ASCA) and ant K-means (AK) to cluster the database, while the ant colony system-based association rules mining algorithm is applied to discover the useful rules for each group. The reason to clustering the database first is that this can dramatically decrease the mining time. In order to assess the proposed method, a database being provided by the National Health Insurance Plan of Taiwan Government is applied. This database has accumulated 12 millions administrative and claims data, which is the largest database in the world. Basically, this work is a cooperation of National Health Research

* Corresponding author. Fax: +886 2 27763996.
E-mail address: rjkuo@ntut.edu.tw (R.J. Kuo).

0957-4174/$ - see front matter © 2006 Published by Elsevier Ltd.
doi:10.1016/j.eswa.2006.08.035

Institute with the National Health Insurance Bureau of Taiwan Government in order to establish a Nation Health Insurance research database. The computational results show that the proposed method not only can extract the useful rules faster, but also can provide more precise rules for the medical doctors.

The rest of this paper is organized as follows. Section 2 summarizes some general background for data mining, clustering analysis, association rule and ant colony optimization system, and the proposed method is presented in Section 3. The result of real world data with the proposed method is illustrated in Section 4. Finally, concluding remarks are made in Section 5.

2. Background

This section will briefly review four aspects of literatures. They include data mining, clustering analysis, association rule mining and ant colony optimization system algorithm. Detailed information is presented in the following subsections.

2.1. Data mining

In the past study, Fayyad et al. had defined the knowledge discovery in database (KDD) as a nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data (Fayyad, Piatetsky-Shapiro, & Smyth, 1996, 1997). By the term process shows that KDD is made up of several steps, which involve the selection, preprocessing, transformation, data mining, and interpretation/evaluation. Data mining is a multi-disciplinary field that is at the intersection of statistics, machine learning, database management, and data visualization, to provide a new perspective in data analysis (Peacock Peter, 1998).

The following five foundation-level analysis domains are the ''reason why'' of using data mining: summarization, predictive modeling, clustering/segmentation, classification, and link analysis (Peacock Peter, 1998). Link analysis refers to a family of methods that are employed to correlate patterns cross-section over time with each other. In the marketing, a link analysis model can provide information about the buyers' behavior. Using the same idea to analyze medical behavior, link analysis can find patterns in patients' visits to doctors. This can be helpful for diagnosis and deciding on drugs. If a medical analyst could find out which groups of sets of items are most likely to be diagnosed in a particular group of patients, he can make several treating strategies, depending on the results of link analysis for their regular uses to make more effects. Because of its importance in medical science, the following study we will focus on this issue.

2.2. Clustering analysis

The goal of clustering analysis is to group similar objects together. There are many methods applying in clustering analysis, such as hierarchical clustering, partition-based clustering, density-based clustering, and artificial intelligence-based clustering. In this subsection, the artificial intelligence-based clustering, which includes artificial neural networks (ANN) and genetic algorithm (GA), was illustrated. The others are introduced in the survey research (Bellaachia, Portnoy, Chen, & Elkahoun, 2002; Witten & Frank, 2000; Berkhin, 2002).

2.2.1. Applications of ANN in clustering analysis

The artificial neural network (ANN) is a system which has been derived through models of a collection of simple nonlinear computing elements whose inputs and outputs are linked to form the network (Kohonen, 1991).

Kohonen's feature maps (also called Self-Organizing Feature Map, SOM) is the most widely applied unsupervised learning scheme. The SOM has two layers that include input layer and output layer. The input layer is fully connected to the output layer that is a two-dimensional layer. Each output layer nodes measures the Euclidean distance of its weights to the incoming input vector. The weights of winning node that has the smallest distance in the output layer are adjusted to be closer to the vector of the input nodes.

Because SOM can map the input vectors with high dimensions into 2-D space, it is easier to visualize the data and cluster analysis. In other word, most applications of cluster analysis with SOM are to observe the mapping network by vision, and then determine the distribution of clusters. Recent years, there ware many studies that improved the efficacy and efficiency of SOM. Such as, the Double SOM that can adjust at learning stage and let the nodes that have similar input vectors produce similar weight vectors and come near (Su & Chang, 2000, 2001). But it may have different results by observing the mapping network with the same data. Resson proposed adaptive double SOM (ADSOM) (Fayyad et al., 1996) that combines features of the popular SOM with two-dimensional position vectors, which serve as a visualization tool to detect the number of clusters presented in the data. ADSOM allowed automating detection of the number of clusters with a novel index that is introduced on the base of hierarchical clustering of the final locations of position vectors. Thereby, reducing human error could be incurred from counting clusters visually.

Adaptive resonance theory (ART) is another widely applied unsupervised learning scheme. ART include ART1, which is applicable for binary input, and Art2 which is used to deal with continuous input (Carpenter & Grossberg, 1987). Unlike traditional SOM, ART network can determine the actual number of cluster with any visual examination.

2.2.2. Application of GA in clustering analysis

GA-based clustering algorithm was proposed by Maulik and Bandyopadhyay (2000). It can improve result of the conventional statistics methods, like K-means, that are easy to find a local minimum (Maulik & Bandyopadhyay,