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,

2000). Krishna and Murty proposed Genetic K-means Algorithm (GKA) for clustering analysis, which defines a biased mutation operator specific to clustering called distance-based-mutation. And they proved that GKA can converge to the best known optimum by using finite Markov chain theory (Krishna & Murty, 1999).

2.3. Association rule mining

The issue of mining association rules was first addressed in 1993 (Agrawal, Imielinski, & Swami, 1993a, 1993b). They pointed out that there are some hidden relationships among the purchased items in transactional databases. For example, there are associations or relationships between items such as bread and milk, which are often purchased together in a single basket transaction. The mining results can help understand the customer's purchase behavior, which might not have been previously perceived.

An association rule is of the form $X \Rightarrow Y$, where X and Y are both frequent itemsets in the given database and the intersection of X and Y is an empty set, i.e., $X \cap Y = \varnothing$. The support of the rule $X \Rightarrow Y$ is the percentage of transactions in the given database that contain both X and Y, i.e., $P(X \cup Y)$. The confidence of the rule $X \Rightarrow Y$ is the percentage of transactions in the given database containing X that also contains Y, i.e., $P(Y|X)$. Therefore, association rule mining is used to find all the association rules among itemsets in a given database, where the support and confidence of these association rules must satisfy the user-specified minimum support and minimum confidence. The problem of association rule mining can be divided into two subproblems:

1. Finding frequent itemsets with their supports above the minimum support threshold.
2. Using frequent itemsets found in the step 1 to generate association rules that have a confidence level above the minimum confidence threshold.

Therefore, many studies of association rule mining concentrate on developing efficient algorithms for frequent itemset discovery. The following subsections summarize some of the most popular algorithms for frequent itemset mining.

2.3.1. Apriori-like algorithm

Agrawal et al. (1993a, 1993b) proposed the well-known algorithm, Apriori, to mine large itemsets to find out the association rules among items. This algorithm employs a level-wise approach, which iteratively generates candidate k-itemsets from previously found frequent $(k-1)$-itemsets, and then checks the supports of candidates to form frequent k-itemsets. The algorithm scans multiple passes over the database. The efficiency and correctness of the level-wise generation of frequent itemsets are based on an important property, called the Apriori Property.

The algorithm is first pass counts item occurrences to find the set of frequent 1-itemsets, denoted as L_1. A subsequent pass, say pass k, consists of two steps; the join and prune steps. In the join step, a set of candidate k-itemsets (denoted as C_k) is generated by joining the frequent itemsets L_{k-1} found in the $(k-1)$th pass with itself. For example, Fig. 1 demonstrates how to find frequent itemsets in min_sup = 2.

2.3.2. FP-growth algorithm

Han, Pei, and Yin (2000) proposed a novel frequent pattern tree (FP-tree) structure, which contains all the compact information for mining frequent itemsets, and then proposed the FP-growth algorithm, which adopts a pattern segment growth approach to prevent generating a large number of candidate itemsets. Their mining method only scans the whole database twice and does not need to generate candidate itemsets, and so it is very efficient.

2.3.3. Parallel mining

Parallel mining (Agrawal & Shafer, 1996) is another technique used to improve the classic algorithm of mining association rules on the premise that there exist multiple processors in the computing environment. The core idea of parallel mining is to separate the mining tasks into several sub-tasks so that each sub-task can be performed simultaneously on various processors, which are embedded in the same computer system or even spread over the distributed systems. Thus; this improves the efficiency of the overall algorithm for mining association rules.

2.3.4. Sampling algorithm

A random sampling technique (Toivonen, 1996) was used to find association rules to reduce the database activity. The sampling algorithm applies the level-based method on the sample with lower minimum support threshold to mine the superset of large itemsets. This method produces exact association rules, but in some cases it does not generate the entire association rules, that is, there might exist some missing association rules. Therefore, this approach

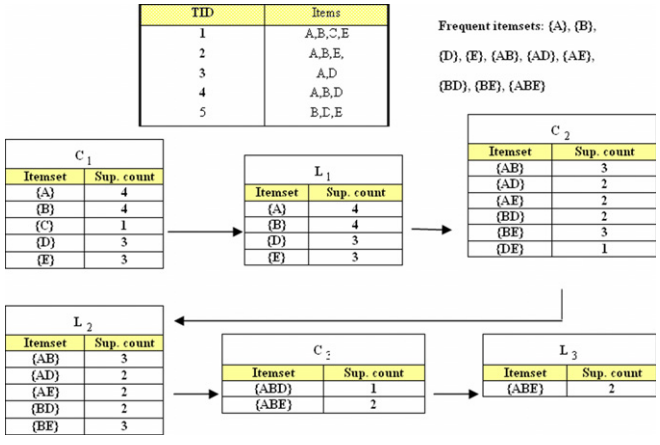

Fig. 1. Frequent itemset mining by Apriori algorithm.

requires only one full pass over the database in most cases, and only two passes in the worst case.

2.3.5. Lattice-based algorithm

Zaki (2000) organized the items into a structure of lattice and presented a set of algorithms including Eclat, MaxEclat, MaxClique, TopDown and AprClique for identifying maximal large itemsets. All of the algorithms attempt to look ahead and identify long large itemsets early to help prune off the number of candidate itemsets considered. There are also another two approaches for mining long large itemsets. Lin and Kedem (2002) proposed Pincer–Search algorithm for mining long large itemsets, whereas Bayardo (1998) proposed the Max–Miner algorithm. Both algorithms attempt to discover the long and large scale patterns through the search effort. The greatest difference between the two methods is in the generation of candidate itemsets. The Max–Miner approach generates the candidate itemsets in polynomial time since it is an NP-hard problem in the Pincer–Search method to ensure that no long candidate itemsets contain any known infrequent itemset.

2.3.6. Partition algorithm

For mining association rules, Savasere, Omiecinski, and Navathe (1995) introduced a partition algorithm that is fundamentally different from the classic algorithm. First, a partition algorithm scans the database once to generate a set of all potentially large itemsets, and then the supports for all the itemsets are measured in the second scan of the database. The key to correctness of the partition algorithm is that a potentially large itemset appears as a large itemset in at least one of the partitions. This algorithm logically divides the database into a number of non-overlapping partitions, which can be held in the main memory. The partitions are considered individually and all large itemsets for that partition are generated. These large itemsets are further merged to create a set of all potential large itemsets. Then these itemsets are generated.

2.3.7. Cluster-decomposition association rule algorithm

Tsay and Chang-Chien (2004) and Zhang and Li (1993) proposed the cluster-based association rule (CBAR), which creates cluster tables by scanning the database once, and then clustering the transaction records by the length of record. Moreover, the large itemsets are generated by contrasts with the partial clusters. They found that CBAR could improve more efficiency with the increasing of database size or the decreasing of minimum support.

2.3.8. Proximus

Koyuturk, Grama, and Ramakrishnan (2005) proposed an efficient framework, PROXIMUS, for error-bounded compression of high-dimensional discrete-attribute data sets. Given a transaction set on a set of items, we can construct a binary transaction matrix by mapping transactions to rows and items to columns and setting entry T_{ij} of transaction matrix T as 1 if item j is in transaction $T_{i\bullet}$. And then, decompose the matrix T to n kinds different transaction sets by finding rank-one approximation of T: $[x_1,\ldots,x_n]\bullet[y_1,\ldots,y_n]^{\mathrm{T}}$. By using this method, the framework can condense the large transaction data to n different kinds of virtual transaction sets, each of them that is associated a weight, which is defined as the number of non-zeros in the corresponding presence vector, i.e., the number of transactions that contain the corresponding pattern. Finally, the framework can reduce the size of transaction sets. In other words, it can have higher performance in mining association rule from transaction sets.

2.4. Ant colony optimization algorithm

2.4.1. The concept of ant colony optimization system

In the real world, ants communicate with others by a trail of chemicals called "pheromones" which are deposited by ants when they search for food. Then, the other ants encounter the previously laid pheromones and decide how many probabilities they will follow. As more and more ants pass by the same path, the pheromones on the shorter path would be increased, but the pheromone would evaporate on the other paths, as illustrated in Fig. 2.

2.4.2. Ant colony system

The ant colony system (ACS) is based on agents that simulate the natural behavior of ants, develop mechanisms of cooperation and learn from experiences (Dorigo & Gambardella, 1997). The heuristics have been shown to be robust and versatile for different problems. In addition, ACS is a population-based heuristics that enables the exploration of the positive feedback between agents as a search mechanism.

ACS is a particular algorithm of ACO whereas the real ants are able to communicate information concerning food sources via an aromatic essence. While searching for food, they secrete a pheromone to mark the path leading to food source. When there are more pheromones on a path, there

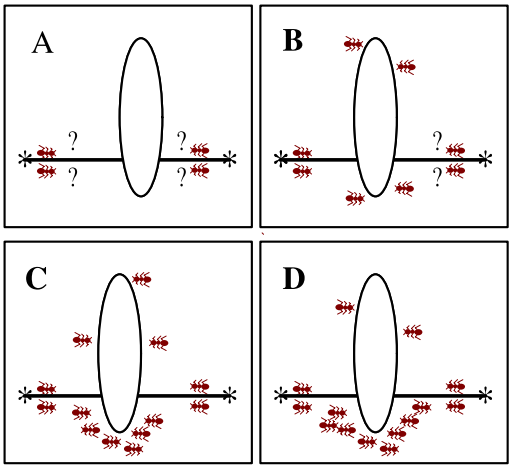

Fig. 2. The behavior of real ants.

is larger probability that other ants will use that path, and therefore the pheromone trail on such a path will grow faster and attract more ants to follow. In the ACS, the method whereas ants select the path is changed, called ACS state transition rule. When ant k in the city r will go to next city s, the selection rule is:

$$s = \begin{cases} \arg\max_{u \in J_k(r)} \{\tau(r,u) \cdot \eta(r,u)^{\beta}\}, & \text{if } q \leqslant q_0 \\ S, & \text{otherwise} \end{cases} \quad (1)$$

where $0 \leqslant q \leqslant 1$ is randomly produced and q with $0 \leqslant q_0 \leqslant 1$ is the random parameter of the system. S is the city by the random-proportional rule selection, which is defined as

$$p_k(r,s) = \begin{cases} \dfrac{\tau(r,s) \cdot \eta(r,s)^{\beta}}{\sum_{u \in J_k(r)} \tau(r,u) \cdot \eta(r,u)^{\beta}}, & \text{if } s \in J_k(r) \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where τ is called pheromone trials, and η is $1/d$ between the nodes. Thus, d represents distance, J_K means that ant k is non-passed city after ant k pass city r, and β is another system parameter in the ant colony system.

These two formulas are overall called pseudo random proportional rules, and they are according to the method of Ant-Q in the ant evolution. There are three models for the station transition rule: pseudo random, pseudo random proportional, and random proportional. Eq. (1) is called the act of exploitation as $q \leqslant q_0$; otherwise s is equal to S, which is called the act of biased exploration.

In the ACS, the pheromone trials are divided into two parts, the ACS global and local updating rules, respectively. The ACS global updating rule is referred to the ANT-cycle method in the ant system. When ants have completed all their tours, the pheromone trial could be renewed, which is called the offline method. The ACS local updating rule is referred to as the ANT-density method in the ant system. When an ant is walking, each step renews the pheromone trail once, called online method.

The ACS global updating rule is presented as

$$\tau(r,s) = (1-\alpha) \cdot \tau(r,s) + \alpha \cdot \Delta\tau(r,s) \quad (3)$$

where

$$\Delta\tau(r,s) = \begin{cases} \frac{1}{L_{gb}}, & \text{if } (r,s) \in \text{global-best-tour} \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

In addition, $0 < \alpha < 1$ is called the pheromone decay parameter, and L_{gb} is the shortest path from first point to current point (In the TSP problem). ACS is the method which means the path that ants find the shortest path from the start to current points. Therefore, it can reach the optimal solution.

The ACS local updating rule is presented in the following equation:

$$\tau(r,s) = (1-\rho) \cdot \tau(r,s) + \rho \cdot \Delta\tau(r,s) \quad (5)$$

where $0 < \rho < 1$ is called the pheromone evaporate parameter, and $\Delta\tau\,(r,s) = \tau_0$.

The ACS local updating rule is similar to the ACS global updating rule. It is increased by a fix quantity of pheromone trails every time. When the pheromone on the original path is bigger than τ_0, the pheromone value on the path is decreased after the local updating rule. This can prevent a larger number of ants using the same path, which causes pheromone trials to stagnate. When ants travel from one to other items, they could do local updating; and when ants finished their travel once, global updating is implemented.

2.4.3. Ant colony system in clustering analysis

Tsai, Wu, and Tsai (2002) proposed the algorithm that was named ant colony optimization with differently favor (ACODF). ACODF algorithm has the following desirable strategies:

1. It uses differently favorable ants to solve the clustering problem.
2. ACODF adopts simulated annealing concept for ants to decreasingly visit the amount of cities and get the local optimal solutions.
3. It utilizes tournament selection strategy to choose a path.

Every ant only needs to visit few cities instead of all of cities. Thus, the ant will reduce visiting the cities every iterations. After several iterations, the closer nodes are, the higher trail intensity will be. On the other hand, the further nodes are, the lower trail intensity will be. Therefore, ants will favor to visit the closer nodes and then reinforcing the trail with their own pheromone. Finally, the clusters will be built by dividing the pheromone that was laid on the edge between the data points.

Kuo and his colleagues proposed the Ant System-based clustering algorithm (ASCA) (Kuo, Cha, Chou, Shih, & Chiu, 2003) and Ant K-means algorithm (AK) (Kuo, Wang, Hu, & Chou, 2005) to solve the problem of clustering analysis. They combined these two algorithms as a two-stage clustering method, which uses ASCA to determine the number of cluster, and then uses AK to optimize the result of clusters. The AK modifies the K-means as locating objects in cluster with the probability, which is updated by the pheromone, while the rule of updating pheromone is according to total within cluster variance (TWCV).

In Yang, Sun, and Huang (2002), they applied the ant colony system (ACS) for clustering problem. Based on ACS, it treats the data (objects or elements) as the ants. Thus, each ant has different properties. Basically, the process of data clustering is the process of ant looking for food.

2.4.4. Ant in association rule mining

The ant system employed for mining association rules is a very new application, although there have been many

applications in data mining. Regarding the application of ant colony system for mining association rules, Su (2002) adopted the technique and concept of Ant System to develop association rules. The developed algorithm is supported by quality data, quantity data, and mix data. According to its results, the ant system must take more time on running the data in assign cycle; and if the data is critical or has time constraints, it may not feasible. Furthermore, there are some parameters in the ant algorithm which need to be pre-determined, which may be time consuming. In order to resolve the foregoing problems, Kuo and Shih (accepted for publication) used the constraints concept to decrease the run time, and let almost all of the parameters be known before running the model. In this study and (Shih, 2004), the algorithms which are based on Ant colony system were proposed to mine the association rules.

3. Methodology

The proposed framework is described in this section. The following subsections will describe the problem definition and the proposed method, Ant System-based Clustering Algorithm (ASCA), Ant K-means Clustering Algorithm (AK), and ACS-based association rule mining algorithm. The mining stages are shown in Fig. 3.

3.1. Clustering algorithm

3.1.1. Definitions and notations

The following terms and notations are used throughout this study:

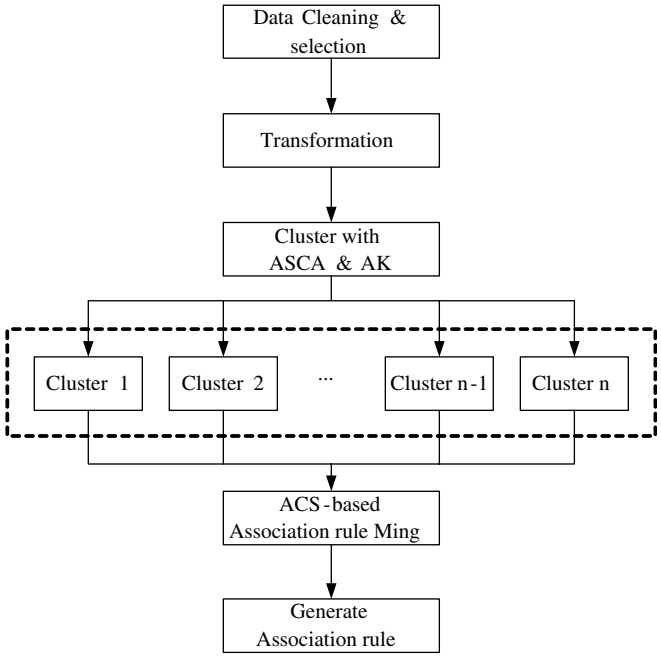

Fig. 3. The mining stages of the study.

Objects	A_1	A_1	⋯	A_1
O_1	1.22	32.5	⋯	56.4
⋮	⋮	⋮	⋯	⋮
O_n	55.6	5.6	⋯	8.4

Fig. 4. The format of the data set.

- Let $\boldsymbol{E} = \{O_1, O_2, \ldots, O_n\}$ be the set of n data or objects, where $\boldsymbol{O}$ is the objects (or data, item) collected from the database. And each object has $\boldsymbol{k}$ attributes, where $k > 0$ (see Fig. 4).
- $\boldsymbol{\alpha}$: The relative importance of the trail, $\alpha \geqslant 0$.
- $\boldsymbol{\beta}$: The relative importance of the visibility, $\beta \geqslant 0$.
- $\boldsymbol{\rho}$: The pheromone decay parameter, $0 < \rho < 1$.
- $\boldsymbol{Q}$: A constant.
- $\boldsymbol{n}$: Number of objects.
- $\boldsymbol{m}$: Number of ants.
- $\boldsymbol{nc}$: Number of clusters.
- $\boldsymbol{T}$ is the set includes used objects. The maximal number recorded by $\boldsymbol{T}$ array will be n, i.e. $\boldsymbol{T} = \{O_a, O_b, \ldots, O_t\}$, where $a, b, \ldots, t$ are the points that ant has been.
- $\boldsymbol{T}_k$: the set $\boldsymbol{T}$ is performed by ant k.
- $\boldsymbol{O}_{\text{center}}(\boldsymbol{T})$: the object which is the center of all objects in T, i.e.,

$$O_{\text{center}}(T) = \frac{1}{n_T} \sum_{O_i \in T} O_i, \tag{6}$$

where n_T is the number of objects in T.

- $\boldsymbol{TWCV}$: Total within cluster variance, i.e.,

$$\sum_{k=1}^{nc} \sum_{i \in k} (O_i, O_{\text{center}}(T_k))^2. \tag{7}$$

3.1.2. Ant system-based clustering algorithm (ASCA)

The algorithm of ASCA is including four sub-procedures, that is ***Divide***, ***Agglomerate_obj, Agglomerate, and Remove***. Following is the subscribing of procedures of ASCA. First, initialize the parameters and group all the objects as a cluster. And then the sub-procedure ***Divide*** will divide the cluster into several sub-clusters and some object which does not belong to any sub-clusters through the consistency of the pheromone and some criterion. After ***Divide***, the ***Agglomerate_obj*** is the next step at this algorithm in order to agglomerate the objects into the suitable sub-cluster. Fourth, ***Agglomerate*** is the sub-procedure to merge the similar two sub-clusters into a cluster. And then run ***Agglomerate_obj*** again. Sixth, after agglomerating the similar object into the suitable sub-cluster, the ***Remove*** sub-procedure tries to remove the un-similar from sub-cluster. Calculate the total within cluster variance (TWCV). If

TWCV is not changed, grouping the non-clustered objects to the closest cluster, and stop the procedure. Otherwise, repeat the sub-procedure ***Divide, Agglomerate_obj, Agglomerate, Agglomerate_obj, Remove***, round and round until TWCV is not changed. The detail algorithm of ASCA is introduced in Kuo et al. (2003) as shown in Fig. 5.

3.1.3. Ant K-means clustering algorithm (AK)

Ant K-means Algorithm (AK) (Kuo et al., 2005) is the second stage of clustering. AK modifies the K-means as locating the objects in a cluster with the probability which is modified by the pheromone. And the rule of updating pheromone is according to total within variance. The process is as following. The first step is initializing the parameters including the number of clusters and its centroid. Then, lay equal pheromone on each path. Third, each ant k chooses the centroid to move with $\boldsymbol{P}$, i.e.,

$$P_{ij}^{k} = \frac{\tau_{ij}^{\alpha}\eta_{ij}^{\beta}}{\sum_{c}^{nc}\tau_{ic}^{\alpha}\eta_{ic}^{\beta}}, \tag{8}$$

where i is the start point, j is the end point (centroid) which ant k chooses to move, c is the centroid and nc is the num-

Procedure ***Ant System_based Clustering Algorithm***

Initialize the parameters.

Group all objects as a cluster.

Do

Divide for all ant k.

Agglomerate_obj for all ant k.

Agglomerate for all ant k.

Agglomerate_obj for all ant k.

Remove for all ant k.

Group the non-clustered objects as a cluster.

Calculating TWCV (Total Within Cluster Variance).

While (TWCV is not chance)

Grouping the objects which are not clustered to the closest group.

Procedure ***Divide***

Lay pheromone on the path by η_{ij} for all i and j, $i \neq j$.

Calculating $\bar{\tau}$.

Updating pheromone by

$$\tau_{ij} \leftarrow (1-\rho)\,\tau_{ij} + \Delta\tau_{ij} \text{ where } \Delta\tau_{ij} = \begin{cases} \frac{1}{d_{ij}} & \text{if } \tau_{ij} > \bar{\tau} \\ 0 & \end{cases}$$

for all i and j, $i \neq j$.

Calculating $\bar{\tau}_i$ for all $i=1,2,3,\ldots,n$.

Each ant k starts at the object i which $i=\boldsymbol{Max}\{\bar{\tau}_i \mid i=1,2,3,..,n\}$, if the object i had been collected by another ant, ant k will stop search.

Each ant k collects object j if $\tau_{ij} \geq \bar{\tau}$ for k=1 to m.

If the number of objects collected by ant k is more than $\boldsymbol{\theta}$, ant k continues collecting object j, or set object j free, *i.e.*:

If $\tau > \bar{\tau}$ where $i \in T_k \quad j \in \{n - T_k \mid k = 1,2,..,m\}$

Else set object j free.

Fig. 5. The procedure of Ant System-based Clustering Algorithm.

Procedure ***Agglomerate_obj***

Let $\boldsymbol{C}$ be the collector which collects O_j to T_k with the following equation.

Let $\boldsymbol{C} = \phi$.

If O_j satisfied with the following equation *i.e.,*

$$D(O_{center}(T_k), O_j) \geq D_{mean}(T_k) + 3Dev(T_k) \quad \text{where } j \in \{n - T_k \mid k = 1,2,...,m\},$$

add O_j to $\boldsymbol{C}$.

Else $\boldsymbol{C} = \phi$.

If ($\boldsymbol{C} \neq \phi$)

If ($\boldsymbol{C} < 2$) Assign O_j to T_k.

Else ($\boldsymbol{C} \geq 2$) Assign O_j to T_k if the distance of $O_{center}(T_k)$ and O_j is minimum.

Else continue.

Procedure ***Agglomerate***

Do

Let $\boldsymbol{C}$ be the collector which collects the T_k satisfied with the following equation.

Let $\boldsymbol{C} = \phi$.

If T_k satisfied with the following equation *i.e.,*

$$D(O_{center}(T_{k_i}), O_{center}(T_{k_j})) \leq D_{mean}(T_{k_i}) + D_{mean}(T_{k_j}) - \xi\, Dev(T_{k_i}) - \xi\, Dev(T_{k_j})$$

for all k_i and k_j, $k_i \neq k_j$ $i, j = 1,2,3,\ldots,m$, add T_k into $\boldsymbol{C}$.

Else $\boldsymbol{C} = \phi$.

If ($\boldsymbol{C} \neq \phi$)

If ($\boldsymbol{C} < 2$) Agglomerate T_{k_i} *and* T_{k_j} as a cluster.

Else ($\boldsymbol{C} \geq 2$) Agglomerate T_{k_i} *and* T_{k_j} as a cluster if the distance of

$O_{center}(T_{k_i})$ and $O_{center}(T_{k_j})$ is minimum.

Else ($\boldsymbol{C} = \phi$) continue.

While (there is no more cluster could be agglomerated)

Procedure ***Remove***

Remove object j from Ant k if $D(O_j, O_{center}(T_k)) > 2Dev(T_k)$, where $j \in T_k$ and $T_k = 1,2,3,...,m$.

Fig. 5 (*continued*)

ber of centroids. Therefore, if the value of P_{ij} is bigger than others, ant k will move from point i to point j, i.e., object i belongs to centroid j. Fourth, update the pheromone by

$$\tau_{ij} \leftarrow \tau_{ij} + \frac{Q}{\text{TWCV}}, \tag{9}$$

where Q is the constant, TWCV is the total within cluster variance. And then, calculate $O_{\text{center}}(T_k)$ where $k = 1, 2, 3, \ldots, nc$. After that, calculate TWCV. If TWCV is changed, go back to third step; otherwise, if TWCV is smaller than smallest TWCV, replace it. The next step, run the procedure ***Perturbation*** to leap from the local minimal solution. If the number of iterations is not reached, go back to third step; otherwise, stop this algorithm. Fig. 6 shows the procedure of Ant K-means algorithm.

Procedure *Perturbation*

Each Ant starts at random object and chooses the centroid randomly of cluster to move for all Ant k.

Calculating $O_{center}(T_k)$ where $k = 1,2,...,nc$ and TWCV.

Procedure *Ant_K-means*

Input the number of clusters and the corresponding centroids, and set the parameter α, β, ρ, number of iterations and ants.

Lay equal pheromone on each path.

While (the number of iterations is not reached)

Do

Updating pheromone by $\tau_{ij} \leftarrow \tau_{ij} + \frac{Q}{TWCV}$.

Each Ant k chooses the centroid to move with P, *i.e.*,

$$P = \frac{\tau_{kc}^{\alpha} \cdot \eta_{kc}^{\beta}}{\sum_{i=1}^{nc} (\tau_{ki}^{\alpha} \cdot \eta_{ki}^{\beta})}.$$

Calculate $O_{center}(T_k)$ where $k = 1,2,...,nc$

Calculate TWCV (Total Within Cluster Variance).

While (TWCV is not changed)

If TWCV is smaller than the smallest TWCV, replace it.

Perturbation.

Fig. 6. The procedure of Ant K-means.

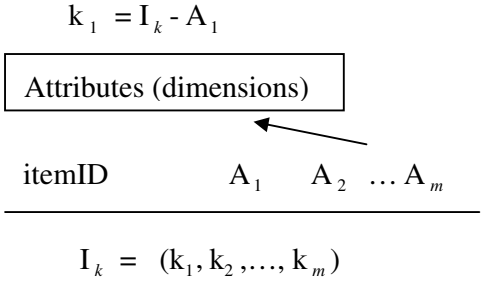

Fig. 7. Multi-dimensional items.

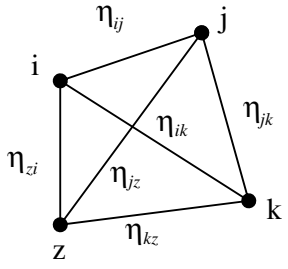

Fig. 8. An simply example.

3.2. The ACS-based association rule mining

In this section, ACS-based association proposed in Fayyad et al. (1996) is introduced.

3.2.1. Problem definition

Let $\Phi = \{I_1, I_2, \ldots, I_n\}$ be a set of all items, where an item is an object with m dimensional attributes ($m \geqslant 1$) that are so-called dimensions (e.g., weight, high, cost, ...etc), as illustrated in Fig. 7. The value k_m is on dimension $A_j[j \in \{1,2,\ldots,m\}]$ of item $I_k - A_j$.

Definition 3.1. Association Rules Mining

If $r(A_i, A_j) = \tau_{ij} \geqslant \tau_{\text{threshold}}$, $(-1 \leqslant \tau_{ij} \leqslant 1)$, represents the degree of relations between A_i, A_j, $\forall\, i, j = 1, \ldots, n$ with $r(A_i, A_j) = r\ (A_j, A_i)$ and $r(A_i, A_i) = 1$, then an association rule is an expression of $A_i \iff A_j$, for any $A_i, A_j \in A$ when $\tau_{ij} \geqslant \tau_{\text{threshold}}$, with $0 \leqslant \tau_{\text{threshold}} \leqslant 1$.

3.2.2. ACS-based association rule mining algorithm

The ACS-based association rule mining algorithm (Shih, 2004) was applied to mine the association rule. The association rules with n items construct a complete graph, where each pair of vertices is joined by an edge, as illustrated in Fig. 8. Let η_{ij} be the frequency between items i and j. The η_{ij} the edges represents the frequency between items.

Let $b_i(t)$ $(i = 1, \ldots, n)$ be the number of ants at item i at time t and let $m = \sum_{i=1}^{n} b_i(t)$ be the total number of ants at time t. All ants will follow:

1. Ant chooses next item j to follow by the state transition rule that is defined by

$$j = \begin{cases} \arg\max_{u \in Z} \{\tau_{iu}(t) \cdot \eta_{iu}^{\beta}\}, & \text{if } q \leqslant q_0 \\ S, & \text{otherwise} \end{cases} \tag{10}$$

where Z is the set of the ant unaccomplished tour and β is a system parameter. In addition, q and q_0 are random

number uniformly distributed in [0, 1] and in parameter ($0 \leqslant q_0 \leqslant 1$), which determines the relative importance of exploitation versus exploration, respectively. If $q \leqslant q_0$, the item of unaccomplished tour j with maximum $\tau_{iu}(t)\eta_{iu}^{\beta}$ value is put at position (exploitation); otherwise the item is chosen according to S (biased exploration).

2. The random variable S is selected according to the probability distribution of the random-proportional rule as following equation:

$$p_{ij}^{k} = \begin{cases} \frac{\tau_{ij}(t)\cdot\eta_{ij}^{\beta}}{\sum_{u\in Z}\tau_{iu}(t)\cdot\eta_{iu}^{\beta}} & \text{if } j \in Z \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

The resulting state transition rules refer to Eqs. (10) and (11), and are called the pseudo-random-proportional rule.

3. Ants can only choose a path that has never been used (increase tabu).
4. After traveling on a path, an ant will lay some pheromone on it (local updating).

Let $\tau_{ij}(t)$ be the intensity of pheromone trail on edge (i,j) at time t, therefore, we can consider an iteration to be when an ant completes the tour, and next iteration will be started at $(t+1)$. Then the pheromone intensity is updated according to

$$\tau_{ij}(t+1) = [(1-\alpha)\cdot\tau_{ij}(t)] + \alpha\cdot\Delta\tau_{ij} \tag{12}$$

where α ($0 \leqslant \alpha \leqslant 1$) is a coefficient for the remaining percentage of pheromone between time t to $t+1$. The association rules α is considered as a time series coefficient, which will decrease the impact levels of the old data and regulate coefficient $\Delta\tau_{ij}$. Therefore, if the data is irrelevant to time, set $\alpha = 0$.

For mining association rules, use Φ-correlation instead of correlation coefficient, as defined below:

$$\Delta\tau_{ij}^{k}(t) = \begin{cases} \frac{1}{L_{gb}}, & \text{if the } k\text{th ant in its tour is global-high-frequency} \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

L_{gb} is proposed as the highest frequency from first ant to kth ant accumulate value. It is also the pheromone intensity among the shortest path ij, while each ant completes its trip at time t.

Now let us summarize our algorithm as follows:

Step 1: Initialization

Set $t = 0$ {t is the time counter};
Set NC = 0 {NC is the iteration counter};
Set $\tau_{ij}(t) = c$ and $\Delta\tau_{ij} = 0$ and $\tau_0 = c$, $\forall\ i,j = 1,\ldots,n$; $i \neq j$;
Set $m = n$ Place the mth ant on the nth nodes (items);
Set $\beta = c$ and $q = c$ and $\rho = c$ and $\alpha = c \in [0,1]$; tabu $(s) = \emptyset$.

Step 2: Multi-dimensional constraints test

Scan the database once and find the complete set $[\mathrm{SAT}_c(\Phi)]$ of itemsets satisfying C.

Step 3: Mining guided by ant colony system

1. Calculate $\eta_{ij}(t) = \mathrm{support}_{ij}(t)$ from set $[\mathrm{SAT}_c(\Phi)]$.
2. Choose next item j by the state transition rule (Eqs. (10)).
3. If $q \geqslant q_0$, then choose the next edge ij until a given step is selected to move to with the transition probability as shown in Eqs. (11).
4. Move the kth ant from node i to the node j and insert that path into tabu (s).
5. Move the kth ant from node i to the node j and change the local pheromone trial by using

$$\tau_{ij}^{k} = (1-\rho)\cdot\tau_{ij} + \rho\cdot\Delta\tau_{ij} \tag{14}$$

where $0 < \rho < 1$ and $\triangle\tau_{ij} = \tau_0$.

6. After running a cycle, we update $\Delta\tau_{ij}^{k}(t)$ to follow Eqs. (13). Then, we calculate Eqs. (12).
7. Set $t = t+1$ and NC = NC + 1, and then repeat steps $1 \sim 5$ until the termination iteration met.
8. According to mining results generate the association rules.

The flows and steps of ASC-based association rule mining are shown in Fig. 9.

4. Model evaluation results and discussion

This section will apply the real world problem to evaluate the proposed method. The procedures and results are provided in the following subsections.

4.1. Data preparation and transformation

The National Health Insurance Plan of Taiwan Government has accumulated 12 million administrative and claims data. It is the largest database in the world. To rapidly and effectively respond to current and emerging health issues, The NHRI (National Health Research Institutes) cooperates with the National Health Insurance Bureau (NHIB) of Taiwan to establish a Nation Health Insurance research database. The NHRI is responsible of protecting the privacy and confidentiality of the data. She also routinely transfers the health insurance data from the NHIB to enable health researchers in order to analyze and improve the health of Taiwan's citizens.

The data used the systematic sampling method to randomly sample a representative database from the entire database. The size of the subset from each month is determined by the ratio of the amount of data in each month to that of the entire year. Then a systematic sampling is performed for each month to randomly choose a representative subset. This sampling database is obtained by combining

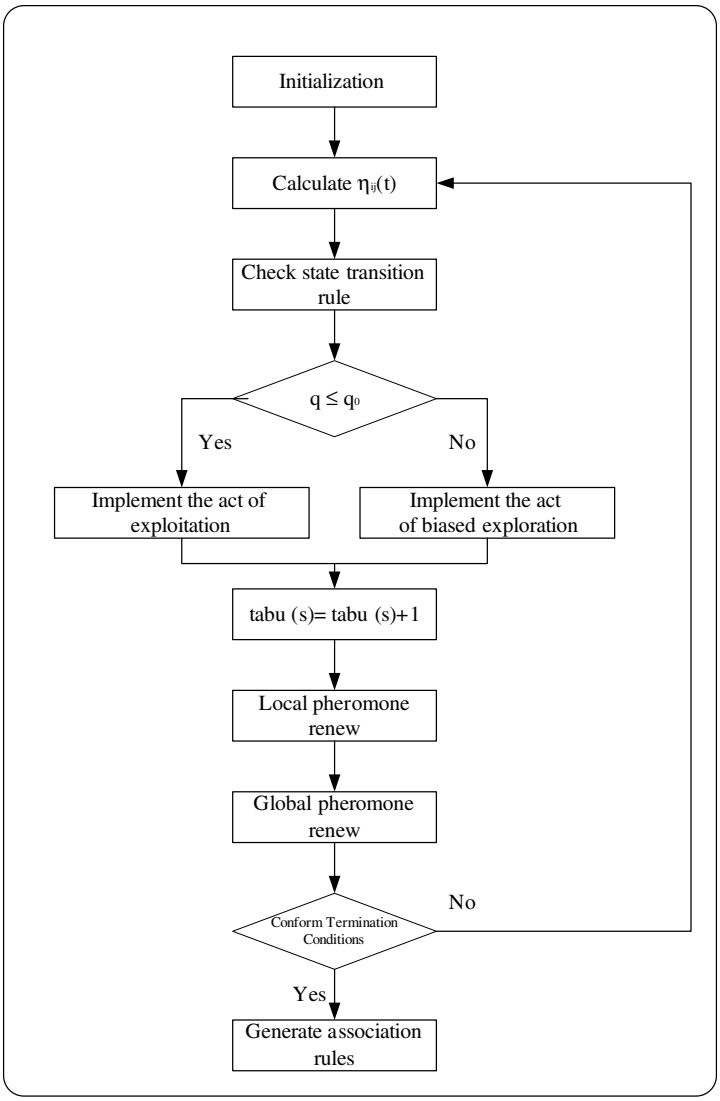

Fig. 9. The flowchart of ACS-based association rule mining algorithm.

the subsets for 12 months. The sampling database of the disease is around 0.2% to the entire database.

In a medical database, the most complete and detailed information are anamnesis data which contain disease name, prescription, patient's detail information, etc. Using this we aim to find the association rules between diseases; and also to detect the fake cases by data mining technology. This process should be able to increase medical quality as well as decrease the cost and the waste of medical resources.

In this study, the ACS-based association rule mining algorithm is employed to find some hidden relationships among disease items in the western medicine database. This study is based in part on data from the National Health Insurance Research Database provided by the Bureau of National Health Insurance, Department of Health and managed by National Health Research Institutes. The interpretation and conclusions contained herein do not represent those of Bureau of National Health Insurance, Department of Health or National Health Research Institutes. Because of the huge resources for executing the ASCA and AK, there are only 1000 data which is mined in the study.

In the data preparation stage, it is necessary to delete the data in disease column whose value is invalid, because this study is concerned with the disease relationships. There are thirty-seven columns in the original medicine database, but this study only concerned with the relationships of disease, so some columns in the database must be deleted. The remaining columns are "outpatient services," "outpatient services date," "patient's birthday," "international classification disease number 1–3 (ICD code)," and "patient's sex."

There are ICD-9-CM and A-Code in the international classification disease number that can be classified into 18 classifications as shown in Table 1 by anatomy and etiology. The 18th classification includes A-code which can not be classified by ICD-9-CM and supplementary classification which have V-code, E-code, and M-code.

Table 1
The classification of ICD code

No.	The disease items	The ICD code
1	Infectious and parasitic diseases	001–139
2	Neoplasm	140–239
3	Endocrine, nutritional, and metabolic diseases	240–279
4	Diseases of the blood-forming organs	280–289
5	Mental disorders	290–319
6	Diseases of the nervous systems and sense organs	320–389
7	Diseases of the circulatory system	390–459
8	Diseases of respiratory system	460–519
9	Diseases of the digestive system	520–579
10	Diseases of the genitourinary system	580–629
11	Complications of pregnancy, childbirth, and the puerperium	630–677
12	Diseases of the skin and subcutaneous tissue	680–709
13	Diseases of the musculoskeletal system and connective tissue	710–739
14	Congenital anomalies	740–759
15	Certain conditions originating in the perinatal period	760–779
16	Symptoms, signs, and ill-defined conditions	780–799
17	Injury and poisoning	800–999
18	Supplementary classification	Others

This study transformed the ICD code of National Health Insurance Research Database into 18 different attributes through Table 1. In order to avoid the data length may influence the attributes, which is the input of clustering analysis and then influence the relationship between data, the ICD code was encoded by Eq. (15).

$$D_k(i) = \frac{C \bullet S_k(i)}{C(i)} \tag{15}$$

where

$D_k(i)$ is the frequency which is relative to the maximum length of total data of the kth dimension (classification) in ith data.

$S_k(i)$ is the frequency of the kth dimension (classification) in ith data.

C is the maximum length of total data,and $C(i)$ is the length of ith data.

Table 2 is an example of decoding. Suppose $C = 3$, and Data ID is 1.

4.2. Clustering analysis

After encoding, clustering analysis was done with the two-stage clustering algorithm, which includes ASCA and AK. The Agglomerate level of ASCA was tuned, and generates different number of clusters shown in Fig. 10. The result, which generated 3 clusters, is chosen for lower cost of association rule mining. And then AK is used to modify the clusters. The curve of the TWCV is drawn in Fig. 11. and the TWCV is 3153.68 at last.

4.3. Association rule mining and analysis

Before mining the association rule from the result of clustering analysis, the effect of the number of data and items were compared. At first, the data sets that had 100, 200, 300, or 400 items were generated randomly. And then the association rules were mined with ACS-based association rule mining algorithm. The searching time and information of data sets were shown in Table 3, the Fig. 12(a) and (b).

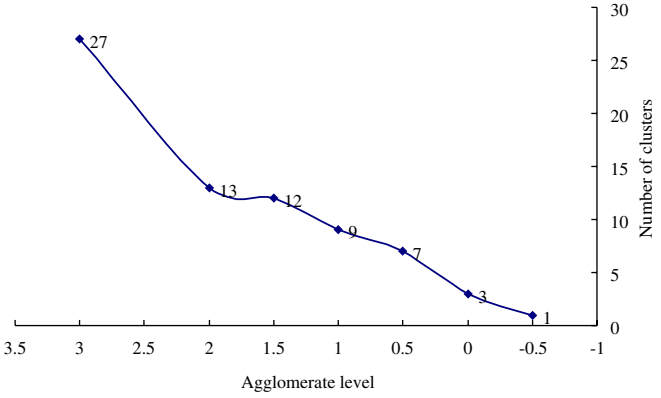

Fig. 10. The number of clusters with different agglomerate level.

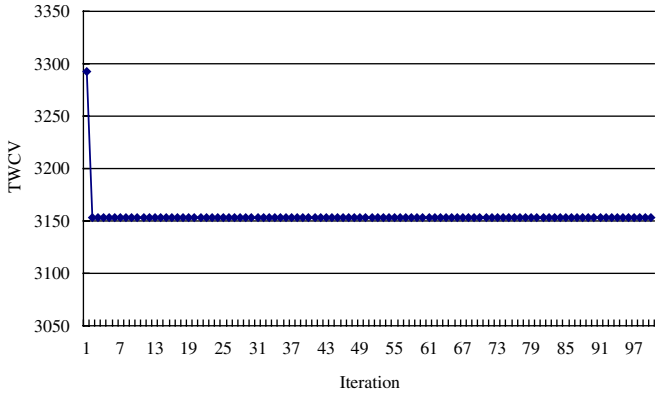

Fig. 11. The curve of the TWCV.

Table 2
An example of decoding

Input data	Data ID	ICD-9-CM						ICD-9-CM						ICD-9-CM					
	1	250.00						461.90						477.90					
Encoded	Data ID	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18
	1	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0

Table 3
Searching time and information of different data sets

Data Set	1	2	3	4	5	6
Number of data	62	111	116	397	305	167
Number of items	100	100	100	200	200	200
Searching time (s)	0.094	0.093	0.109	0.735	0.7	0.703
Data set	7	8	9	10	11	12
Number of data	364	689	870	597	1193	1614
Number of items	300	300	300	400	400	400
Searching time (s)	2.282	2.43	2.406	5.328	5.797	5.718

Fig. 12 shows that the amount of items has a great influence upon the efficiency of searching time, but it has less upon the amount of data. So, an inference about improving the performance of ACS-based association rule mining algorithm by reducing the amount of items could be drawn. The study that applied clustering analysis to reduce the amount of items could ameliorate the performance of ACS-based association rule mining algorithm.

The ACS-based association rule mining system was executed in Intel Pentium 4 3.0 G with 1024 MB ram. The searching time is shown in Table 4. The total searching time in all the clusters was 3.282 s., and it spent 3.875 s in mining the complete data. From mentioned above, the method, which mines association rule from different clusters, solves 15.31% time in mining complete data. But it spends more than 1 h, which is hugely long time, in clustering analysis.

Table 5 shows the association rules with top 5 pheromone, which were built up by ACS-based association rule mining system in all clusters, support and confidence of association rules. Compare clustered with non-clustered, and it can be found that the support and confidence of association rule in clustered are higher. Thus, the attributes of association rules are more similar between each other, and the rules are found easier. In this way, it is more helpful to study the pathology in some group of patients. From mentioned above, the proposed framework, which uses clustering analysis at first and then mines the association rule by ACS-based association rule mining algorithm, can not only improve the efficiency of performance but make the rules hidden in data easier to find.

Table 4
The searching time in different clusters

Cluster	1	2	3	Total
Number of data	49	808	143	1000
Number of items	68	507	162	642
Searching time (s)	0.157	2.75	0.375	3.875

From mentioned above, the proposed method can find the useful rule. For example, the rules such as "Essential hypertension, unspecified ==> Other and unspecified hyperlipidaemia" in cluster 2, has lower confidence, which is less than 10%, and may be overlooked easier before analysis. But the rules are found by the proposed method.

Table 6 is the result extracted from the complete data in ACS-based association mining system. The result shows that most rules are the same as the one extracted from the larger clusters, cluster 2. There are 808 data in the cluster, which is 80.8% of total data. The rules, which experts considered as the useful rules with robust relation, are extracted from not only the above-mentioned clusters but also the others. For example, "Trichiasis ==> Conjunctivitis, unspecified", which are extracted from cluster 3, is the rules with robust relationship. But it must set the threshold of pheromone as a lower value to generate a very large number of rules in the complete data. Thus many rules to examine the generated ones may be overlooked readily because of the big size of information. By contract, from the proposed method, it is easier to find out the hidden rules, which may occur less but have robust relationship. In other words, it can meet the same effect with lower cost.

Although the proposed method can find the useful and important rule and is executed in higher efficiency, the method also produces some useless rules and noise, in. For example, the rules, "Conjunctival xerosis ==> Chronic conjunctivitis, unspecified" in cluster 1, "Hypertrophy (benign) of prostate ==> Hypertensive heart disease, benign without congestive heart" and "Hypertrophy (benign) of prostate ==> Calculus of kidney" in cluster 3, go against the results of reaching in the past. The possible reasons are summed up as following:

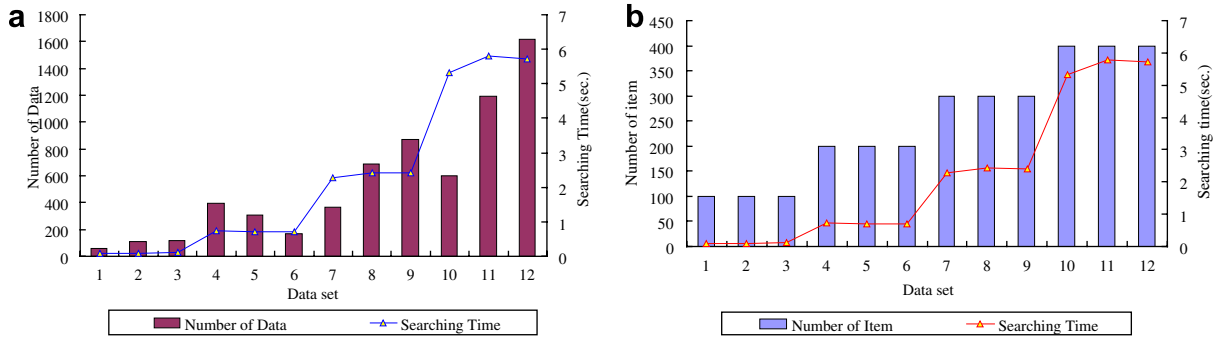

Fig. 12. Performance of ASC-based association rule mining on various amount data and amount itemsets. (a) Performance on various amounts of data. (b) Performance on various amounts of itemsets.

Table 5
The result of the proposed method

	Association rule	Clustered			Support	Confidence
		Pheromone	Support	Confidence		
1	Conjunctival xerosis ==> Chronic conjunctivitis, unspecified	4.016184	16.32653	62.50000	0.80000	62.50000
	Conjunctivitis, unspecified ==> Other specified disorders of eye and adnexa	3.011721	18.36735	33.33333	1.10000	27.27273
	Nonsenile cataract, unspecified ==> Conjunctivitis, unspecifie	1.996815	6.12245	66.66667	0.30000	66.66667
	Trichiasis ==> Conjunctivitis, unspecified	1.603509	8.16327	50.00000	0.40000	50.00000
	Other specified disorders of eye and adnexa ==> Age-related macular degeneration	1.012361	6.12245	33.33333	0.30000	33.33333
2	Acute upper respiretory infections of unspecified site ==> Essential hypertension, unspecified	8.640093	6.68317	16.66667	5.90000	15.25424
	Essential hypertension, unspecified ==> Other and unspecified hyperlipidaemia	8.002158	15.47030	8.00000	14.10000	7.09220
	Chronic ischemic heart disease, unspecified ==> Diabetes mellitus	7.202158	5.19802	21.42857	4.60000	19.56522
	Chronic ischemic heart disease, unspecified ==> Hypertensive heart disease, unspecified, without congestive	6.988804	5.19802	16.66667	4.60000	15.21739
	Headache ==> Dizziness and giddiness	5.999774	4.33168	17.14286	3.70000	16.21622
3	Diabetes mellitus (no complication) ==> Chronic renal failure	4.965272	11.18881	31.25000	11.70000	4.27350
	Urinary tract infection, site not specified ==> Vaginitis and vulvovaginitis, unspecified	4.00394	16.08392	17.39130	2.30000	17.39130
	Hypertrophy (benign) of prostate ==> Hypertensive heart disease, benign without congestive heart	3.973273	21.67832	12.90323	3.10000	12.90323
	Hypertrophy (benign) of prostate ==> Calculus of kidney	3.840274	21.67832	12.90323	3.10000	12.90323
	Diabetes mellitus(no complication) ==> Essential hypertension, unspecified	3.20137	11.18881	25.00000	11.70000	34.18803

Table 6
The result of the complete data

Association rules	Pheromone	Support	Confidence
Asthma, unspecified, without mention of status asthmaticus ==> Allergic rhinitis case unspecified	10.91202	2.7000%	40.7407%
Essential hypertension ==> Hyperlipidemia	9.998129	14.1000%	7.0922%
Allergic rhinitis case unspecified ==> Acute bronchitis	8.986932	4.3000%	20.9302%
Allergic rhinitis case unspecified ==> Asthma, unspecified, without mention of status asthmaticus	8.800364	4.3000%	25.5814%
Menopausal syndrome ==> Osteoporosis	7.998769	2.4000%	33.3333%
Upper respiratory infection ==> Essential hypertension	7.200365	5.9000%	15.2542%
Chronic hepatitis ==> Diabetes mellitus	6.400364	3.7000%	21.6216%
Headache ==> vertigo	5.999409	3.8000%	15.7895%
Other and unspecified hyperlipidaemia ==> Hypertensive heart disease, unspecified, without congestive	5.991731	4.0000%	15.0000%
Upper respiratory infection ==> Headache	5.952014	5.9000%	10.1695%
Asthma, unspecified, without mention of status asthmaticus ==> Acute sinusitis, unspecified	5.761401	2.7000%	22.2222%

1. The ICD codes are not good enough to classify the diseases and can not describe the relationship between the diseases.
2. The habits of Taiwanese to take medical treatment.
3. The error of observing the mapping network by vision.

5. Conclusions

In the early of 21st century, the developing of science and technology lets the medicine be prosperous and makes a huge change for the environments. Thus, it is quite difficult to predict what will happen in the further. Especially there are more and more previously unknown diseases, like SARS and bird flu, which were found recently. As mentioned above, human beings have to fight with the germs more and more hardly. Therefore, developing a decision support system which is about patient treatments and extracting the important relationships or association rules between diseases has become a very critical issue. This also can provide another way, which is different from the medicine and biology, to help diagnose the diseases for finding out the treatments.

According to the above findings, this study has developed a method which is able to discover more useful and accurate rules from the medical database fast. In order to avoid the missing knowledge in dividing the data, we divide the medical database into several clusters by ant colony system and then mine the hidden knowledge from the clustered data also via ant colony system. This can not only let the researchers pay more attention on some important groups and find out the hidden relation in the groups easier, but also avoid the important relationship ignored in the large database. The evaluation results using National Health Insurance Database have shown the proposed method's feasibility.

Although the result in this study shows the promising application, there are some issues that should be further

solved. Because this study just mines the relation between the ICD codes, it is suggested to add in the numerical data of medical examination and fuzzy the numerical data in preparation stage. In the clustering analysis stage, the proposed method utilized ASCA and AK to build up the cluster. Therefore, it may be desirable to apply other cluster method, like ART2, ADSOM or other two-stage methods to cluster the data. Besides, there many similar rules generated from the mining process, so it is feasible to apply other technology, such as the Fuzzy theorem, to merge the similar rules.

Acknowledgements

This study is partially supported by the National Science Council of Taiwan Government under Contract Number: NSC94-2416-H-027-001. Her support is appreciated.

References

Agrawal, R., Imielinski, T., & Swami, A. (1993a). Database mining: a performance perspective. *IEEE Transactions on Knowledge and Data Engineering, 5*(6), 914–925 (Special issue on Learning and Discovery in Knowledge-Based Databases).

Agrawal, R., Imielinski, T., & Swami, A. (1993b). Mining association rules between sets of items in large databases. In *Proc. ACM-SIGMOD int. conf. management of data (SIGMOD'93), May, Washington, USA* (pp. 207–216).

Agrawal, R., & Shafer, J. C. (1996). Parallel mining of association rules. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 962–969.

Bayardo, R. J. (1998). Efficiently mining long patterns from database. In *Proceedings of the ACM SIGMOD international conference on management of data, Washington, USA* (pp. 85–93).

Bellaachia, A., Portnoy, D., Chen, Y., & Elkahoun, A. G. (2002). E-CAST: a data mining algorithm for gene expression data. In *2nd workshop on data mining in bioinformatics, July* (pp. 49–54).

Berkhin, P. (2002). Survey of clustering data mining techniques. Accrue Software, Inc. Available from http://www.accrue.com/products/researchpapers.html.

Carpenter, G. A., & Grossberg, S. (1987). ART2: self-organization of stable category recognition codes for analog input pattern. *Applied Optics, 26*, 4919–4930.

Dorigo, M., & Gambardella, L. M. (1997). Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Transactions on Evolutionary Computation, 1*, 53–66.

Fayyad, U. (1997). Data mining and knowledge discovery in databases: implications for scientific databases. In *Scientific and statistical database management, 1997 proceedings, ninth international conference* (pp. 2–11).

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in database. *American Association for Artificial Intelligence*(August), 37–54.

Han, J., Pei, J., & Yin, Y. (2000). Mining frequent patterns without candidate generation. In: *Proceedings of the ACM SIGMOD international conference on management of data, Dallas, TX, USA* (pp.1–12).

Kohonen, T. (1991). Self-organizing maps: optimization approaches. In T. Kohonen, K. Makisara, O. Simula, & J. Kangas (Eds.), *Artificial neural networks* (pp. 981–990). Amsterdam, The Netherlands: Elsevier.

Koyuturk, Mehmet, Grama, Ananth, & Ramakrishnan, Naren (2005). Member, compression, clustering, and pattern discovery in very high-dimensional discrete-attribute data sets. *IEEE Transactions on Knowledge and Data Engineering, 17*(4), 447–461.

Krishna, K., & Murty, M. (1999). Genetic *K*-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 29*(3), 433–439.

Kuo, R.J., Cha, C.L., Chou, S.H., Shih, C.W., & Chiu, C.Y. (2003). Integration of ant algorithm and case based reasoning for knowledge management. In *Proceedings of International Conference on IJIE, November 10–12, 2003, Las Vegas, USA, in CD-R.*

Kuo, R. J., & Shih, C. W. (accepted for publication). Association rule mining through the ant colony system for national health insurance research database in Taiwan. *Computers and Mathematics with Applications*.

Kuo, R. J., Wang, H. S., Hu, T.-L., & Chou, S. H. (2005). Application of ant K-means on clustering analysis in data mining. *International Journal of Computers and Mathematics with Applications, 50*, 1709–1724.

Lin, D., & Kedem, Z. (2002). Pincer-search: an efficient algorithm for discovering the maximum frequent set. *IEEE Transactions on Knowledge and Data Engineering, 14*(3), 553–566.

Maulik, H., & Bandyopadhyay, S. (2000). Genetic algorithm-based clustering technique. *Pattern Recognition, 33*, 1455–1465.

Peacock Peter, R. (1998). Data mining in marketing: Part 1. *Marketing Management*, 9–18.

Savasere, A., Omiecinski, E., & Navathe, S. (1995). An efficient algorithm for mining associate rules in large databases. In *Proceedings of the international conference on very large data bases, Zurich, Switzerland* (pp. 432–444).

Shih, C. W. (2004). Applying ant colony system in data mining under multi-dimensional constraints. Master Thesis of National Taipei University of Technology, Taiwan, ROC.

Su, B. D. (2002). Discovering association rules through ant systems. Master Thesis of National Chin-Hwa Univeristy, Taiwan, ROC.

Su, M. C., & Chang, H. T. (2000). Fast self-organizing feature map algorithm. *IEEE Transactions on Neural Networks, 11*(3), 721–733.

Su, M. C., & Chang, H. T. (2001). A new model of self-organizing neural networks and its application in data projection. *IEEE Transactions on Neural Networks, 12*, 153–158.

Toivonen, H. (1996). Sampling large databases for association rules. In *Proceedings of the international conference on very large data bases, Mumbai (Bombay), India* (pp. 134–145).

Tsai, C. F., Wu, H. C., & Tsai, C. W. (2002). A new clustering approach for data mining in large databases. In *Proceedings of the international symposium on parallel architectures, algorithms and networks (ISPAN'02)* (pp. 1087–4089). IEEE Computer Society.

Tsay, Y. J., & Chang-Chien, Y. W. (2004). An efficient cluster and decomposition algorithm for mining association rules. *Information Sciences, 160*, 161–171.

Witten, I. H., & Frank, E. (2000). *Data mining: practical machine learning tools and techniques with java implementations*. Morgan Kaufmann Publishers.

Yang, X. B., Sun, J. G., & Huang, D. (2002). A new clustering method based on ant colony algorithm. In *Proceedings of the 4th world congress on intelligent control and automation, June* (pp. 2222–2226).

Zhang, X., & Li, Y. (1993). Self-organizing map as a new method for clustering and data analysis. In *Proc. IJCNN'93, int. joint conf. on neural networks* (pp. 2448–2451).

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering, 12*(3), 372–390.