Hindawi Publishing Corporation
Advances in Artificial Neural Systems
Volume 2012, Article ID 835903, 11 pages
doi:10.1155/2012/835903

Research Article

Hemodialysis Key Features Mining and Patients Clustering Technologies

Tzu-Chuen Lu and Chun-Ya Tseng

Department of Information Management, Chaoyang University of Technology, Wufeng District, Taichung 41349, Taiwan

Correspondence should be addressed to Tzu-Chuen Lu, tclu@cyut.edu.tw

Received 3 March 2012; Revised 4 June 2012; Accepted 8 June 2012

Academic Editor: Anke Meyer-Baese

The kidneys are very vital organs. Failing kidneys lose their ability to filter out waste products, resulting in kidney disease. To extend or save the lives of patients with impaired kidney function, kidney replacement is typically utilized, such as hemodialysis. This work uses an entropy function to identify key features related to hemodialysis. By identifying these key features, one can determine whether a patient requires hemodialysis. This work uses these key features as dimensions in cluster analysis. The key features can effectively determine whether a patient requires hemodialysis. The proposed data mining scheme finds association rules of each cluster. Hidden rules for causing any kidney disease can therefore be identified. The contributions and key points of this paper are as follows. (1) This paper finds some key features that can be used to predict the patient who may has high probability to perform hemodialysis. (2) The proposed scheme applies k-means clustering algorithm with the key features to category the patients. (3) A data mining technique is used to find the association rules from each cluster. (4) The mined rules can be used to determine whether a patient requires hemodialysis.

1. Introduction

The human kidney is located on the posterior abdominal wall on both sides of the spinal column. The main functions of the kidney include metabolism control, waste and toxin excretion, regulation of blood pressure, and maintaining the body's fluid balance. All blood in the body passes through the kidney 20 times per hour. When renal function is impaired, the body's waste cannot be metabolized, which can result in back pain, edema, uremia, high blood pressure, inflammation of the urethra, lethargy, insomnia, tinnitus, hair loss, blurred vision, slow reaction time, depression, fear, mental disorders, and other adverse consequences. Furthermore, an impaired kidney will produce and secrete erythropoietin. When secretion of red blood cells is insufficient, patients will have the anemia. The kidney also helps maintain the calcium and phosphate balance in blood, such that a patient with renal failure may develop bone lesions.

When renal function is abnormal, toxins can be produced, damaging organs and possibly leading to death. To extend or save the lives of patients with impaired kidney function, kidney replacement is typically utilized, including kidney transplantation, hemodialysis (HD), and peritoneal dialysis (PD). Although kidney transplantation is the most clinically effective method, few donor kidneys are available and transplantation can be limited by the physical conditions of patients. Notably, HD can extend the lives of kidney patients.

Although medical technology is mature, factors causing diseases are changing due to changing environments. Any factor may potentially lead to disease. When the detection index of a patient exceeds the standard and kidney disease has been diagnosed, patients must go the hospital for kidney replacement therapy. For instance, a doctor may recommend that high-risk patients adjust their habits by, say, stopping smoking, controlling blood pressure, maintaining normal

urination, controlling urinary protein levels, maintaining normal sleeping patterns, controlling blood sugar levels, reducing the use of medications, avoiding reductions in the body's resistance, maintaining low body fat levels, and reducing the burden on the kidneys.

However, improving one's physical condition and diet are insufficient. To control one's physical condition, periodic health examinations at a hospital have become a common disease-prevention strategy. Doctors may offer advice to patients based on health examination results to reduce disease risk.

Many scholars have applied data mining techniques for disease prediction. These techniques include clustering, association rules, and time-series analysis. Different analyses may require different mining techniques. Selection of an appropriate mining technique is the key to obtaining valuable data. However, choosing a data mining technique is very difficult for general hospitals, especially when dealing with different forms of original data. Therefore, to help medical professionals identify hidden factors that cause kidney diseases, this work applies a novel hemodialysis system (HD system). The HD system may identify factors not previously known.

General medical staff may perform routine examinations for particular factors associated with a particular disease and ignore other factors that may be associated with other diseases, such as kidney diseases. For example, staff may only assess blood urea nitrogen (BUN) and creatinine (CRE) levels and CRE clearance (CC). However, increasing amounts of data indicate that some hidden rules and relationships may exist. Therefore, this work uses an entropy function to identify key features related to HD. By identifying these key features, one can determine whether a patient requires HD. This work uses these key features as dimensions in cluster analysis. When patients requiring HD are classified into the same group, and the other patients are classified into the other group, the key features can effectively determine whether a patient requires HD. The proposed data mining scheme finds association rules of each cluster. Hidden rules for causing any kidney disease can therefore be identified.

2. Literature Review

2.1. Hemodialysis. Hemodialysis is also called dialysis. An artificial kidney discharges uremic toxins and water to eliminate uremic symptoms. In an HD system, a semipermeable membrane separates the blood and dialysate. The human blood continues passing through on one side of an artificial kidney and the dialysate carries away uremic toxins on the other side. Finally, the cleaned blood will back into the body. This continuous cycle eventually purifies blood.

A doctor may recommend that patient undergo dialysis according to the difference between acute and chronic. If kidney failure is acute, the doctor will recommend that the patient undergo dialysis before the occurrence of uremic toxins accumulate. For chronic kidney failure, medical treatment is first utilized and HD may be initiated after uremia occurs. Additionally, a doctor may assess according to the causes of kidney failure, kidney size, anemic state, degradation of kidney function, and recovery. Moreover, each examination indicator will be assessed. The most commonly used indicators are BUN concentration, CRE concentration, CC, urine-specific gravity, and osmotic pressure [1, 2].

2.1.1. Blood Urea Nitrogen (BUN). Blood urea nitrogen is the metabolite of proteins and amino acids excreted by the kidneys. The BUN concentration in blood can be used to determine whether kidney function is normal. The normal BUN range is 10–20 mg/dL. If the BUN concentration exceeds 20 mg/dL, this is called high azotemia. However, the BUN concentration may increase temporarily because of dehydration, eating large amounts of high-protein foods, upper gastrointestinal bleeding, severe liver disease, infection, steroid use, and impaired kidney blood flow. When the BUN concentration is high and the CRE concentration is normal, kidney function is normal. Although the BUN concentration can be used as an indicator of kidney function, it is not as accurate as the CRE concentration and CC.

2.1.2. Creatinine (CRE). Creatinine is mainly a metabolite of muscle activity and daily production is excreted through the kidneys. Daily CRE production cannot be fully excreted and the CRE concentration increases when TRY kidney function is impaired. As the CRE concentration increases, kidney function decreases. Because CRE is a waste generated by muscle metabolism, the CRE concentration is associated with the total amount of muscle or weight but is not related to diet or water intake. The CRE concentration may reflect kidney function more accurately than the BUN concentration. When the CRE concentration is in the normal range, it does mean that kidney function is normal; that is, CC is a better tool when assessing kidney function. The compensatory capacity of the kidney is large. For example, although the CRE concentration may increase from 1.4 mg/dL to 1.5 mg/dL, kidney function may have declined by more than 50%.

2.1.3. Creatinine Clearance (CC). Creatinine clearance is widely used and is an accurate estimation of kidney function. Creatinine Clearance is the amount of CRE cleared per minute. The CC for a healthy person is 80–120 mL/min; the average is 100 mL/min. Kidney failure is minor when the CC is 50–70 mL/min and moderate when CC is only 30–50 mL/min. If CC is <30 mL/min, kidney failure is severe and uremic symptoms will develop gradually. When CC is <10 gradually, a patient must start dialysis. By collecting all the urine produced within 24 hours, CC can be determined easily. Notably, CC is derived as follows:

$$\mathrm{CC} = \frac{\text{Urine CRE}_{\text{concentration}}\ (\text{mg\%}) \times 24\,\text{hours urine volume (c.c.)}}{\text{Blood CRE}_{\text{concentration}}\ (\text{mg\%}) \times 1440\,(\text{minutes})}. \quad (1)$$

2.1.4. Urine-Specific Gravity and Osmotic Pressure. Urine-specific gravity and osmotic pressure reflects the ability of the kidney to concentrate urine. If the specific gravity of urine is ≤1.018 or each urine-specific gravity gap is ≤0.008, the ability of the kidney to concentrate urine is impaired. Moreover, the ratio of osmolality to blood osmotic pressure must exceed 1.0; otherwise, the ability of the kidney to concentrate urine is impaired. If the ratio of urine to blood osmotic pressure is ≤3 after water fasting for 12 hours, the ability of the kidney to concentrate urine is impaired. Abnormal urine concentration function usually occurs in patients with analgesic nephropathy.

Doctors recommend patients undergo dialysis when their BUN concentration exceeds 90 mg/dL, the CRE concentration exceeds 9 mg/dL, and CC is <0.17 mL/sec, or the CRE concentration exceeds 707.2 mg/dL. However, when the BUN concentration begins increasing, the kidney is very fragile. That is, the kidney that has been damaged exceeds 1/3 when HD is required [3]. Thus, indexes such as the albumin globulin ratio (A/G ratio) of kidney function (Table 1), red blood cell (RBC) count in blood tests (Table 2), or white blood cell (WBC) count by urinalysis (Table 3) are related to kidney function [1]. This work proposes an effective scheme that identifies unknown key features to predict HD. This work uses the entropy function to identify key features that are strongly related to HD and applies the k-means clustering algorithm to these key features to group patients.

Hung proposed an association rule mining with multiple minimum supports for predicting hospitalization of HD patients [4]. Hung used this association rule to analyze factors that may lead to HD to reduce the number of patients hospitalized for kidney impairment.

Hung relied on routinely examined HD indexes for patients per month, including BUN, CRE, uric acid (UA), natrium (Na), potassium (K), calcium (Ca), phosphate (IP), and alkaline phosphatase levels and analyzed 667 derived variables, such as protein ratio, to determine whether monocytes infected or a patient was undernourished. Hung obtained 9 rules from 5,793 records. For instance, diabetic patients with high cholesterol levels were hospitalized most. Inadequate dialysis was a high risk factor for hospitalization. If patient is female, aged 40–49, infected with monocytes, and had a recent hemoglobin (Hb/Ht) test value that was too low, the frequency of hospitalization was high. If hematocrit (Ht) was abnormal twice in the last three months, average platelet volume (MPV) was abnormal twice, and total protein (TP) was abnormal once, the probability of hospitalization was 93%. If TP, glutamic oxaloacetic transaminase (GOT), and glutamic pyruvic transaminase (GPT) of patients were abnormal twice in the last three months and uric acid was also abnormal, hospitalization risk was 100%.

Huang analyzed risk of mortality for patients on long-term HD in 2009 [5]. Huang used the Classification and Regression Tree, Mann-Whitney *U* Test, Chi-square Test, Pearson Correlation, and the Nomogram to analyze 992 patients on long-term HD. Albumin level and age were the factors most strongly related to mortality. Huang clustered and analyzed patients. If a patient had good nutrition and was young, mortality of diabetic patients was 5.45 times that of nondiabetic patients. However, if a patient was malnourished and older, albumin and CRE levels were the factors most strongly related to mortality. Thus, albumin

Table 1: Kidney function test features.

Kidney function test items		Reference	Units
Blood urea nitrogen	BUN	5–25	mg/dL
Creatinine	CRE	0.3–1.4	mg/dL
Uric acid	UA	2.5–7.0	mg/dL
Albumin-globulin in ratio	A/G ratio	1.0–1.8	
Creatinine clearance/24 hrs urine	CC	M: 71–135 F: 78–116	mL/min
Renin	Penin	0.15–3.95	pg/mL/hr
Creatinine urine	Creatinine urine	60–250	mg/dL
Natrium	Na	135–145	meq/L
Potassium	K	3.4–4.5	meq/L
Calcium	Ca	8.4–10.6	mg/dL
Phosphorus	IP	2.1–4.7	mg/dL
Alkaline phosphatase	ALP	27–110	U/L

Table 2: Blood test features.

Blood test items		Reference	Units
Hemoglobin	Hb	M: 14–18 F: 12–16	g/dL
Red blood cell	RBC	M: 450–600 F: 400–550	mil/mm^3
White blood cell	WBC	5000–10000	mm^3
Hematocrit	Hct	M: 40–55 F: 37–50	%
Platelets	PLT	15–40.0	10^3/uL
Mean corpuscular volume	MCV	83–100	u^3
Mean corpuscular hemoglobin	MCH	27–32.5	uug
Mean corpuscular hemoglobin concentration	MCHC	32–36	%
Reticulocyte	Reticulocyte	0.5–2.0	%
Malaria	Malaria	(−)	
Erythrocyte sedimentation Rate.	ESR	M: 1–15 F: 1–20	mm/hr
Differential count	DC		
Band	Band	0–2	%
Neutrophils	Neutrophils	50–70	%
Lymphocytes	Lymphocytes	20–40	%
Monocytes	Monocytes	2–6	%
Eosinophils	Eosinophils	1–4	%
Basophils	Basophils	0–1	%
Bleeding times	BT	0–3	Minute
Coagulation times	CT	2–6	Minute
Blood type	Blood type		
Rhesus factor	Rh Factor	(+)	
Blood pressure	BP		mm/Hg
Height	Height		cm
Weight	Weight		kg

Table 3: Urine test features.

Urine test items		Reference	Units
Color/appearance	Color/appearance		
Reaction pH	Reaction PH	5.5–8.5	
Protein	Protein	<(+)	mg/mL
Sugar	Sugar	(−)	g/dL
Bilirubin	BIL	(−)	
Urobilinogen	URO	≤1; 4	umol/L
Urine red blood cells	RBC	0–3	/HPF
Urine white blood cells	WBC	0–5	/HPF
Pus cell	Pus cell	0-1	/HPF
Epith cell	Epith cell	M: 0–3 F: 0–15	/HPF
Casts	Casts	Not found	/LPF
Ketones	Ketones	(−)	mmol/L
Crystals	Crystals	− ~ (±)	/LPF
Bacteria and other	Bacteria and other	−	/HPF

level, age, diabetes status, and CRE level can help predict risk of mortality.

Yeh et al. used a data mining technique to predict hospitalization of HD patients in 2011 [6]. The availability of medical resources and dialysis quality may decline when too many patients are admitted to a hospital. Therefore, Yeh et al. used analysis of the C4.5 decision tree and the multiple minimum support (MS) association rule mining technology for analysis. The C4.5 decision tree was used to eliminate null values and association rule mining was used to identify hospitalization of HD patients. According to the records of hospitalized patients, hospitalized patients seldom have a chronic disease or may not have a chronic disease, but doctors only determine whether a patient should be hospitalized during an examination.

Lin used hospital records of patients combined with the association rule and the time-series analysis to establish a health-management information system for chronic diseases [7]. Lin found that occluded cerebral arteries may lead to cerebral thrombosis and a cerebral embolism. After examination by a doctor, the rule is effective in avoiding a second stroke. Additionally, ill-defined heart diseases still require improvement. Lin used data mining to provide the chronic disease patients' family members and medical staffs for controlling their disease.

These scholars usually used well-known blood tests as mining rules. This work uses an effective and novel scheme to identify some previously unknown features to predict HD. The entropy function is applied to identify features that are strongly related to HD, and the k-means clustering algorithm is applied with these key features to group patients.

2.2. Entropy Function. Information gain, proposed by Quinlan in 1979 [8], is a basis of the decision tree constructed by Interactive Dichotomiser 3 (ID3). Information gain can also be utilized to determine differences in feature attributes and other classification attributes. Further, it is usually used to select the split point of ID3.

We assume a classification problem that includes N data records, m feature dimensions, and k clusters. The measurement of a single feature's information gain must be determined based on two correlated values, called entropy; the difference between two correlated values is called information entropy

$$\text{Entropy}(N) = \sum P_t \times \log\left(\frac{1}{P_t}\right) = -\sum p_t \times \log(p_t), \quad (2)$$

$$\text{Entropy}\left(D_j\right) = \sum_{v=1}^{|D_j|} \frac{D_{jv}}{N} \times \text{Entropy}\left(D_{jv}\right), \quad (3)$$

$$\text{Gain}\left(D_j\right) = \text{Entropy}(N) - \text{Entropy}\left(D_j\right). \quad (4)$$

In (2), Entropy (N) is the total information content of whole problems, and this total information content is taken as a basis of single feature information gain, in which P_t is the probability of occurrence of t classification in N dataset.

In (3), Entropy (D_{jv}) is the information content of the j feature dimension, the v value, and classification and information quantity, D_{jv} is the j feature dimension, including v kinds of values, and the j feature dimension has $|D_j|$ values.

In (4), Gain (D_j) is a classification problem, the information gain received by the j feature dimension. Through (2)–(4), the information gain of each feature for a classification problem is found. This work then evaluates all threshold settings and collects the features with the greatest information gain to form a feature set for classification. Entropy is used to identify key features and cluster HD patients to determine the accuracy of key features.

2.3. Clustering Algorithm. Although many clustering techniques have been proposed, the k-means algorithm is the most representative and widely applied [9]. The k-means algorithm is also called the generalized Lloyd algorithm (GLA) [10]. The k-means algorithm transforms each data record into a data point and random numbers are utilized to generate the initial cluster center to determine which data point belongs to which cluster point. The divided data points are used to calculate the distance between a data point and the cluster center, such that a data point will belong to one cluster center when the data point is closer to one cluster center than another cluster center. The newly recomputed cluster center is the average among all data points in a cluster, and the new cluster center is taken as a basis for the next iteration. This process is repeated until no change occurs. The steps of the k-means algorithm are as follows.

(1) Use random numbers to generate the initial cluster centers $C_i = \{1, 2, \ldots, k\}$.

(2) Calculate the Euclidean distance $d(X, C_i)$ for each data point $X = \{x_1, x_2, \ldots, x_m\}$ and each cluster center C_i. The point with the shortest distance is classified in to C_i, and the distance formula is as follows:

$$d(X, C_i) = \sqrt{\sum_{j=1}^{m} \left(x_j - c_{ij}\right)^2}. \quad (5)$$

(3) Recompute the new cluster center C_i. If the movement of all data points in a cluster stop moving, all clustering work stops; otherwise, steps (1) and (2) are repeated for clustering.

2.4. Association Rule. An association rule is a widely used technique. It progressively scans a database to identify rules for the relationships between items. For instance, the probability that people will buy bread after buying milk is milk → bread (support = 50% and confidence = 100%); support means that the probability of a consumer buying both milk and bread is 50%, and confidence means that the probability of a consumer buying bread after buying milk is 100%.

Agrawal et al. developed the Apriori algorithm in 1994 [11]. The Apriori algorithm is one of the most popular data mining methods, where I is all itemsets, each data record is $X = \{x_1, x_2, \ldots, x_m\}$, and $X \subseteq I$. The expression of the association rule is $x_1 \rightarrow x_2$ (support, confidence), where $x_1 \subseteq I$, $x_2 \subseteq I$, and $x_1 \cap x_2 = \varphi$. Support and confidence affect mining results most. Support is the occupied percentage for N data records and the probability of occurrence of both x_1 and x_2 is $(x_1 \cup x_2)/N$. Confidence is the probability of x_1 and x_2 and is called a strong association rule.

First, set the threshold of minimum support and minimum confidence to generate frequently occurring items, where L_b represents frequently occurring b-itemsets, and all generated L_b frequent itemsets are combined to generate candidate itemsets. Only the support and confidence values that are greater than the minimum support and minimum confidence thresholds are retained. This process is repeated until all L_b frequent itemsets are identified.

3. Proposed Algorithms

This work applies a novel and effective scheme to find key features that predict HD. This work uses the entropy function to find the key features that are strongly related to HD and applies the k-means clustering algorithm with these key features to group patients. Furthermore, the proposed scheme applies the data mining technique to identify association rules from each cluster. These rules can be used to warn patients who may require HD. Figure 1 shows the system architecture, which is divided into four procedures.

These procedures are as follows.

(1) The input procedure, which should be handled very carefully, can determine the disease target and input various sources and formats into a database. This procedure has a marked impact on the subsequent procedure.

(2) The preprocess procedure is divided into two subprocedures. For quantitative processing, one subprocedure, data are converted into an appropriate analytical form; for example, a string form is converted into a numeric form, or a numeric form is converted into a similar spacing. For selecting features, the other subprocedure, this work uses the entropy function to find the key features that are strongly related to diseases.

(3) The mining procedure is also divided in two subprocedures. For clustering analysis, one subprocedure, the clustering algorithm is applied to these key features to group patients. For the association rule, the other subprocedure, the Apriori algorithm is applied to find the association rule in each cluster.

(4) The output procedure may express the entire mining result, and a medical professional will explain the mining result, and find any factor that may cause a disease.

3.1. Input Procedure. Examination information is from many sources, such as a hospital information system (HIS), laboratory information system (LIS), or Excel report. These different systems may have different data storage formats. For example, in the A database, gender is 1 for male and 2 for female, but in the B database, M is for male and F is for female. Thus, an error may occur while collecting data. Therefore, one should apply the preprocess process to ensure that information is correct, complete, and sufficient. The preprocess process is divided into five steps.

(1) Unified data storage format: to simplify mining, all information must be in the same format.

(2) Irrelevant data: if one does not specify the mining topic, mining efficiency and even accuracy will be adversely affected.

(3) Incorrect data: incorrect data may be caused by a source error or login error; thus, one should modify or remove.

(4) Formats do not match: to smooth information mining, information must be converted into an appropriate format when necessary.

(5) Incomplete data: incomplete data is a common problem; for example, some information may be lost, lacking for a certain period.

3.2. Preprocess Procedure. Data are standardized to improve analytical accuracy. A standard value may be applied to an item such as triglycerides (TG). If the TG level is ≥201 mg/dL, it exceeds and the standard is 100; if TG is normal it is in the range of 20–200 and the standard is 50; if TG is smaller than <19 mg/dL, it is lower than the standard and the standard is 0. If data are consecutive, a packing normalization method is used; its formula is as follows:

$$v'_j = \left\lfloor \left(\frac{v_j - \min_j}{\max_j - \min_j} \right) \times Q_j \right\rfloor, \quad (6)$$

where v_j represents raw data, $\min_j$ is the minimum value of j, $\max_j$ is the maximum value of j, v'_j is the packing normalized value, and Q_j is quantified distance. Table 4 shows example data after quantization.

Table 4 is a normalized form used to derive information gain and in association rule analysis, and it can effectively

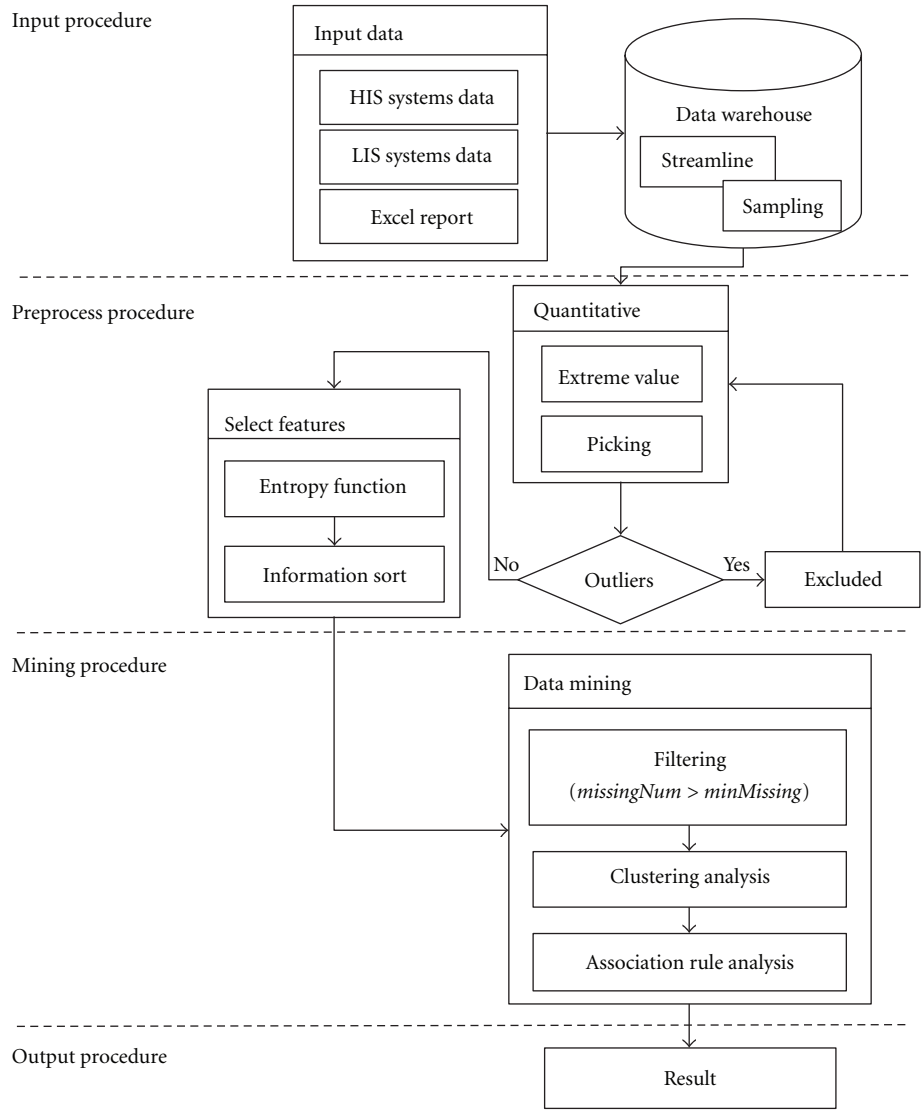


Figure 1: The system architecture.

differentiate between patients. This work simultaneously uses extreme value normalization; its formula is

$$v_j'' = \frac{v_j - \min_j}{\max_j - \min_j} \times 100, \tag{7}$$

where v_j represents raw data, $\min_j$ is the minimum value of j, $\max_j$ is the maximum value of j, and v_j'' is the packing normalized value. For instance, if the WBC value is 1, max = 10.7, and min = 3.5, then $v'' = [(1 - 3.5)/(10.7 - 3.5)] \times 100 = 38.89\%$ can be derived by applying (7).

In the entire database, the maximum and minimum values of each item markedly affect the quantification result, and the values are called outliers. If outliers exist, anomalies will also exist; for example, suppose that Q of CRE is 80, and CRE values are generally 0.37–2.99; however, a polarization datum may occur when a record is 6990. After quantization, values in the range of 0.37–2.99 will be quantified as 1, and the value recorded as 6990 will be assigned 80. Therefore, this work creates a mechanism to remove outliers. To avoid the influence of outlier values, this work sets a minNum threshold for each record. For example, assume minNum = 3 is the threshold. The total number of hemoglobin (HB), which is quantified as 2 (HB = 2), is 9; however, that of HB, which is quantified as 0 (HB = 0), is 1. This means that most data are assigned to HB = 2, and only 1 datum is assigned to HB = 0. The total number of quantified values that are smaller than minNum is the extreme value. This scheme replaces the extreme value with the average value.

3.3. Information Gain Analysis. This work uses dialysis item to identify information gain. For example, 6 patients are on dialysis (Dialysis = 1) (Table 4), the occurrence probability is $P_1 = 6/15$, and information gain is $P_1 \times \log(1/P_1) = (6/15) \times \log(6/15) = 0.528771$. When 9 patients are nondialysis (Dialysis = 0), occurrence probability is $P_0 = 9/15$, information gain is $P_0 \times \log(1/P_0) = (9/15) \times \log(9/15) = 0.442179$, and total information gain of P_0 and P_1 is 0.970951.

Table 4: Packing method normalized data.

Q_j	Sex	Age	WBC	RBC	HB	BUN	CRE	UA	GOT	GPT	TP	ALB	GLO	A/G	TG	Dialysis
	2	5	4	3	3	4	2	2	4	5	2	2	2	2	3	2
1	1	3	1	1	1	3	1	0	0	0	0	0	0	0	1	1
2	0	1	3	1	0	0	0	0	0	1	1	0	1	0	1	1
3	0	1	1	0	0	1	0	0	1	1	0	0	0	0	1	1
4	0	2	0	1	0	0	0	0	2	0	0	0	0	0	2	0
5	1	3	1	1	1	2	1	0	3	4	1	1	1	0	1	0
6	1	1	1	1	2	1	1	1	0	0	0	0	0	0	1	1
7	0	4	2	0	0	1	1	0	0	0	1	0	1	0	1	0
8	1	1	1	1	2	3	1	0	2	4	0	0	0	1	2	1
9	1	2	0	1	2	2	1	1	1	2	0	0	0	1	1	0
10	0	2	3	1	0	2	0	1	2	1	0	0	1	0	2	0
11	1	0	1	2	2	1	0	0	1	1	1	1	1	0	1	0
12	0	0	1	1	0	3	0	0	0	0	0	0	0	0	1	0
13	0	2	1	2	0	0	0	0	0	1	1	0	1	0	1	1
14	1	2	2	0	1	1	1	1	1	0	0	1	0	1	1	0
15	1	2	3	1	2	3	1	1	1	1	0	1	0	1	1	0

Table 5: Calculation information gain of sex relative to dialysis.

Sex j	Dialysis	Count (D_{jv})	$P_{D_{jv}}$	$P_{D_{jv}} \times \log(1/P_{D_{jv}})$	Entropy (D_{jv})	Entropy (D_j)
0	0	4	4/7	0.46	0.99	0.459773
	1	3	3/7	0.52		
1	0	5	5/8	0.42	0.95	0.509031
	1	3	3/8	0.53		
			Sum			0.968804

Next, this work calculates the information gain of each item relative to dialysis item. Take Sex (Table 5) as an example. The Sex of 7 women is 0 (Sex = 0) and only 4 records with non-dialysis (Dialysis = 0), the probability is $P_{D_{jv}} = 4/7$ of Sex = 0 and Dialysis = 0, and information gain is 0.46. Three records have Sex = 0 and Dialysis = 1; thus, the probability $P_{D_{jv}} = 3/7$, and information gain is 0.52. Total information gain of 0.46 and 0.52 is 0.99. Information gain of the women is $0.99 \times (7/16) = 0.459773$ because the probability of Sex = 0 is 7/16. After summing the information gain of the women (Sex = 0) and men (Sex = 1), total information gain is 0.968804, where $0.968804 = 0.459773 + 0.509031$. Next, via (3), which is Entropy (N) − Entropy (D_j), Gain $(D_j) = 0.970951 - 0.968804 = 0.002147$.

The information gain of each item related to dialysis can be obtained and ranked, and the association rule can be mined using the top few items as key features. Take Table 6 as an example. Assume that the top three items are chosen. Thus, Age, WBC, and BUN are taken as key features.

3.4. Data Mining Procedure

3.4.1. Missing Values. Some patients may have missing values. If their records are removed directly, some import information may be lost. Thus, this work applies a second filter before data mining analysis. This research sets minMissing as the threshold and takes missingNum as a null value of each record. If missingNum > minMissing, then the record is removed. Otherwise, missingNum $\leqq$ minMissing, the record will be retained and the missing null values will be replaced by the mean value. For instance, Age, WBC, and BUN are the top three key features when records are missing records. Assume minMissing is 1. When a record for which missingNum > 1, the record is removed; otherwise, the record is retained and the missing null values are replaced by the mean value.

3.4.2. Clustering. This work uses key features for clustering, where $x_1, x_2, \ldots, x_m$ as m key features, $X = \{x_1, x_2, \ldots, x_m\}$ are patient records, x_j is a key feature in X, $1 \leq j \leq m$, and k is the cluster number. The k-means process is as follows.

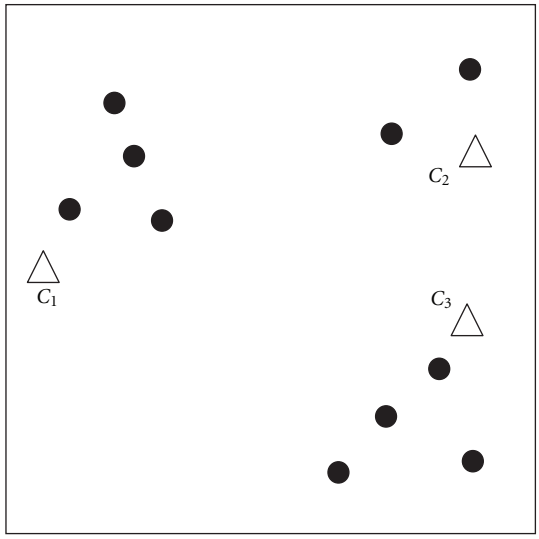
(1) First, randomly generate k initial cluster centers $C_i = \{c_1, c_2, \ldots, c_m\}$. Figure 2(a) has ten solid circles, $N = 10$, which are the locations of each record, and three triangles, $k = 3$, which are the locations of cluster centers C_i.

(2) Apply (5), $d(X, C_i) = \sqrt{\sum_{j=1}^{m} (x_j - c_{ij})^2}$, to calculate the distance between each patient's data point X and the cluster center C_i. When some X distance d_1 is less than d_i, X will be classified to C_1.

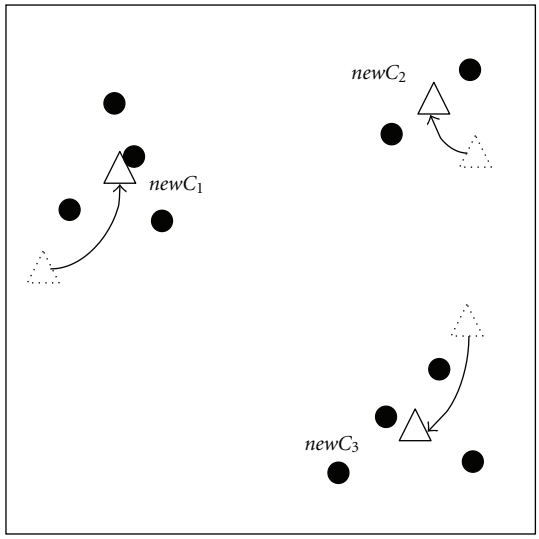
(3) Let $\widehat{C}_i = \{X_{c_i1}, X_{c_i2}, \ldots, X_{c_iS}\}$ be a cluster center membership, where S is the total number of members in C_i, and X_{c_iu} is u patient's data point in C_i. Thus,

Table 6: Information gain of each item.

Items	Sex	Age	WBC	RBC	HB	BUN	CRE	UA	GOT	GPT	TP	ALB	GLO	A/G	TG
Gain	0.002	0.577	0.329	0.14	0.06	0.28	0.05	0.09	0.18	0.2	0.05	0.24	0.02	0.06	0.03



(a) Initial dataset and cluster center (Before)



(b) Center displacement (After)

Figure 2: The diagram of clustering algorithm.

newC$_i$ will be added to the sum of X_{c_iS} in each $\widehat{C}_i$, and newC$_i$ = $\{(\sum_{u=1}^{s} x_{u1}/S), (\sum_{u=2}^{s} x_{u2}/S), \ldots, (\sum_{u=1}^{s} x_{uj}/S)\}$ can then be obtained. This function can also be taken as a new cluster center.

(4) Repeat steps (2) and (3) until each C_i remains the same.

3.4.3. Association Rule. Next, the proposed scheme finds each clustering characteristic rule using Apriori association rule analysis. We assume that the total number of records in cluster C_i is S, and each cluster membership is $\widehat{C}_i = \{X_{c_i1}, X_{c_i2}, \ldots, X_{c_iS}\}$; thus, the u patient's data point $X_{c,u}$ is in the $\widehat{C}_i$, and the j key features x_{uj} are in $X_{c_iu} = \{x_{u1}, x_{u2}, \ldots, x_{um}\}$. Next, the association rule is used to analyze each cluster C_i.

(1) First, set the values of minimum support minSup and minimum confidence minConf.

(2) Convert the normalization table into an extreme values table.

(3) Find the candidate set. We assume $\alpha = \text{item}_{jp}^{i}$, where item_{jp}^{i} is the p quantified value of the j key feature in $\widehat{C}_i$, $1 \leqq p \leqq Q$, and Sup(α) denotes the occurrence probability of item_{jp}^{i} in $\widehat{C}_i$. If Sup(α) $\geqq$ minSup, then item_{jp}^{i} becomes a candidate itemset L_Z and proceed to the next step.

(4) Through candidate set $L_z = \{\alpha_1, \alpha_2, \ldots, \alpha_y\}$, generate a set of two items, $\widehat{L}_y = \{\alpha_1 \cup \alpha_2, \alpha_1 \cup \alpha_3, \ldots, \alpha_1 \cup \alpha_y, \alpha_2 \cup \alpha_3, \ldots, \alpha_{y-1} \cup \alpha_y\}$; however, α_A and α_B cannot be the same item. Calculate the occurrence probability of each group, Sup($\alpha_A \cup \alpha_B$). If Sup($\alpha_A \cup \alpha_B$) > minSup, it becomes a member of frequent itemset L_{Z+1}.

(5) Take L_Z as a candidate set and repeat step (4) until the candidate set is null.

(6) Generate the association rule of the frequent itemset. If the confidence of the rule exceeds minConf, the rule is set up and the process is as follows.

(i) Let α^* be one of the frequent itemsets L_f, $\alpha^* = (R_A \cup R_B)$.

(ii) Generate rules $R_A \rightarrow R_B$ and $R_B \rightarrow R_A$.

In the case of A clustering C_A, where minSup = 2 and minConf = 0.5, the key features are item$_1$ = Age, item$_2$ = WBC, and item$_3$ = BUN, and $S = 7$ is the total number of records in C_A. Thus, this work finds the frequent itemsets L_1 using the minSup and minConf thresholds. The proposed scheme merges two items by L_1 as a candidate set, where j = Age, $p = 3$ in α_1, and j = WBC and $p = 1$ in α_3, and then calculates Sup($\alpha_1 \cup \alpha_3$). If Sup($\alpha_1 \cup \alpha_3$) $\geqq$ minSup, then let α_1 and α_3 be the two frequent itemsets until no more frequent itemsests are found.

Next, the quantified values are converted back into their original values if all rules are found; the formula is

$$v_j = \frac{v'_j}{Q_j} \times \left(\max_j - \min_j\right) + \min_j, \tag{8}$$

where v'_j is a quantified value, $\min_j$ is the minimum value of j, $\max_j$ is the maximum value of j, v_j is the original value,

and Q_j is a quantified interval. Take WBC = 1 → Age = 3 as an example rule. If the $\max_j$ of WBC is 10.7, the $\min_j$ value is 3.5, and Q_j is 4; then the original value of WBC = 1 is WBC $= 1/4 \times (10.7 - 3.5) + 3.5 = 5.3$. If the $\max_j$ value of Age is 68, the $\min_j$ value is 30, and Q_j is 5; then the original value of Age = 3 is Age $= 3/5 \times (68 - 30) + 30 = 52.8$. Through (8), the association rule WBC = 1 → Age = 3 can be transformed into WBC = 5.3 → Age = 52.8.

4. Experimental Results

This experiment uses health examination records provided by hospitals. The data are mainly for outpatient dialysis and general outpatients. The hospital has 105 records with many values missing. This is because each patient does not undergo all examinations. Therefore, data must first be filtered to eliminate records with missing values. This work adopts BUN and CRE, which are related to kidney function, as the first filter. If any null value occurs in BUN or CRE, the record is removed. In total, 18,166 records are retained after the first filtering.

The purpose of quantification in the preprocess procedure is to convert values into a continuity value or significant difference value from a finite interval. This work sets interval Q_j for each item based on recommendations by medical staff. Table 7 shows the intervals.

4.1. Choose Key Features. The mining result does not make sense when too many items are used. The proposed scheme uses the Entropy function to identify the top 4 key features between each item and dialysis; these features are are UA, AST (GOT), TG, and K (Blood).

4.2. Mining Procedure

4.2.1. Clustering Analysis. Based upon the above clustering algorithm, this work applies the k-means clustering algorithm with these key features to group patients. Before the experiment, records with many missing values were filtered out, leaving 7118 records. Table 8 shows the cluster grouping result. For example, 1169 patients are classified into the first group. The average indicator values are UA = 6.54, AST (GOT) = 24.48, TG = 119.79, and K (Blood) = 5.10, and the average density of the first group is 13.26. The average difference among all groups is 27.02, which is the best result of 100 random trial runs.

4.2.2. Association Rule Analysis. This work identifies the top four items related to dialysis as TG, AST (GOT), UA, and K (Blood); AST (GOT) is the main indicator of liver function. These four items are adopted as key features and the association rule technique is applied to analyze each group rule after clustering, where minSup = 35% and minConf = 65%. The association rules of the four clusters are shown in Table 9.

4.3. Summary. This work uses the clustering algorithm and the association rule algorithm to identify some previously unknown features of HD patients and possible association rules. This work then evaluates all threshold settings and collects the features with the greatest information gained to form a feature set for classification. Entropy is used to identify key features and cluster HD patients to determine the accuracy of key features. During the clustering process, the clustering algorithm is applied on these key features to group patients, and the entropy function can effectively determine clustering analysis with the key features. Furthermore, this work applies the apriori algorithm to find the association rules of each cluster. Hidden rules for causing any kidney disease can therefore be identified.

Table 7: Each interval of item.

ID	Item	Interval
1	TG	50
2	AST (GOT)	20
3	Ch	50
4	ALT (GPT)	20
5	UA	2
6	K (Boold)	2
7	BUN	5
8	Amylase (B)	50
...	...	...

Table 8: Clustering results.

Cluster	UA	AST (GOT)	TG	K (Blood)	Density
Cluster-1	6.54	24.48	119.72	5.10	14.14
Cluster-2	6.16	30.12	138.92	3.92	11.59
Cluster-3	4.47	24.72	112.33	4.07	11.22
Cluster-4	8.40	28.03	228.72	4.20	20.91

This experiment adopts the health examination records provided by one general hospital of Taiwan. During the experiment process, the experimental results will be discussed with medical staffs. From the experimental results, we can find that if BUN is in the range of 58.5–61.5 (60 ± 1.5) and Na (Blood) is in the range of 137.5–140.25 (140 ± 2.5), patients have a high risk of receiving a dialysis. The BUN is reported to be a reliable indicator of high risk, but the Na (Blood) is not clearly defined. Therefore, the Na (Blood) needs for further analysis and clarification. Conversely, if UA is in the range of 6.25–6.75 (6.5 ± 0.25), TG is in the range of 134.75–184.75 (159.75 ± 25), and K (Blood) is in the range of 3.89–4.39 (4.14 ± 0.25), or AC-GLU is in the range of 111–161 (136 ± 25), patients have a low risk of receiving a dialysis.

The medical staffs express that the UA, TG, and AC-GLU will definitely affect the possibility of patients to receive a dialysis, but K (Blood) is not clearly defined to create an influence on patients. The factor should be further analysis. At last, there is one more special feature, AST (GOT) because it appears both in the groups of high risk and low risk. The medical staffs express, actually AST (GOT) is not directly related to HD. Thus, AST (GOT) is not a key factor to determine whether a patient requires HD.

Table 9: Association rule of each cluster-k.

α^*	Sup (α^*)	Conf.
Cluster-1 ($k = 1$)		
BUN = 60 ± 1.5 → Dialysis = Yes	487	91%
Dialysis = Yes → AST (GOT) = 24.5 ± 10	708	74%
AST (GOT) = 24.5 ± 10 → Dialysis = Yes	523	73%
Na (Blood) =140 ± 2.5 → Dialysis = Yes	455	70%
Dialysis = Yes → BUN = 60 ± 1.5	487	69%
Na (Blood) = 140 ± 2.5 → AST (GOT) = 24.5 ± 10	434	66%
Cluster-2 ($k = 2$)		
CRE = 0.85 ± 0.15 → Dialysis = No	487	91%
UA = 6.5 ± 0.25 TG = 159.75 ± 25 → Dialysis = No	1341	97%
AC-GLU = 136 ± 25 → Dialysis = No	1265	94%
TG = 159.75 ± 25 → Dialysis = No	1696	93%
UA = 6.5 ± 0.25 → Dialysis = No	1920	93%
AST (GOT) = 45 ±10 → Dialysis = No	1479	92%
K (Boold) = 4.14 ± 0.25 → Dialysis = No	1938	91%
TG = 159.75 ± 25 Dialysis = No → UA = 6.5 ± 0.25	1341	79%
TG = 159.75 ± 25 → UA = 6.5 ± 0.25	1378	76%
TG =159.75 ± 25 → UA = 6.5 ± 0.25 Dialysis = No	1341	74%
UA = 6.5 ± 0.25 Dialysis = No → TG = 159.75 ± 25	1341	70%
UA = 6.5 ± 0.25 → TG = 159.75 ± 25	1378	67%
UA = 6.5 ± 0.25 → TG = 159.75 ± 25 Dialysis = No	1341	65%
CRE = 0.85 ± 0.15 →Dialysis = No	487	91%
UA = 6.5 ± 0.25 TG = 159.75 ± 25 → Dialysis = No	1341	97%
Cluster-3 ($k = 3$)		
CRE = 0.85 ± 0.15 → Dialysis = No	732	100%
CRE = 0.85 ± 0.15 K (Boold) = 5 ± 0.25 → Dialysis = No	560	100%
K (Boold) = 4.14 ± 0.25 → Dialysis = No	910	95%
AST (GOT) = 24.5 ± 10 K (Boold) = 4.14 ± 0.25 → Dialysis = No	507	94%
AST (GOT) = 24.5 ± 10 → Dialysis = No	505	92%
AST (GOT) = 24.5 ± 10 → Dialysis = No	679	86%
CRE = 0.85 ± 0.15 → K (Boold) = 4.14 ± 0.25	560	77%
CRE = 0.85 ± 0.15 Dialysis = No → K (Boold) = 4.14 ± 0.25	560	77%
CRE = 0.85 ± 0.15 → K (Boold) = 4.14 ± 0.25 Dialysis = No	560	77%
AST (GOT) = 24.5 ± 10 Dialysis = No → K (Boold) = 4.14 ± 0.25	507	75%
Dialysis = No → K (Boold) = 4.14 ± 0.25	910	74%
AST (GOT) = 24.5 ± 10 → K (Boold) = 4.14 ± 0.25	539	68%
Cluster-4 ($k = 4$)		
AST (GOT) = 45 ± 10 K (Boold) = 4.14 ± 0.25 → Dialysis = No	364	98%
K (Boold) = 4.14 ± 0.25 → Dialysis = No	503	91%
AST (GOT) = 45 ± 10 → Dialysis = No	537	90%
K (Boold) = 4.14 ± 0.25 Dialysis = No → AST (GOT) = 45 ± 10	364	72%
Dialysis = No → AST (GOT) = 45 ± 10	537	71%
AST (GOT) = 45 ± 10 Dialysis = No → K (Boold) = 4.14 ± 0.25	364	68%
K (Boold) = 4.14 ± 0.25 → AST (GOT) = 45 ± 10	372	68%
Dialysis = No → K (Boold) = 4.14 ± 0.25	503	67%
K (Boold) = 4.14 ± 0.25 → AST (GOT) = 45 ± 10 Dialysis = No	364	66%

5. Conclusion

Medical staffs try to find some information from patient's health examination records to reduce the occurrence of disease. However, some hidden information may be ignored because of the human observation or the restriction of book. Although there are many data mining techniques that have been proposed, most of them are focused on some known items. Seldom techniques in regard with searching for hidden key features are proposed. The reason is because the examination items are too many but incomplete. It is hard to find out the association rule by using system.

This research will help medical staffs to find some unknown key features to predict the hemodialysis. We apply k-means clustering algorithm with these key features to group the patients. Furthermore, the proposed scheme applies data mining technique to find the association rule from each cluster. The rules can help the patients to detect any occurrence possibility of disease.

Acknowledgment

The authors would like to thank the National Science Council of the Republic of China, Taiwan, for financially supporting this paper under Contract no. NSC 99-2622-E-324-006-CC3.

References

[1] DrKao, "Normal Test Values," 2010, http://www.drkao.com/1st_site/health_wap/normal_main.htm.

[2] Green Cross, "How to Detect Renal Function," 2010, http://www.greencross.org.tw/kidney/symptom_sign/kid_func.html.

[3] Shin Kong Wu Ho-Su Memorial Hospital, 2010, http://www.skh.org.tw/mnews/178/4-2.htm.

[4] K. C. Hung, *Multiple minimum support association rule mining for hospitalization prediction of hemodialysis patients [M.S. thesis]*, Computer Science and Information Engineering, 2004.

[5] S. Y. Huang, *The evaluation & analysis of the risk of mortality for patients receiving long-term hemodialysis proposal [M.S. thesis]*, Graduate Institute of Biomedical Informatics, 2009.

[6] J. Y. Yeh, T. H. Wu, and C. W. Tsao, "Using data mining techniques to predict hospitalization of hemodialysis patients," *Decision Support Systems*, vol. 50, no. 2, pp. 439–448, 2011.

[7] Y. J. Lin, *Applying data mining in health management information system for chronic desease [M.S. thesis]*, Department of Computer Science and Information Management, 2008.

[8] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[9] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithms: analysis and implementation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 881–892, 2002.

[10] J. Z. C. Lai, T. J. Huang, and Y. C. Liaw, "A fast k-means clustering algorithm using cluster center displacement," *Pattern Recognition*, vol. 42, no. 11, pp. 2551–2556, 2009.

[11] R. Agrawal, R. Srikant, H. Mannila et al., "Fast discovery of association rules," in *Advances in Knowledge Discovery and Data Mining*, pp. 307–328, 1996.