Empirical Evaluation of Ensemble Approach for Diagnostic Knowledge Management

Ha, Sung-Ho* · Zhang, Zhen-Yu**

Contents

I. Introduction	4.2 Associative Patterns among Lab Tests
II. Literature Review	4.3 Clinical Classification Knowledge for Chest Pain Diseases
2.1 Medical Data Mining	V. Validation and Comparison
2.2 Chest Pain Research	5.1 Classification Performance of the Ensemble Model
III. Research Methodology	5.2 Validation of Association Rules
3.1 Emergency Department Process	5.3 Justification of Using Association Rule Mining
3.2 Diagnosis Intelligence Using Data Mining	VI. Conclusions and Future Research
3.3 Feature Selection from Lab Tests	References
3.4 Ensemble Strategy for Clinical Classification	<Abstract>
IV. Data Analysis and Results	
4.1 Data Description and Preprocessing	

I. Introduction

The hospital's emergency department is a complex unit in which the fight between life and death is only a breath away. The emergency department has been frustrated by the problems of overcrowding, long waiting time, patient care delays, and high costs over decades. Accordingly,

to solve these problems has become the hottest issue in this area. Several internal or external factors have contributed to the long processing time and patient care delay: patient characteristics, emergency department staffing patterns, access to health care providers, patient arrival time, management practices, and testing and treatment strategies chosen (Fromm et al., 1993). Understanding these factors well is an important step to improve

* Corresponding author, Professor, School of Business, Kyungpook National University

** Ph.D student, School of Business, Kyungpook National University

the efficiency of patient care in an emergency department.

Most hospitals today have employed certain kinds of hospital information systems to manage their healthcare or patient data. These systems typically generate vast amounts of data in the forms of number, text, chart, and image. This raises an important question: “How can healthcare practitioners turn that data into useful information that would enable to make intelligent clinical decisions?” Considering the fast growth of data content, size, and diversity, researchers have focused on techniques to find useful information from collections of data during previous decades. Although its application to medical data analysis has been relatively limited until recently, the term ‘data mining’ has been increasingly used in the medical literature (Fayyad, 2002). The goal of predictive data mining in clinical medicine is to derive models that use specific patient information to support clinical decision-making. Data mining models can be applied to building of decision-making procedures such as prognosis, diagnosis, and treatment planning, which once evaluated and verified, could then be embedded within clinical information systems.

Accordingly, the purpose of this study is as follows: first, using data mining techniques, this study focuses on generating the association rules that help physicians to decide which lab tests they should select, which can reduce lab-testing time and cost in the emergency department. Second,

this study aims to build an ensemble of classifiers that supports to make a complex diagnosis, which can help physicians to formulate clinical decisions more quickly and more accurately. The organization of the paper is as follows: Section 2 explains medical data mining appeared in the literature and its application to the emergency department. Section 3 illustrates a hybrid decision model as a research methodology used in this study and section 4 applies the methodology to the real emergency data. Section 5 evaluates the methodology and compares it with other techniques. Section 6 provides conclusions and future directions.

II . Literature Review

1. Medical Data Mining

Medical data mining has been applied to accurate classification and rapid prediction for prognosis and diagnosis of patients in a specialized medical area (Cios and Moore, 2002). It has been also used for training unspecialized doctors to solve a specific diagnostic problem (Kononenko, 2001). Among several algorithms for classification and prediction tasks, a decision tree is one of the frequently used techniques in medical data mining area. While it is easy to find many cases to prove the decision tree to be useful in the business domain, the decision tree enables to predict prognoses and diagnoses in the domain

of medicine, by using tree-structured models or in the form of 'IF condition-based-on attribute-values THEN outcome-value' to identify useful features of importance.

Khan et al. (2008) used decision trees to extract clinical reasoning in the form of medical expert's actions that are inherent in a large number of electronic medical records. The extracted data could be used to teach students of oral medicine a number of orderly processes for dealing with patients with different problems, depending on the time. Yun (2008) utilized a C4.5 algorithm to build a decision tree in order to discover the critical causes of type II diabetes. She has learned knowledge about the illness regularity from diabetes data, and has generated a set of rules for diabetes diagnosis and prediction.

The Apriori algorithm was useful to figure out large item sets and thus to generate association rules from medical data. Abdullah et al. (2008) adopted an association algorithm to find the relationship between diagnosis and prescription. They stated that purchases and medical bills had much in common. Tan et al. (2009) used the Apriori algorithm to mine the rules for the compatibility of drugs from prescriptions to cure arrhythmia in the traditional Chinese medicine database. The experimental results showed that the drug compatibility obtained by the Apriori algorithm generally was consistent with the traditional Chinese medicine for that disease.

Emergency data is the data collected by emergency department environment. They are

more critical to human life than routine medical data. Most diseases included in the emergency data are fatal diseases. Ceglowski et al. (2007) discovered 'treatment pathways' through mining medical treatment procedures in the emergency department. They found that the workload in the emergency department varied depending on the number of presented patients, and was not affected by the type of procedure carried out.

Rossille et al. (2008) have presented a complementary perspective on the activities for specific patient groups by the emergency department: patients over 75 year old or less than 75 year old. She thought once validated, these views would be used as decision support tools for delivering better care to this population. Lin et al. (2010) found a way to raise the accuracy of triage through mining abnormal diagnostic practices in the triage. A two-stage cluster analysis (Ward's method, K-means) and a decision tree analysis were done on abnormal diagnoses in an emergency department.

Artificial neural networks have been applied to the fields of clinical diagnosis mostly to render the complex and fuzzy cognitive process of diagnosis. Many previous studies have shown the suitability of neural networks in the design of clinical decision support systems and biomedical applications. Ellenius and Groth (2000) assessed chest-pain patients and declared the limits of critically-sized systematic errors by calculating the decrease in diagnostic performance of neural networks. Guven and Kara (2006) concentrated

on the diagnosis of subnormal eye through the analysis of Electrooculography signals with the help of neural networks.

Support vector machines (SVM) are recently of increasing interest to biomedical researchers. It is not only because of the characteristics of well-founded theory, but also the superiority in practical applications. Conforti and Guido (2005) developed kernel-based SVM classifiers to aid the early diagnosis of acute myocardial infarction. By running a 10-fold cross validation procedure, the performance of their classifier was 97.5%. Majumder et al. (2005) made use of SVM for classification by integrating the recursive feature elimination.

2. Chest Pain Research

Chest pain is one of the most common reasons why people visit emergency rooms. Chest pain is particularly important because it may announce the existence of a serious and occasionally life-threatening disease. In addition, it can be complicated by the frequent disassociation between signal strength (symptoms) and seriousness of underlying pathology.

Considering the ratio of the outbreak of chest pain resulting from each body organ, it is easy to find that cardiac diseases are the most common cause of chest pain (45%), followed by musculoskeletal (14%), psychiatric (8%), gastrointestinal (6%), and pulmonary (5%) (Qiao, 2009). In fact, cardiac diseases are the third most

common diseases in Korea (22,347 people died in 2009) and they have maintained the status over the last ten years (Statistics Korea, 2010). Angina pectoris (AP) and acute myocardial infarction (AMI) are two types of acute illness, which can easily lead to death.

An accurate diagnosis and treatment of patients with chest pain is a difficult task of emergency physicians. A procedure to diagnose patients with acute chest pain should identify high-risk patients quickly for a fast cure (Erhardt et al., 2002). Bassan et al. (2000) evaluated the efficiency of a systematic diagnostic approach to chest pain patients in an emergency room in the relationship between diagnosis of acute coronary syndrome and hospitalization rates in the high-cost care units. Martinez-Selles et al. (2008) described the characteristics of patients with chest pain and evaluated the usefulness of the CPU-65 index for risk stratification of chest pain. The most frequent diagnosis for patients discharged from the emergency department was atypical chest pain and respiratory infection. Ridker et al. (2003) examined a group of healthy postmenopausal women over three years to assess the risk of cardiovascular events associated with base-line levels of markers of inflammation. Conroy et al. (2003) initiated a SCORE project to develop a risk-scoring system for use in the clinical management of cardiovascular risk in an European clinical practice.

III. Research Methodology

1. Emergency Department Process

There have been many studies that focus on redesigning and enhancing efficiency of the emergency department process. Based on these studies, a common process of current emergency department can be summarized as follow: a regular patient enters the emergency department, picks a number, and remains in the waiting area. When their number is called, the patient is assessed by a triage nurse who screens for apparent critical symptoms (high blood pressure, fever, and so on). If the patient is found to be in critical condition, they are transferred to the Intensive Care Unit for immediate care. Otherwise, the triage nurse assigns a triage code depending on the patient's condition (1 to 5, 1 being most critical).

Once a triage code is assigned, the patient waits for a physician's assessment. The waiting period depends on the availability of physicians and examination rooms. After the first assessment, lab tests may be required by the physician (blood test, lung scanner, and so on). If not, the patient is discharged to go home or transferred to another department. A patient with complete lab test results has to wait again for a second assessment by the physician requesting the tests. After the second assessment, the patient may be discharged to go home with a prescription or transferred for admission. Patients arriving by ambulance are

transferred directly to the trauma room without triage (Duguay, 2007).

2. Diagnosis Intelligence Using Data Mining

Based on the process in the emergency department stated above, the emphasis of this study focuses on the steps between 'Lab tests required' → 'Tests completed' → 'Waiting' → 'Physician assessment.' They are where the problems of congestion, long waiting time occur seriously over decades in the emergency department. If some unnecessary tests can be removed without any loss of diagnostic accuracy, the cost caused from congestion and long waiting time can be reduced.

Fig. 1 illustrates the research methodology, which consists of three stages, such as data collection, feature selection, and an ensemble strategy. The first stage of the methodology is collecting data in the form of electronic medical record (EMR) from an emergency department of a hospital. The EMR records include the information of patients, lab tests, and diagnosis.

Feature selection is the second stage of importance in this study. It reduces the dimension of EMR data in such a way that selected features can represent the most significant aspects of data. This study reduces the number of lab tests necessary to diagnose chest diseases accurately. In doing so, this study extracts association relationships between lab tests and diagnosis,

[Fig. 1] Research methodology

which are revealed by the Apriori algorithm in the form of association rules. The association rule here is an implication of the form $X \rightarrow Y$, which means 'If a patient takes a lab test X, then he will take a lab test Y.' The rule $X \rightarrow Y$ has to satisfy pre-specified minimum support and minimum confidence levels (Zaki, 2004).

Domain knowledge about diagnostic tests for the chest pain diseases helps select the lab tests associated with those diseases. Thus, the second stage includes such lab tests, which domain knowledge mentions they are important, from a set of lab tests chosen from the association rules that have recorded higher scores on support and confidence levels (e.g., above 0.9). For example, if X is one of the critical tests mentioned in the

domain knowledge, then, a test Y will be selected from the rules in the form of $X \rightarrow Y$ or $Y \rightarrow X$, whose support and confidence values are higher than 0.9.

In the third stage, by using the lab tests selected in the second stage and other information of patients (e.g., previous diagnosis and medical records), an ensemble strategy of classification algorithms is adopted to classify diagnosis of chest pain. Several different types of classification algorithms, including decision trees (DT), neural networks (NN), and support vector machines (SVM) are chosen to use (Shmueli et al., 2010). Specifically, DT algorithms involve C5.0 (Entropy index) and Classification and regression tree (Gini and Twoing indexes). NN

algorithms include Multilayer perceptron (MLP), Radial basis function network (RBFN), and Exhaustive prune. SVM algorithms involve RBF and Polynomial kernels.

After generating each single classifier, it needs to aggregate independent trained classifiers into DT ensembles, NN ensembles, and SVM ensembles through using an appropriate aggregation method, voting and confidence-weighted voting. Furthermore, the ensembles with higher performance can be collected to generate the final ensemble model. The one with the best performance will be the best final ensemble model.

3. Feature Selection from Lab Tests

The association algorithm finds the associative relationship between lab tests. Whenever a patient gets medical lab tests, associations occur between lab tests regarding diagnosis. The Apriori algorithm is chosen to analyze the associations and select the most critical features.

The information of lab tests and diagnosis is extracted from the EMR, and then the chest pain diseases involved in the data are clustered into several kinds, such as acute myocardial infarction, angina pectoris, and so on. The clustered lab test data according to each disease is analyzed by the Apriori algorithm that generates association relationships between lab tests. The frequent association rules that contain critical lab tests revealed by domain experts are selected

from the generated association rules, which form a knowledge base called ‘Lab tests association base.’

Association rules mining using Apriori is a two-step process, where in the first step frequent item-sets (in this case, lab tests) are discovered and in the second step, association rules are derived from the frequent item-sets (Kantardzic and Zurada, 2005). For the first step, the algorithm makes multiple scans over the data. In the first pass, the support of individual items is counted and frequent items are determined. In each subsequent pass, a set of item-sets found to be frequent in the previous pass is used for generating new frequent item-sets, called candidate item-sets, and their actual support is counted during the pass over the data. At the end of the pass, those satisfying minimum support constraint are collected, and they become the seed for the next pass. This process is repeated until no new frequent item-sets are found.

The second step is to generate the desired association rules from the frequent item-sets (lab tests). All subsets of every frequent item-set f are enumerated and for every such subset a , a rule of the form $a \rightarrow (f-a)$ is generated if the ratio of $\text{support}(f)$ to $\text{support}(a)$ is at least minimum confidence. Given a set of transactions $\{D = T_1, \dots, T_n\}$, and a set of items $\{I = I_1, \dots, I_m\}$, such that any transaction T in D is a set of items in I , an association rule is an implication $A \rightarrow B$ where the antecedent A and the consequent B are subsets of a transaction T in D , and A and B have no common

items. Statistical interestingness can be measured according to various criteria, which are related to the observed frequency of the rules. The support for a rule $A \rightarrow B$ is obtained by dividing the number of transactions, which satisfy the rule, $N_{A \rightarrow B}$, by the total number of transactions, N :

$$\text{Support}(A \rightarrow B) = \frac{N_{A \rightarrow B}}{N} \dots\dots\dots (1)$$

The confidence of the rule $A \rightarrow B$ is obtained by dividing the number of transactions, which satisfy the rule by the number of transactions, which contain the body of the rule, A :

$$\begin{aligned} \text{Confidence}(A \rightarrow B) &= \frac{N_{A \rightarrow B}}{N_A} \\ &= \text{support}(A \rightarrow B) / \text{support}(A) \dots\dots\dots (2) \end{aligned}$$

The lift takes the confidence of a rule and relates it to the support for the rule's consequent:

$$\text{Lift}(A \rightarrow B) = \text{confidence}(A \rightarrow B) / \text{support}(B)$$

$$\begin{aligned} &= \text{support}(A \rightarrow B) / \text{support}(A) \text{support}(B) \\ &\dots\dots\dots (3) \end{aligned}$$

A lift value greater than one indicates there is a positive association, whereas a value less than one indicates there is a negative association.

4. Ensemble Strategy for Clinical Classification

Fig. 2 illustrates the detailed process of an ensemble strategy for clinical classification and diagnosis. With the selected lab tests discovered by the 'Feature selection from lab tests' stage and other information (diagnosis, patient, and medical record data), an ensemble strategy with several different types of classification algorithms are adopted to use.

This study specifically selects several DTs, NNs, and SVMs as individual classifiers. C5.0 is

[Fig. 2] The procedure of generating ensembles of classifiers.

one of popular decision-tree learning tools (Quinlan, 1993). The distinctive characteristic of a C5.0 model is how the division rule split is chosen for the units belonging to a group, corresponding to a node of the tree. The criterion is expressed as the following Equation 4:

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} \times Entropy(S_i) \dots\dots\dots (4)$$

where n is the number of attributes, $|S_i|$ is the number of cases in the partition S_i , $|S|$ is the total number of cases in S . Entropy impurity is the usual choice as the following Equation 5:

$$I(m) = - \sum_{i=1}^{k(m)} \pi_i \log \pi_i \dots\dots\dots (5)$$

where π_i is the fitted probabilities of the levels present in node m , which are at most $k(m)$.

Classification and regression tree (CART) partitions data into two subsets so that the records within each subset are more homogeneous than in the previous subset (Giudici, 2003). There are three different impurity measures used to find splits for CART models, including Gini, Twoing, and Least-squared deviation. For symbolic targets, Gini or twoing can be the choice. The Gini index, $g(t)$, at a node t in a CART tree is defined as:

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t) \dots\dots\dots (6)$$

where i and j are categories of the target field. The

Towing index, $\Phi(s, t)$, for a split s at a node t is defined as:

$$\Phi(s, t) = [\sum_j |p(j|t_L) - p(j|t_R)|]^2 \dots\dots\dots (7)$$

where t_L and t_R are the nodes generated by the split.

MLP is a hierarchical structure of several perceptrons and overcomes shortcomings of the single-layer neural networks (Mitchell, 1999). The training algorithm for MLP requires differentiable, continuous nonlinear activation functions:

$$o = \frac{1}{(1 + e^{-s})} \dots\dots\dots (8)$$

where s is the sum of products of weights and inputs.

A RBF neural network has an input layer, a hidden layer, and an output layer. Neurons in the hidden layer contain Gaussian transfer functions whose outputs are inversely proportional to the distance from the center of neurons. The Exhaustive prune method is related with a prune method. It starts with a large network and prunes the weakest units in the hidden and input layers as training proceeds.

SVM generates input-output mapping functions from a set of labeled training data (Kantardzic and Zurada, 2005). For classification, nonlinear kernel functions are often used to transform input data to a high-dimensional feature space. The RBF kernel,

such as an exponential (Gaussian) kernel uses the Euclidean distance between vectors $\mathbf{x}$ and $\mathbf{y}$:

$$K_{RB}(\vec{x}, \vec{y}) = k(\|\vec{x} - \vec{y}\|_2^2) = \exp(-\frac{1}{2\sigma^2} \|\vec{x} - \vec{y}\|_2^2) \dots\dots\dots (9)$$

where $k: R_0^+ \rightarrow R$ is a kernel profile. For a Polynomial kernel,

$$K_{Pl}(\vec{x}, \vec{y}) = (\gamma \vec{x} \cdot \vec{y} + r)^d, \gamma > 0 \dots\dots\dots (10)$$

The ensemble approach uses ‘Simple voting’ and ‘Confidence-weighted voting.’ With simple voting, if two out of three models predict ‘yes’, then ‘yes’ wins by a vote of 2 to 1. In the case of confidence-weighted voting, votes are weighted based on the confidence value for each prediction. Thus, if one model predicts ‘no’ with a higher confidence than the two ‘yes’ predictions combined, then ‘no’ wins.

IV. Data Analysis and Results

1. Data Description and Preprocessing

The EMR dataset included eleven attributes available from the database of the emergency department, which were patient number (PNO), diagnosis, date of arrival, time of arrival, age, gender, symptom, ocode, subcode, oname, and testing results. To protect privacy of the patient, an artificial PNO was assigned uniquely to the

individual patient. ‘Ocode’ is a high-level classification code and ‘subcode’ is a low-level classification code that represents each lab test. To facilitate analysis, the data were sorted in ascending order of PNO by using EXCEL 2007 and rearranged to put the same patient’s information in the same row.

The majority of patients were men. In terms of the patient age, more than 56.25% of patients exceeded 60 years. When chest pain diseases were clustered into three groups, such as AMI, AP, and others, the sample data showed that AMI was the most typical disease (45.82%). As previous studies have pointed out, the AMI disease has been a threat to human life with a higher risk. There were 106 patients with the AP (22.18%) disease that has been the second critical disease and there were 153 patients with other chronic diseases (32.01%) that include Cardiomyopathy, Pneumothorax, and Respiratory tuberculosis, for example.

Because the majority of patients were the old, this study focused on them in particular. Out of 478 patients, 180 patients who were older than 60 years old were selected for the further analysis.

2. Associative Patterns among Lab Tests

As mentioned above, there were a series of 410 lab tests conducted in the focal emergency room. Not all patients, however, need to undergo all the lab tests since they arrive. The patient’s condition could get worse while waiting to be tested. If a

[Table 1] A sample of co-occurrence patterns between lab tests at the sub-code level.

Confidence	Support	Lift	Association rules at the sub-code level
One antecedent -> One descendent			
100	99.39	1	J503942_01 -> J252640_01
Two antecedents -> One descendent			
99.87	94.18	1.01	J252630_01 and J252640_01 -> J25283A_13
Three antecedents -> One descendent			
95.24	91.81	1.27	J252630_01 and J252640_01 and J503942_01 -> J151530_01

patient gets testing unnecessary for diagnosis, it could be a waste of time and money. In order to reduce the number of lab tests and to reduce time spent on tests, the most critical lab tests need to be revealed to diagnose the disease. In doing so, the domain knowledge in the field of medicine was combined to discover important lab tests.

According to Kenneth and Sharon (2006), and Ren (2008), the AP and AMI diseases belong to the Acute Coronary Syndrome (ACS). Creatine Kinase (CK), Creatine Kinase MB fraction (CK-MB), and Troponin are very sensitive and critical tests to diagnose the ACS diseases. Based on their findings, association rules which contained each of CK (sub-code J252630_01), CK-MB (sub-code J252640_01), and Troponin (sub-code J503942_01) were discovered by the Apriori algorithm. In the process of analyzing the relationship between lab tests, sub-codes indicating specific lab tests were set as both input and target variables. Both minimum support and minimum confidence values were set at 0.9 to discover the rules showing a higher frequency.

There were 4,650 rules that met the threshold (support=0.9 and confidence=0.9). As a result of using the domain knowledge for the diagnosis of ACS, the number of rules which included three crucial lab tests (CK, CK-MB, and Troponin) has been reduced to 3,585 from 4,650. Table 1 shows a sample of association rules at the sub-code level. Consider the relationship between lab tests J503942_01 and J252640_01, for example. The support value for the rule, which means 'If a patient takes a lab test J503942_01, then he also takes a test J252640_01,' was calculated as follows:

$$\text{Support}(J503942_01 \rightarrow J252640_01) = \frac{325}{327} = 0.9939$$

The confidence of the rule depended on its consequent and antecedent:

$$\text{Confidence}(J503942_01 \rightarrow J252640_01) = \frac{325}{326} = 0.9969$$

From these association rules, it was found that 56 lab tests had a strong association with the critical lab tests such as CK, CK-MB, and Troponin.

3. Clinical Classification Knowledge for Chest Pain Diseases

For the building of diagnosis intelligence by using an ensemble approach, 180 patients' records were split into two kinds of data sets at random: a training dataset (126 records, 70%) and a validation dataset (54 records, 30%). Input variables included the selected 59 lab tests mentioned above and demographic information (i.e., gender and age) of each patient (Kannel and Vasan, 2009). The target variable included three types of chest pain diseases: AP, AMI, and other chest pain diseases.

Table 2 shows the classification accuracy of each classifier and of each ensemble obtained on the training set of chest pain data. In the case of single classifier classification, the SVM (Polynomial) algorithm recorded the highest accuracy on the training data (95.42%). Among the ensembles, the SVM ensemble showed the best accuracy (96.18%), regardless of voting

methods. The accuracy performance of both final ensembles was quite well for the training data set and all of them scored above 95% accuracy.

[Table 3] Confusion matrix for the best final ensemble (training data set)

		Classified diseases		
		Others	AP	AMI
Actual diseases	Others	42	0	1
	AP	0	41	2
	AMI	0	0	40

Especially for the final ensemble, a simple voting achieved the best classification accuracy (97.62%), which rose 2.2% compared to the best single classifier, SVM(Polynomial).

Table 3 summarizes the confusion matrix for the best final ensemble obtained on the training data set. Each row represented the actual diseases while each column represented the classified diseases in the confusion matrix. The diagonal values showed correct classifications.

The classifications had an error rate of

[Table 2] Classification accuracy of single and ensemble classifiers (training data).

Type of classifier	Aggregation method	Classifier	Classification accuracy (%)
Single classifier		SVM (RBF)	94.66
		SVM (Polynomial)	95.42
Ensemble of classifiers	Voting	SVM ensemble	96.18
		NN ensemble	67.18
	Confidence-weighted voting	SVM ensemble	95.42
		NN ensemble	76.34
Ensemble of classifiers	Voting	Final ensemble	97.62
	Confidence-weighted voting	Final ensemble	96.95

[Table 4] Classification accuracy of single and ensemble classifiers (validation data)

Type of classifier	Aggregation method	Classifier	Classification accuracy (%)
Single classifier		SVM (RBF)	87.50
		SVM (Polynomial)	83.93
Ensemble of classifiers	Voting	SVM ensemble	85.71
		NN ensemble	76.57
	Confidence-weighted voting	SVM ensemble	82.14
		NN ensemble	80.36
Ensemble of classifiers	Voting	Final ensemble	92.59
	Confidence-weighted voting	Final ensemble	91.07

$100 \times (1+2)/126 = 2.38\%$. The classification accuracy was $100\% (=40/40)$ for the AMI disease, $95.35\% (=41/43)$ for the AP disease, and $97.67\% (=42/43)$ for other chest pain diseases. Total accuracy reached 97.62%.

V. Validation and Comparison

[Table 5] Confusion matrix for the best final ensemble obtained (validation data set)

		Classified diseases		
		Others	AP	AMI
Actual diseases	Others	15	0	2
	AP	0	17	0
	AMI	1	1	18

[Fig. 3] Gain charts for the considered models obtained on the validation data.

1. Classification Performance of the Ensemble Model

For the validation of the best final ensemble model, the records of 54 patients with chest pain were prepared separately from the training data set. Table 4 lists the classification accuracy of each single classifier and of each ensemble generated on the validation set of chest pain data.

Among the single classifier models, the SVM(RBF) model showed the best performance on the validation data set (87.50%). For the voting ensemble, both SVM and DT ensembles had the best classification accuracy (85.71%). As shown in the training data set, the final ensemble model generated by voting exhibited the best accuracy (92.59%). Table 5 shows the confusion matrix for the best final ensemble obtained on the validation data set.

The classification accuracy was $90\% (=18/20)$ for the AMI disease, $100\% (=17/17)$ for the AP disease, and $88.24\% (=15/17)$ for other chest pain diseases. Total accuracy reached 92.59%. Fig. 3 shows the gain charts for single classifier models and for the best final ensemble model being

compared together. The gain chart of the best final ensemble model suggests that it has better performance.

2. Validation of Association Rules

The 59 important lab tests have been identified by using association rules based on domain knowledge. The support and confidence of all the association rules were above 0.9. The derived lab tests were fed into the classifiers as input variables. This subsection examined the impact of altering of support values on changes in the classification performance, while the confidence value remained to be fixed at 0.9. detecting chest pain diseases (AMI, AP, and others), as the value of support gradually decreased to 0.2 from 0.9 (refer to Fig. 4). When the value of support was set as 0.9, the classification accuracy was best (92.59%). As the value of support was getting lower, the accuracy performance of the best final ensemble model for decreased until it reached 71.69%, when the support value was 20%. The stronger associations between lab tests, the higher accuracy the ensemble model achieves.

[Fig. 4] Box plot as changes in the support value.

3. Justification of Using Association Rule Mining

[Table 6] Confusion matrix for the comparative ensemble model (three input variables).

		Classified diseases		
		Others	AP	AMI
Actual diseases	Others	16	1	0
	AP	2	14	1
	AMI	1	6	13

From the studies conducted by Kenneth and Sharon (2006), and Ren (2008), it is known that the three types of tests (i.e., CK, CK-MB, and Troponin) are critical to diagnose chest pain diseases, especially for the acute coronary syndrome. Association rule mining discovered additional 56 lab tests that were strongly associated with the three critical tests. By using all of these lab tests, the ensemble model then achieved 92.59% of the classification accuracy. Thus, this subsection evaluated the design choice of using association rule mining before carrying out the disease classification. To do so, a comparative ensemble model was built by using the three lab tests (CK, CK-MB, and Troponin) only as input variables. The same training data set and validation data set were used for this comparison. The comparative ensemble model adopted the same learning parameters.

Table 6 summarizes the confusion matrix for the comparative ensemble model obtained on the validation data set. Classification accuracy reached 79.63%. It was found that the accuracy

was lower than that of the proposed method, combining association rule mining. The classifications had an error rate of $100 \times (1+2+1+1+6)/54 = 20.37\%$. The classification accuracy was $65\% (=13/20)$ for the AMI disease, $82.35\% (=14/17)$ for the AP disease, and $94.12\% (=16/17)$ for other chest pain diseases.

VI. Conclusions and Future Research

The purpose of this study was to build a hybrid model to derive diagnosis intelligence to aid in clinical decision making in a hospital environment. This study used an Apriori algorithm and an ensemble approach to combine different classification algorithms, and analyzed actual data collected from a hospital to classify chest pain diseases, which can help physicians to make clinical decisions faster and more accurately.

Through evaluation, this study showed that the final ensemble model performed well (classification accuracy of 92.59%) in diagnosis of chest pain diseases. In order to show the advantage of combining the association rule mining and the ensemble model, a further analysis was conducted by using the three critical lab tests only as input variables. The classification accuracy of using 59 lab tests as input variables had better accuracy. By integrating medical domain knowledge into the research, unnecessary

lab tests could be filtered out so that this has brought higher accuracy to the classification of diseases and savings on time and money spent on the processing of lab tests.

Future research should consider several improvements. First, this study adopted a hybrid decision-model approach, combining the association rule mining and the ensemble of classifiers. However, another combination of data mining techniques is still possible to accomplish the same task of medical data mining. To think of an alternative methodology will be a good challenge. Second, as the accumulation and diversification of medical data increase, the need for adaptive methods of medical decision-making has been gradually expanding. Considering the current situation of hospital information systems, more complete and comprehensive systems will be necessary for the future. Such a system should have medical intelligence mentioned in this study, which can be enhanced and expanded. For example, it can incorporate other medical attributes such as image or audio, which can help to raise the accuracy of classification and prediction of diseases. Third, the data used in this study were collected from a single hospital in a city during one year. Due to the geographical and temporal limitation, the data may not be typical for all chest pain patients. Even though the distribution of the sample data obtained was consistent with that of patients with chest pain in the country, those efforts are needed to generalize the classification model and methodology.

References

- Abdullah, U., Ahmad, J., and Ahmed, A., "Analysis of Effectiveness of Apriori Algorithm in Medical Billing Data Mining," *Proceedings of 4th International Conference on Emerging Technologies*, 2008, pp.327-331.
- Bassan, R., Scofano, M., Gamarski, R., Dohmann, H.F., Pimenta, L., Volschan, A., Araujo, M., Clare, C., Fabricio, M., Sanmartin, C.H., Mohallem, K., Gaspar, S., and Macaciel, R., "Chest Pain in the Emergency Room: Importance of a Systematic Approach," *Arquivos Brasileiros de Cardiologia*, Vol.74, No.1, 2000, pp.13-29.
- Ceglowski, R., Churilov, L., and Wasserthiel, J., "Combining Data Mining and Discrete Event Simulation for a Value-Added View of a Hospital Emergency Department," *Journal of the Operational Research Society*, Vol.58, 2007, pp.246-254.
- Cios, K. J., and Moore, G.W., "Uniqueness of Medical Data Mining," *Artificial Intelligence in Medicine*, Vol.26, No.1, 2002, pp.1-24.
- Conforti, D., and Guido, R., "Kernel-Based Support Vector Machine Classifiers for Early Detection of Myocardial Infarction," *Optimization Methods and Software*, Vol.20, No.2-3, 2005, pp.401-413.
- Conroy, R.M., Pyorala, K., Fitzgerald, A. P., Sans, S., Menotti, A., De Backer, G., De Bacquer, D., Ducimetiere, P., Jousilahti, P., and Keil, U., "Estimation of Ten-Year Risk of Fatal Cardiovascular Disease in Europe: the SCORE Project," *European Heart Journal*, Vol.24, No.11, 2003, pp.987-1003.
- Duguay, C., and Chetouane, F., "Modeling and Improving Emergency Department Systems Using Discrete Event Simulation," *Simulation*, Vol.83, No.4, 2007, pp.311-320.
- Ellenius, J., and Groth, T., "Transferability of Neural Network-Based Decision Support Algorithms for Early Assessment of Chest-Pain Patients," *International Journal of Medical Informatics*, Vol.60, No.1, 2000, pp.1-20.
- Erhardt, L., Herlitz, J., and Bossaert, L., "Task Force on the Management of Chest Pain," *European Heart Journal*, Vol.23, 2002, pp.1153-1176.
- Fayyad, U.M., "Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases," *Proceedings of Ninth International Conference on Scientific and Statistical Database Management*, Vol.2, 2002, pp.11-13.
- Fromm, R.E., Gibbs, L.R., McCallum, W.G.,

- Niziol, C., Babcock, J.C., and Gueler, A.C., "Critical Care in the Emergency Department: A Time-Based Study," *Critical Care Medicine*, Vol.21, No.7, 1993, pp.970-976.
- Guyen, A., and Kara, S., "Classification of Electro-Oculogram Signals Using Artificial Neural Network," *Expert Systems with Applications*, Vol.31, No.1, 2006, pp.199-205.
- Kannel, W.B., and Vasan, R.S., "Is Age Really a Non-Modifiable Cardiovascular Risk Factor," *American Journal of Cardiology*, Vol.104, No.9, 2009, pp.1307-1310.
- Kantardzic, M.M., and Zurada, J., *Next Generation of Data-Mining Applications*, John Wiley & Sons, Hoboken, 2005.
- Kenneth, H.B., and Sharon, A.S., "Chest Pain: A Clinical Assessment," *Radiologic Clinics of North America*, Vol.44, No.2, 2006, pp.165-179.
- Khan, F.S., Anwer, R.M., Torgersson, O., and Falkman, G., "Data Mining in Oral Medicine Using Decision Trees," *World Academy of Science, Engineering and Technology*, Vol.37, 2008, pp.225-230.
- Kononenko, I., "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective," *Artificial Intelligence in Medicine*, Vol.23, No.1, 2001, pp.89-109.
- Lin, W.T., Wang, S.T., Chiang, T.C., Shi, Y.X., Chen, W.Y., and Chen, H.M., "Abnormal Diagnosis of Emergency Department Triage Explored with Data Mining Technology: An Emergency Department at a Medical Center in Taiwan Taken as an Example," *Expert Systems with Applications*, Vol.37, No.4, 2010, pp.2733-2741.
- Majumder, S.K., Ghosh, N., and Gupta, P.K., "Support Vector Machine for Optical Diagnosis of Cancer," *Journal of Biomedical Optics*, Vol. 10, No.2, 2005, pp.24-34.
- Martinez-Selles, M., Bueno, H., Sacristan, A., Estevez, A., Ortiz, J., Gallego, L., and Fernandez-Aviles, F., "Chest Pain in the Emergency Department: Incidence, Clinical Characteristics and Risk Stratification," *Revista Espanola de Cardiologia*, Vol.61, No.9, 2008, pp. 953-959.
- Mitchell, T.M., "Machine Learning and Data Mining," *Communications of the ACM*, Vol.42, No.11, 1999, pp.30-36.
- Qiao, D.D., "Clinical Diagnosis of Life Threatening Chest Diseases Based on 156 Cases," *Journal of Chinese Modern Medicine*, Vol.32, No.4, 2009, pp.11-13.
- Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, 1993.

Ren, H., "Clinical Diagnosis of Chest Pain,"
Chinese Journal for Clinicians, Vol.36,
No.1, 2008, pp.5-7.

Ridker, P.M., Hennekens, C.H., Buring, J.E., and
Rifai, N., "C-Reactive Protein and
Other Markers of Inflammation in the
Prediction of Cardiovascular Disease in
Women," *New England Journal of
Medicine*, Vol.342, 2000, pp.836-843.

Rossille, D., Cuggia, M., Arnault, A., Bouget, J.,
and Le Beux, P., "Managing an
Emergency Department by Analyzing
HIS Medical Data: A Focus on Elderly
Patient Clinical Pathways," *Health
Care Management Science*, Vol.11,
No.2, 2008, pp.139-146.

Shmueli, G., Patel, N.R., and Bruce, P.C., *Data
Mining for Business Intelligence:
Concepts, Techniques, and Applications
in Microsoft Office Excel with XLMiner*,
John Wiley & Sons, Hoboken, 2010.

Statistics Korea, <http://www.kostat.go.kr/>, 2010.

Tan, Y., Yin, G.F., Li, G.B., and Chen, J.Y.,
"Mining Compatibility Rules from
Irregular Chinese Traditional Medicine
Database by Apriori Algorithm,"
*Journal of Southwest JiaoTong
University*, Vol.15, No.4, 2009,
pp.288-293.

Yun, Y.P., "Application and Research of Data
Mining based on C4.5 Algorithm,"
*Harbin University of Science and
Technology*, Vol.10, No.8, 2008,

pp.311-315.

Zaki, M.J., "Mining Non-Redundant Association
rules," *Data Mining and Knowledge
Discovery*, Vol.9, No.3, 2004,
pp.223-248.

하성호(Ha, Sung-Ho)

현재 경북대학교 경영학부
교수로 재직 중이며 한국과학
기술원에서 박사학위를 취득
하였다. 관심분야는 데이터마
이닝, e비즈니스, 지식경영, 의
료경영, 지능의사결정 등이다.

장전위(Zhang, Zhen-Yu)

경북대학교 일반대학원 경
영학부에서 석사과정을 취득하
고 동대학원 박사과정에 재학
중이다. 관심분야는 데이터마
이닝, 진단의사결정, 지식경영
등이다.

<초록>

진단지식관리를 위한 앙상블 기법의 실증적 평가

하성호 · 장전위

지난 수십 년 간 연구자들은 효과적인 진료지원시스템을 개발하기 위해 다양한 도구와 방법론들을 제안하였고 지금도 새로운 방법론과 도구들을 계속적으로 개발하고 있다. 그 중에서 흉통으로 응급실에 내원한 노인환자에 대한 정확한 진단은 중요한 이슈 중의 하나였다. 따라서 많은 연구자들이 의사의 진단 능력을 향상시키기 위한 지능적인 의료의사결정과 시스템 개발에 투신하고 있지만 전통적인 의료시스템에 따른 대부분의 진료의사결정이 단일 분류기(classifier)에 기반하고 있어 만족스런 성능을 보여주지 못하고 있는 것이 현실이다. 따라서 이 논문은 앙상블 전략을 활용하여 의사들이 노인환자들의 흉통을 더 정확하고 빠르게 진단하는데 있어 도움을 줄 수 있게 하였다. 의사결정나무, 인공신경망, SVM 모델을 결합한 앙상블 기법을 실제 응급실에서 수집한 응급실 자료에 적용하였고, 그 결과 단일 분류기를 사용하는 것에 비해 월등히 향상된 진단 성과를 보이는 것을 관찰할 수 있었다.

Key Words: 진단, 지식관리, 데이터마이닝, 앙상블 기법.

* 이 논문은 2011년 8월 5일 접수되어 1차수정(2011년 8월 31일)을 거쳐 2011년 9월 4일 게재 확정되었습니다.