International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:4, No:1, 2010

A Hybrid Data Mining Method for the Medical Classification of Chest Pain

Sung Ho Ha and Seong Hyeon Joo

***Abstract*—**Data mining techniques have been used in medical research for many years and have been known to be effective. In order to solve such problems as long-waiting time, congestion, and delayed patient care, faced by emergency departments, this study concentrates on building a hybrid methodology, combining data mining techniques such as association rules and classification trees. The methodology is applied to real-world emergency data collected from a hospital and is evaluated by comparing with other techniques. The methodology is expected to help physicians to make a faster and more accurate classification of chest pain diseases.

***Keywords*—**Data mining, medical decisions, medical domain knowledge, chest pain.

I. Introduction

IN recent years, the hospital's emergency department has been frustrated by the problems of overcrowding, long processing time, delayed patient treatment, and high costs. These problems have been caused from several internal and external factors, including patient characteristics, staffing patterns of emergency department, access to health care providers, patient arrival time, management practices, and testing and treatment strategies selected by emergency department [1]. Understanding these factors well is the first step to solve them and as a result, to improve the efficiency of patient care in emergency departments.

Most hospitals today have employed many different types of hospital information systems to manage their healthcare or patient data. These systems typically generate vast amounts of data in the form of number, text, chart, and image. 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.

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 over the past few years. The goal of predictive data mining in clinical medicine is to derive models that use patient information to support specific clinical decisions. 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 in clinical information systems [2].

Therefore, the purposes of this study are as follows: using data mining techniques, this study focuses on generating the association rules that help physicians to decide which lab tests patients should be tested by, which can eliminate unnecessary lab tests to classify chest pain diseases and reduce testing time and cost in the emergency department. This study then aims at building a classification scheme that supports to make a complex diagnosis, which can help physicians to formulate clinical decisions more quickly and accurately. The organization of this paper is as follows: Section 2 explains medical data mining and its applications to the clinical decisions. Section 3 illustrates the research methodology used in this study and section 4 applies the methodology to real-world emergency data. Section 5 evaluates the methodology and compares its performance with other techniques. Section 6 provides conclusions and future directions.

Sung Ho Ha is with Kyungpook National University, Daegu, 702-701 Republic of Korea (Corresponding author, phone: 82-53-950-5440; fax: 82-53-950-6247; e-mail: hsh@mail.knu.ac.kr).

Seong Hyeon Joo is with Kyungpook National University, Daegu, 702-701 Republic of Korea.

II. Literature Review

Medical data mining has been applied to accurate classification and rapid prediction for prognosis and diagnosis of patients in a specialized medical area [3]. It has been also used for training unspecialized doctors to solve a specific diagnostic problem [4]. Among several algorithms for classification and prediction tasks, a decision tree is one of the most frequently used techniques in a 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, 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. [5] 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 time. Yun [6] utilized a C4.5 algorithm to build a decision tree in order to discover the critical causes of type II diabetes. She has learned about the illness regularity from diabetes data, and has generated a set of rules for diabetes diagnosis and prediction.

Abdullah et al. [7] adopted an association algorithm to find the relationship between diagnosis and prescription. They

stated that purchases and medical bills have much in common. Therefore, the Apriori algorithm was useful to figure out large item sets and to generate association rules in medical billing data. Tan et al. [8] 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 is generally consistent with the traditional Chinese medicine for that disease.

Ceglowski et al. [9] discovered 'treatment pathways' through mining medical treatment procedures in the emergency department. They found that the workload in the emergency department varies depending on the number of presented patients, and is not affected by the type of procedure carried out. Delphine et al.'s [10] has presented a complementary perspective on the activities of the emergency department for specific patient groups: over 75 year old and under 75 year old patients. She thought once validated, these views would be used as decision support tools for delivering better care to this population. Lin et al. [11] 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 performed on 501 abnormal diagnoses done in an emergency department.

III. Research Methodology

A. Emergency Department Process

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 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 [12].

B. Classification Knowledge using Data Mining

Fig. 1 illustrates the research methodology, which consists of three stages. In the first stage, the information of lab tests and diagnoses is collected from the electronic medical data and the association relationships between the lab tests are extracted. The association rules are generated by the Apriori algorithm. The association rule here is an implication of the form $X \rightarrow Y$, which means 'If a patient takes the lab test X, then he will have the lab test Y.' The rule $X \rightarrow Y$ has to meet pre-specified minimum support and minimum confidence levels.

Domain knowledge about diagnostic tests for a specific disease helps select the lab tests associated with that disease (for example, chest pain). Thus, the second stage selects such lab tests from the generated association rules, which have recorded higher scores (e.g., above 0.9) on support and confidence and have included lab tests of importance mentioned in the domain knowledge. For example, if X is one of the critical tests mentioned in the domain knowledge, then, a test Y will be selected from all the rules with the form of $X \rightarrow Y$ or $Y \rightarrow X$, whose support and confidence values are greater than pre-specified minimum support and minimum confidence levels (e.g., 0.9).

In the third stage, a classification tree model and its classification rules are generated to classify a chest pain disease, with using the lab tests selected in the second stage and the information of patients (e.g., previous diagnosis and medical records). In this study, a C5.0 algorithm, one of classification tree models, is chosen to use. Classification rules have the form of 'IF A, then B' where A is the combination of lab tests and B is the classified disease.

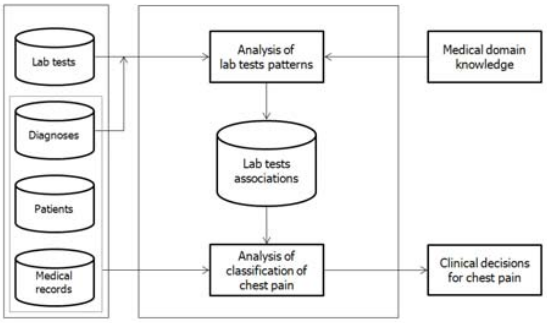

Fig. 1 Research methodology with three stages

1. Lab Tests Patterns

The association algorithm finds the associative relationships between lab tests. Whenever a patient gets medical lab tests, associations also occur between lab tests. The association analysis is applicable for analyzing the relationship among lab tests and the Apriori algorithm is chosen to use in this analysis.

Association rules mining using Apriori is a two-step process, where in the first step, frequent test-sets are discovered and in the second step, association rules are derived from the frequent test-sets [13]. For the first step, the algorithm makes multiple scans over the data. In the first pass, the support of individual test is counted and frequent tests are determined. In each subsequent pass, a set of test-sets, found to be frequent in the previous pass, is used for generating new frequent test-sets, called candidate test-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 repeats until no new frequent test-sets are found.

The second step is to generate the desired association rules from the frequent test-sets. All subsets of every frequent test-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 Support (*f*) to Support (*a*) is at least minimum confidence. Given a set of transactions $\{D = T_1, \ldots, T_n\}$, and a set of tests $\{I = I_1, \ldots, .I_m\}$, such that any transaction *T* in *D* is a set of tests 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 tests.

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*:

$$\text{Support } (A \rightarrow B) = N_{A->B} / N \quad (1)$$

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*:

$$\text{Confidence } (A \rightarrow B) = N_{A->B} / N_A \quad (2)$$

2. Classification of Chest Pain

With the 'selected lab tests' generated by the 'Lab tests patterns' stage, and other information including diagnosis, patient, and medical record data, classification tree models can be generated for the classification of chest pain. This study uses the C5.0 algorithm to generate classification rules.

The main distinctive characteristic of the C5.0 model is how the division is chosen for the units belonging to a group, corresponding to a node of the tree [14]. Let t_r, $r = 1, \ldots, s$, indicate the child groups generated by the subdividing and let p_r indicate the proportion of observations, among those in *t*, that are allocated to each child node with $\sum p_r = 1$. The criterion function is expressed as

$$\Phi(s,t) = I(t) - \sum_{r=1}^{s} I(t_r) p_r \quad (3)$$

where $\Phi(s,t)$ is a measure of the performance gain in subdividing a parent node *t* in a number of child nodes *s*. $I(t)$ represents an impurity function. High values of the criterion function imply that the chosen partition is a good one. Impurity refers to a measure of variability of the response values of the observations. Entropy impurity is a usual choice.

$$I(m) = -\sum_{i=1}^{k(m)} \pi_i \log \pi_i \quad (4)$$

where π_i are the fitted probabilities of the levels, which are present at the node *m* and are at most *k(m)*.

IV. Data Analysis and Results

A. Description of Data

Data involved in this study were based on the electronic medical records of chest pain patients who had received treatment in an emergency department of a hospital. Raw data of 478 patients were extracted from the database of the emergency department. There were 410 kinds of lab tests. Because the same patient sometimes took several tests at a time and some patients took the same test more than once, therefore the total number of lab tests received by 478 patients reached 16,581 records.

B. Association Patterns of Lab Tests

There were 410 kinds of lab tests in the focal emergency environment. No patients, however, need to undergo all the lab tests since they arrived. The patient's condition could get worse while waiting to be tested. If the patient gets tests unnecessary for the diagnosis, it can be a waste of time and money. In order to reduce the number of lab tests, and to reduce time spent on unnecessary tests, the most critical lab tests need to be revealed from all the lab tests. For this reason, domain knowledge from the medical literature was filed to find important lab tests.

According to Butler et al. [15] and Ren [16], acute myocardial infarction (AMI) belongs to an Acute Coronary Syndrome (ACS) and is the most common disease causing chest pain. Creatine Kinase (CK), Creatine Kinase MB fraction (CK-MB), and Troponin are very sensitive and critical tests to diagnose the ACS. Based on these findings, association rules which contain 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 relationships between lab tests, sub-codes indicating specific lab tests were set as both input and target variables.

As a result, among the generated rules (4432 rules), there were 327 rules including three crucial lab tests (CK, CK-MB, and Troponin). To use domain knowledge for the diagnosis of ACS, thus, has reduced the number of rules discovered. Consider the relationship between lab tests J503942_01 and J252640_01, for example. The support value for the rule, which says 'If a patient takes a lab test J503942_01, then he also takes a test J252640_01,' is calculated as follows:

Support (J503942_01 → J252640_01) = 325/327 = 0.9939

The confidence of the rule depends on the consequent and antecedent of the rule:

Confidence (J503942_01 → J252640_01) = 325/326 = 0.9969

Depending on the high threshold of support (0.9) and confidence (0.9), the number of rules has been further reduced from 327 to 246. From these association rules, it was found that 34 lab tests had a strong association with the critical lab tests, such as CK, CK-MB, and Troponin. They are alanine amino transferase, aspartate amino transferase, alkaline phosphatase,

bilirubin direct, total bilirubin, albumin, total protein, total cholesterol, creatinine, blood urea nitrogen, glucose, inorganic phosphorus, total calcium, lactate dehydrogenase, cardiolipin, K, Na, plasma thromboplastin, activated partial thromboplastin time, mean platelet volume, large unstained cell, basophil, mean corpuscular hemoglobin concentration, platelet, neutrophil, lymph corpuscles, mononucleosi, eosinophils, mean corpuscular hemoglobin (methacholine chloride), mean corpuscular volume (mean cell volume), hematocrit, hemoglobin, red blood cell count, and white blood cell count.

C. Classification Rules of Chest Pain

Of 478 patients with diagnostic records, 219 patients were diagnosed with AMI, 106 patients were diagnosed with angina pectoris (AP), and the remaining 153 patients were diagnosed with other chest-pain diseases. For the classification of chest pain, 478 records were split into two data sets at random: a training dataset (289 records, 60%) and a validation dataset (189 records, 40%).

C5.0 adopted a boosting method and a local/global pruning in training. A branch of the tree was split only if the sub-branches contained at least five records. Input variables included 37 lab-testing results and demographic information (i.e., gender, age) of patients. Entropy was the impurity measurement and the target response variable was the diagnostic results that were coded as follows: AMI indicates acute myocardial infarction, AP means angina pectoris, and OTH signifies other chest-pain diseases.

The ages of patients were transformed into nine codes: ages less than 20 were coded as '1'; ages between 20 and 30 were coded as '2'; …, and ages between 90 and 100 were coded as '9'. In order to improve the accuracy of classifying diseases, such as AMI and AP, the misclassification costs of these two diseases were raised, compared to other diseases.

As a result of learning, C5.0 has produced all 19 rules and an example is shown as follows:

If
  AGE > 2
  J503942_01 > 0.150
  J151530_01 <= 45.600
  J25283A_09 > 3.100
  J10101D 04 <= 1.200
Then
  AP (Angina pectoris) (95.56%)

To understand the rule above, J503942_01 (TROPONIN-I) is a rapid qualitative test for the detection of cardiac Troponin-I from whole blood or serum samples. J151530_01, partial thromboplastin time (PTT), is a blood test that measures the time it takes blood to clot. J25283A_09 (Albumin) generally refers to any protein with water solubility, which is moderately soluble in concentrated salt solutions and experiences heat coagulation. J10101D_04 (Hemoglobin) is the iron-containing oxygen-transport metalloprotein in the red blood cells of vertebrates and the tissues of some invertebrates.

V. Validation of Models and Comparison

A. Classification Performance of the C5.0 Model

Table 1 shows the confusion matrix for the classification model (C5.0), obtained on the validation data set (189 records of patients). Of 189 records of patients, there were 35 records of the patients with OTH, 57 patients with AP, and 97 patients with AMI. In the confusion matrix, the column represents classified values while the row represents observed values: OTH means patients with other chest pain diseases; AMI means patients with acute myocardial infarction; and AP indicates patients with angina pectoris. The values on the diagonal are correct classifications.

TABLE I
Confusion Matrix for the C5.0 Model

		Classified		
		OTH	AMI	AP
Actual	OTH	29	2	4
	AMI	0	96	1
	AP	3	1	53

These classifications had an error rate of $100 \times (2 + 4 + 1 + 3 + 1)/189 = 5.82\%$. Total classification accuracy reached 94.18%. The classification accuracy for each disease is 29/35 = 82.86% for other chest pain; 96/97 = 98.97% for AMI; and 53/57 = 92.98% for AP.

B. Comparison of Classification Accuracy

This subsection compares the classification accuracy of the C5.0 model with that of other models, such as Support Vector Machine (SVM) and Neural Network (NN). SVM adopted a radial basis function as the kernel function and set the regularization parameter = 10, the regression precision = 0.1, and the RBF gamma = 0.1. NN placed 19 neurons in the first hidden layer, 5 neurons in the second hidden layer, and 3 neurons in the output layer. NN set the learning rate to 0.9 and the momentum to 0.3, which were decaying gradually to 0.01.

The same training data set (289) and validation data set (189) which had been applied to the C5.0 model were also used for each comparative model. First al all, the models were compared in terms of their gains. Fig. 2 shows the gain charts for the three models being compared. It appears that the gain charts for all three models are rather similar and they have a similar performance.

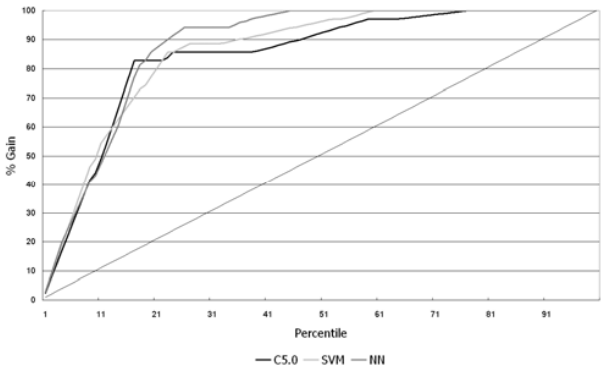

Fig. 2 Gain charts for the considered models obtained on the validation data

To decide between the curves, more information is needed about the classification performance. Thus, this study measured the classification accuracy rates for the models. On the validation set, the C5.0 model has the highest accuracy, followed by the NN model and the SVM model, as shown in Table 2. To summarize, the C5.0 seems to be the best-performing model.

TABLE II
COMPARISON OF CLASSIFICATION ACCURACY RATES FOR EACH MODEL

	C5.0	SVM	NN
Classification accuracy (%)	94.18	85.19	88.89

Table 3 shows the confusion matrices for each model obtained on the validation data set.

TABLE III
CONFUSION MATRICES FOR EACH MODEL

			Classified		
			OTH	AMI	AP
C5.0					
	Actual	OTH	29	2	4
		AMI	0	96	1
		AP	3	1	53
SVM					
	Actual	OTH	30	4	1
		AMI	6	87	4
		AP	8	5	44
NN					
	Actual	OTH	30	4	1
		AMI	2	94	1
		AP	5	8	44

VI. CONCLUSION

The purpose of this study was to build a hybrid data mining model to extract classification knowledge for chest pain to aid in clinical decisions in an emergency department. This study utilized real world data collected from an emergency department of a hospital and used an Apriori algorithm to identify 37 lab tests and a C5.0 algorithm to generate a classification rule base for the classification of chest pain, which can help physicians to make clinical decisions faster and more accurately.

Through training and evaluation, the experimental results showed that the generated classification rules performed well in the classification of chest pain, whose accuracy rate reached 94.18%. By comparing these results with other algorithms, SVM and NN, it has been realized that the C5.0 algorithm has the best performance. Many lab tests unnecessary for classifying chest pain could be filtered out from the 410 lab tests in the raw data, which leads to reducing the waste of time and cost during lab tests for a patient.

Future research should consider several improvements. First, this study adopted a hybrid data mining approach, combining an association rule mining and a classification tree mining. However, another combination of data mining techniques is still possible to accomplish the same task of medical data mining. To think of another methodology will be a good challenge.

Second, as the diversification of medical data increases, the need for adaptive methods of medical decision-making has been gradually expanding. Considering the current situation of emergency department information systems, more complete and comprehensive systems will be necessary for the future. Such a system should have not only medical intelligence mentioned in this study, but also the capability that can be further 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 for 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 whole country, those efforts to generalize the model and methodology are needed.

REFERENCES

[1] R. E. Fromm, L. R. Gibbs, W. G. McCallum, C. Niziol, J. C. Babcock, A. C. Gueler, and R. L. Levine, "Critical care in the emergency department: a time-based study", *Crit. Care Med*., vol. 21, pp. 970-976, 1993.
[2] B. Riccardo and Z. Blaz, "Predictive data mining in clinical medicine: Current issues and guidelines," *International Journal of Medical Informatics*, vol. 77, pp. 81-97, 2008.
[3] G. Masuda, N. Sakamoto, and R. Yamamoto, "A framework for dynamic evidence based medicine using data mining," In *Proc. 15th IEEE Symposium on Computer-Based Medical Systems*, IEEE press, 2002, pp. 117-122.
[4] I. Kononenko, "Machine learning for medical diagnosis: history, state of the art and perspective," *Artificial Intelligence in Medicine*, vol. 23, pp. 89-109, 2001.
[5] F. S. Khan, R. M. Anwer, O. Torgersson, and G. Falkman, "Data mining in oral medicine using decision trees," *International Journal of Biological and Medical Sciences*, vol. 4, pp. 156-161, 2009.
[6] Y. P. Yun, "Application and research of data mining based on C4.5 Algorithm," *Master thesis*, Haerbin University of Science and Technology, 2008.
[7] U. Abdullah, J. Ahmad, A. Ahmed, "Analysis of effectiveness of apriori algorithm in medical billing data mining," In *Proc. 4th International Conference on Emerging Technologies*, IEEE press, 2008, pp. 327-331.
[8] Y. Tan, G. F. Yin, G. B. Li, and J. Y. Chen, "Mining compatibility rules from irregular Chinese traditional medicine database by Apriori algorithm," *Journal of Southwest JiaoTong University*, vol. 15, 2007.
[9] R. Ceglowski, L. Churilov, and J. Wasserthiel, "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, pp. 246-254, 2007.
[10] R. Delphine, M. Cuggia, A. Arnault, J. Bouget, and P. L. Beux, "Managing an emergency department by analyzing HIS medical data: a focus on elderly patient clinical pathways," *Health Care Management Science*, vol. 11, pp. 139-146, 2008.
[11] W. T. Lin, S. T. Wang, T. C. Chiang, Y. X. Shi, W. Y. Chen, and H. M. Chen, "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, pp. 2733-2741, 2010.
[12] C. Duguary, and F. Chetouane, "Modeling and improving emergency department systems using discrete event simulation," *Simulation*, vol. 83, pp. 311-320, 2007.
[13] M. J. Zaki, "Mining non-redundant association rules," *Data Mining and Knowledge Discovery*, vol. 9, pp. 223-248, 2004.

[14] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Morgan Kaufmann, CA: San Francisco, 2005.
[15] K. H. Butler and S. A. Swencki, "Chest pain: a clinical assessment," *Radiologic Clinics of North America*, vol. 44, pp. 165-179, 2006.
[16] H. Ren, "Clinical diagnosis of chest pain," *Chinese Journal for Clinicians*, vol. 36, 2008.

Sung Ho Ha is a professor of business administration at Kyungpook National University in Korea. He received his PhD in industrial engineering from Korea Advanced Institute of Science and Technology. He is an editorial member of several international journals. His research interests include machine learning, data mining, electronic commerce, and total quality management.

Seong Hyeon Joo is a PhD candidate student at Kyungpook National University. His interests include data mining, e-business, and e-learning.