IEMS Vol. 4, No. 1, pp. 102-108, June 2005.

A Comparative Study of Medical Data Classification Methods Based on Decision Tree and System Reconstruction Analysis

Tzung-I Tang†
Department of Information & Electronic Commerce
Kainan University, Taiwan
Tel: 886-3-341-2500 ext. 1122, E-mail: michael@mail.knu.edu.tw

Gang Zheng
Department of Computer Science
Tianjin University of Technology, Tianjin, China
E-mail: zheng-gang@eyou.com

Yalou Huang
Computer Science Institute
Nankai University, Tianjin, China
E-mail: yellow@nankai.edu.cn

Guangfu Shu
Institute of Systems Science
Chinese Academy of Science, Beijing, China
E-mail: guangfu-shu@yahoo.com

Pengtao Wang
Department of Computer Science
Tianjin University of Technology, Tianjin, China
E-mail: wangpengtao@eyou.com

Abstract. This paper studies medical data classification methods, comparing decision tree and system reconstruction analysis as applied to heart disease medical data mining. The data we study is collected from patients with coronary heart disease. It has 1,723 records of 71 attributes each. We use the system-reconstruction method to weight it. We use decision tree algorithms, such as induction of decision trees (ID3), classification and regression tree (C4.5), classification and regression tree (CART), Chi-square automatic interaction detector (CHAID), and exhausted CHAID. We use the results to compare the correction rate, leaf number, and tree depth of different decision-tree algorithms. According to the experiments, we know that weighted data can improve the correction rate of coronary heart disease data but has little effect on the tree depth and leaf number.

Keywords: data mining, decision tree and system analysis, data classification

1. INTRODUCTION

Data mining techniques have been applied to medical services in several areas, including prediction of effectiveness of surgical procedures, medical tests, medication, and the discovery of relationships among clinical and diagnosis data (Prather et al., 1997, Aslandogan and Mahajani, 1988). Our study is concerned with the analysis of coronary heart disease data. Coronary heart disease has become more prevalent in recent years, prompting

† : Corresponding Author

scholars to devote more attention to its risk factors. Early diagnosis and treatment is one of the best approaches to reducing the disease's death rate.

The paper uses data from 1,723 coronary heart disease patients' cases. Each case contains about 71 attributes. The data come from a hospital clinic's observations and allow us to get a good classification of patients' status and behavior, from which we can determine the relationships among the factors. We also want to find a data mining method to analyze the medical data. We will use a system-reconstruction method to do data preprocessing and use decision-tree algorithms to do the classification. We want to compare the classification correction rate on weighted and not weighted data, which is preprocessed by the system-reconstruction method.

In this paper, first we introduce the system-reconstruction method and show how the coronary heart disease data are to be processed, and we discuss the theory and algorithms of decision trees, including, ID3, C4.5, CART, CHAID, and Exhausted-CHAID. We also apply these methods to medical data mining problems by designing some experiments to compare the correction rate, tree depth, and leaf number of weighted and not weighted data gotten by decision tree.

2. Data-Preparation Methods

This study uses the mixed variable system reconstruction and prediction method, which is the most up-to-date study harvest in the field of system theory to weight data on coronary heart disease patients. We get the weight of every factor in the data and use the results in the decision tree analysis. The system reconstruction analysis used in data mining is a kind of system analysis method, based on the constraint analysis theory of Ashby (1965). Through the effort of many scientists from the United States, the Netherlands, Germany, and Japan (Klir, 1976), the leader, theory and methodology architecture (Cavallo and Klir, 1979) was principally established. At that time, the main concerns were how to better partition the whole system into sub-systems and how to better ensure the characteristics of the whole system from local characteristics.

Zwick (2000) and others have addressed the first concern. Approaches to the second question have been developed by Jones (1989), who designed a computation method that uses characters reconstruction and a condition function that reflects system features to ensure main and sub-system conditions (main reflects local). According to the need to study real-world problems, he developed a mixed-variable (continuum variable, discrete variable, and classification variable) factor-analysis method through variable reconstruction analysis (Shu 1997, 1998). This method increases the precision of quantitative results and makes the foundation of quantitative synthesize. At the same time, he presented reconstruction-prediction, evaluation, optimization, and decision-support methods (Shu 1997). These methods are applicable to many fields, including medicine, environmental studies, economy and finance, marketing strategy, industrial management, and talent prediction.

2.1 Theory Description

Our model to analyze patients' data and get the weight of every factor is as follows:

(1) Compute the importance degree of factor congregate state in quota level, maximum and minimum value.

$$\max \Phi_{k,l} = \sum_{n=1}^{N} \max edc^{*}(k, l_k, n, l_n) \quad l_n \in \Omega_n^f \qquad (1)$$

$$\min \Phi_{k,l} = \sum_{n=1}^{N} \min edc^{*}(k, l_k, n, l_n) \quad l_n \in \Omega_n^f \qquad (2)$$

Where, edc is the information entropy distance of property function between hypothesis system and original system, edc^{*}=1/edc, Ω_n^f is all levels' number collection of factor n.

(2) Compute the value generated from the importance degree of factor congregate state in quota level for the sample T.

$$\Phi_{k,l}(T) = \sum_{n=1}^{N} edc^{*}[k, l_R(T), n, l_n(T)] \qquad (3)$$

(3) Compute the realizing degree of every quota level for sample T.

$$\Phi_{k,l_k}(T) = \frac{\Phi_{k,l_k}(T) - \min \Phi_{k,l_k}}{\max \Phi_{k,l_k} - \min \Phi_{k,l_k}} \qquad (4)$$

(4) We use $\Phi_{k,l_k}(T)$, to forecast trends and select the maximum value of the forecasting level.

(5) Compute the forecasting value

When trend forecasting at low levels

$$W_2^f(T) = \frac{\Phi_{2,1}(T) \cdot E_0^{'} + \Phi_{2,2}(T) \cdot E_1^{'}}{\Phi_{2,1}(T) + \Phi_{2,2}(T)} \qquad (5)$$

W_2^f is the predicate value of sample T, $E_0^{'}$ is low level edge, $E_1^{'}$ is middle value.

In case of trend forecasting at middle levels,

$$W_2^f(T) = \frac{\Phi_{2,1}(T) \cdot E_0^{'} + \Phi_{2,2}(T) \cdot E_1^{'} + \Phi_{2,2}(T) \cdot E_2^{'}}{\Phi_{2,1}(T) + \Phi_{2,2}(T) + \Phi_{2,3}(T)} \qquad (6)$$

$E_2^{'}$ is high level edge,

In case of trend forecasting at high levels,

$$W_2^f(T) = \frac{\Phi_{2,2}(T)\cdot E_0^{'} + \Phi_{2,3}(T)\cdot E_2^{'}}{\Phi_{2,2}(T) + \Phi_{2,3}(T)} \quad (7)$$

2.2 Patient case digitalization

First, we will explain how we get the patient's data. The 1,723 patients' data records include the following nine groups: ache, alleviation method, sign, personal history, blood fat, electrocardiogram, ultrasonic cardiogram (UCG), and Holter. Parameters include gender, temperament, age, ache character, position, time, cause, pulse, blood pressure, family history, smoking history, smoking amount, alcohol history, high blood pressure history, diabetes history, blood sugar, uric acid, and arrhythmia. Table 1 shows our patient-case digitalization method.

Table 1. Patient case digitalization

……	Sign					……
……	pulse	Blood pressure Systole	Blood pressure Diastole	Speed	Rhythm	……
……	V19	V20	V21	V22	V23	……
……	number	number	number	Times/s	1:yes -1:no	……

2.3 Result of data reconstruction analysis

Table 2. Patients factor weight sorting table

Order	Factor	Factor weight (0-1)
1	Blood pressure systole high	1
2	Cause of drinking	0.891833
3	Blood pressure diastole high	0.785103
4	Female	0.712903
5	Atrial premature beats, Atrial tachycardia	0.650113
6	Myocardial enzyme GOT serious	0.51834
7	Cause of sleeping	0.511743
8	Myocardial enzyme CKMB serious	0.503684
9	No taking glonoine	0.494534
10	More locations of ache	0.476548
:	:	:
71		

Factors' weight by reconstruction analysis is based on the real diagnosis result. Because the pathogenic of coronary heart disease and the phenomena in diagnosing are diversiform, some factors take a big role in the heart disease of patients, while others do not. Therefore we get a weight table from recon-struction analysis that is based on the real diagnosis result; it can be used in the next phase of analysis. Table 2 is the weight list for the factors of the sample problem.

2.4 Data for next analysis

After we use the system reconstruction method to analyze the patient's data, we get the factor weight, which is important for the next step in our analysis.

When we digitalize the data, we do not consider the relationship among the factors. In fact, they are related, and their degrees of importance to coronary heart disease are different. Since we have the weight of data, we can process the original digitalized data with the factor weight, and the data will be used in the next analysis. Then we get two data sets, weighted and not weighted, which we use in decision-tree analysis and in our attempt to learn which can get better results.

3. Decision-tree method

3.1 Algorithm Description

The decision tree induction algorithm has been used broadly for several years. It is an approximation discrete function method and can yield lots of useful expressions. It is one of the most important methods for classification. This algorithm's terms follow the "tree" metaphor. It has a root, which is the first split point of the data attribute for building a decision tree. It also has leaves, so that every path from root to leaf will form a rule that is easily understood.

Since the decision tree is built by given data, the data value and character will be more important. For example, the amount of data will affect the result of the tree-building procedure. The type of attribute value will also affect the tree model. Decision trees need two kinds of data: training and testing. Training data, which are usually the bigger part of data, are used for constructing trees. The more training data collected, the higher the accuracy of the results. The other group of data, testing, is used to get the accuracy rate and misclassification rate of the decision tree.

Many decision-tree algorithms have been developed. One of the most famous is ID3 (Quinlan 1986, 1983), whose choice of split attribute is based on information entropy. C4.5 is an extension of ID3 (Prather *et al*. 1997). It improves computing efficiency, deals with continuous values, handles attributes with missing values, avoids over fitting, and performs other functions. CART (classification and regression tree) is a data-exploration and prediction algorithm similar to C4.5, which is a tree-construction algorithm (Martínez and Suárez, 2004). Breiman *et al*. (1984) summarized the classification and

regression tree. Instead of information entropy, it introduces measures of node impurity. It is used on a variety of different problems, such as the detection of chlorine from the data contained in a mass spectrum (Berson and Smith, 1997). CHAID (*Chi*-square automatic interaction detector) is similar to CART, but it differs in choosing a split node. It depends on a *Chi*-square test used in contingency tables to determine which categorical predictor is farthest from independence with the prediction values (Bittencourt and Clarke, 2003). It also has an extended version, Exhausted-CHAID.

Although decision trees may not be the best method for classification accuracy, even people who are not familiar with them find them easy to use and understand. Figure 1 shows a binary decision tree. It gives us an impression of a decision. It uses a circle as the decision node and a square as the terminal node. Each decision node has a condition that is represented by a function F, and the parameter is the split point of the split attribute. Each terminal node has a class label C, the value of which represents a class. It is apparent that it is easy to use decision trees to interpret the tree to rules, from which we can do analysis, and easy to interpret the representation of a nonlinear input-output mapping (Jang 1994).

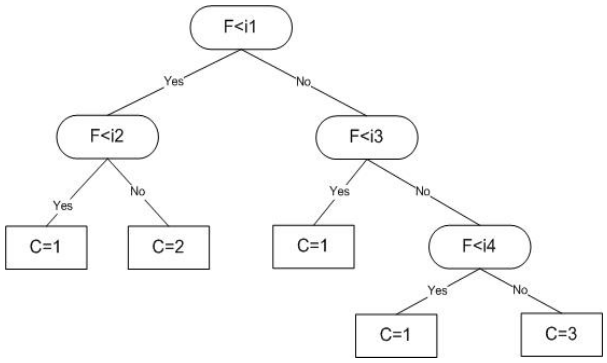

Figure 1. A typical binary decision tree

Lots of works address the splitting node choosing method and optimization of tree size, but less attention has been given to the weight of the data attributes. In this study, we use a system-reconstruction analysis method to get the weight of each attribute, which we use to reform raw data. After that, we use the decision-tree algorithm mentioned above to build a decision tree, from which we can find the decision-accuracy and misclassification rates.

3.2 Induction of decision trees algorithm, ID3

ID3 is a typical decision-tree algorithm. It introduces information entropy as the splitting attribute's choosing measure. It trains a tree from root to leaf, a top-down sequence. Each path from that form is a decision rule. We will discuss the theory of ID3 below.

(1) Define function and expression:

Definition 1. D is defined as a training data set whose attributes are divided into two parts: non-target and target. The non-target attribute is named as Q (Q_1, …,Q_m), where each attribute Q_i(1<=i<=m) takes k_i values $\{a_{i_1},\ldots,a_{i_k}\}$. The target attribute (usually just one attribute) is named as C. Suppose it has l values; thus we get l classes, C={C_1, …,C_l}. Let D^j be a subset in D whose class is C_j and |D| be the number of elements in D. The information entropy of data set D is defined as:

$$E(D) = \sum_{j=1}^{l} (P_j \log_2 P_j) \tag{8}$$

Where P_j is the proportion of D belonging to class j

$$P_j = \frac{|D^j|}{|D|} \tag{9}$$

Definition 2. The measure of the impurity in a collection of training examples is defined as information gain, Gain (D, Q_i), of attribute Q_i :

$$Gain(D, Q_i) = E(D) - E(D, Q_m) \tag{10}$$

$$E(D, Q_m) = \sum_{j=1}^{k_i} (P_{ij} \cdot E(D_{ij})) \tag{11}$$

Where, D_{ij} is the obtained jth subset which is divided by attribute Q_i on D, and

$$P_j = \frac{|D^j|}{\sum_{j}^{K_i} |D_{ij}|} \tag{12}$$

(2) Processing

The target of the ID3 algorithm is to search the attribute with maximum information gain, and to use the attribute as the splitting attribute. Thus, the definition of information entropy becomes an important case to study, for perfect entropy is more reasonable in classification.

3.3 Regression tree algorithm, C4.5

C4.5 is an extended version of ID3. It improves appropriate attribute selection measure, avoids data over fitting, reduces error pruning, handles attributes with different weight, improves computing efficiency, handles missing value data and continuous attributes, and performs other functions.

It is based on the idea of ID4, instead of information gained in ID3, and it introduces an information gain ratio.

We also use the data set used in ID3 to explain the theory of C4.5.

Definition 3. V has n values which are not repeated, shown as $\{V_1,\ldots,V_n\}$, and D is separated into subsets $D_1, D_2,\ldots,D_n$.
$|D|$ is the example number of data set D.
$|T_i|$ is the number of example that $V=V_i$,
$|C_j| = freq(C_j, T)$, number of example on C_j
$|C_{jv}|$ is the number of example on C_j where $V=V_i$.
Probability of C_j:

$$P(C_j) = \frac{|C_j|}{|T|} = \frac{freq(C_j, T)}{|T|} \tag{13}$$

Probability when $V = v_i$: $P(v_i) = \frac{|T_i|}{|T|}$

Probability of C_j when $V = v_i$: $P(C_j \mid v_i) = \frac{|C_{jv}|}{|T_i|}$

Definition 4. Information gain ratio

(1) Class information entropy:

$$E(C) = -\sum_j P(C_j)\log(p(C_j) = -\sum_j \frac{|C_j|}{|T|}\log(\frac{|C_j|}{|T|}) = -\sum_{j=1}^{k} \frac{freq(C_j, T)}{|T|} \times \log_2(\frac{freq(C_j, T)}{T}) = info(T) \tag{14}$$

(2) Class condition entropy:

$$E(C \mid V) = -\sum_j P(v_i)\sum_i P(C_j \mid v_i)\log P(C_j \mid v_i) = -\sum_j \frac{|T_j|}{|T|}\sum_i \frac{|C_{jv}|}{|T_i|}\log\frac{|C_{jv}|}{|T_i|} = \sum_{i=1}^{n}\frac{|T_i|}{|T|} \times info(\mathrm{T_i}) = info_v(\mathrm{T}) \tag{15}$$

(3) Information gain

$$\mathrm{Gain(C,V) = E(C) - H(C|V) = info(T) = infov(T)}$$

(4) Information entropy of attribute V

$$E(V) = -\sum_i P(v_i)\log(P(v_i)) = -\sum_{=1}^{n}\frac{|\mathrm{T_i}|}{|\mathrm{T}|} \times \log_2(\frac{|\mathrm{T_i}|}{|\mathrm{T}|}) = split - \mathrm{info(V)} \tag{16}$$

(5) Information gain ratio

$$\mathrm{Gain\text{-}ratio(V) = E(C,V)/E(V) = gain(V)/split\text{-}info(V)} \tag{17}$$

C4.5 uses an information gain ratio select attribute to split, which yields better information gain than ID3.

3.4 CART

CART (classification and regression tree) is another decision tree algorithm developed by Breiman (Breiman *et al*. 1984). The tree is constructed based on the training set and then pruned by the minimum cost-complexity principle (Jang 1994). Unlike C4.5, which uses information entropy as the measurement of choosing a splitting attribute, it uses impurity. Some key theories are shown below (Prather *et al*. 1997):

Impurity,

$$i(t) = -\sum_{j=1}^{k} p(w_j \mid t)\log p(w_j \mid t) \tag{18}$$

Best division,

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R) \tag{19}$$

If there is no significant decrease in the impurity measurement, and the next divisions cannot be completed, node t will be the terminal node. The class w_j related to terminal node t is that which maximizes the conditional probability $p(w_j \mid t)$ (Prather *et al*. 1997).

3.5 *Chi*-squared automatic interaction detector, CHAID

CHAID is one of the oldest tree-classification methods. It was originally proposed by Kass (1980). According to Ripley, 1996, the CHAID algorithm is a descendent of THAID developed by Morgan and Messenger, 1973. CHAID grows non-binary trees through a relatively simple algorithm that is particularly well suited for the analysis of larger data sets, and it has been particularly popular in marketing research. The algorithm proceeds as follows:

(1) Preparing predictors: create categorical predictors out of any continuous predictors by dividing the respective continuous distributions into a number of categories with an approximately equal number of observations.
(2) Merging categories: cycle through the predictors to determine for each predictor the pair of (predictor) categories that is least significantly different with respect to the dependent variable; for classification problems (where the dependent variable is categorical as well), it will compute a Chi-square test (Pearson Chi-square). If the statistical significance for the respective pair of predictor categories is significant, then it will compute a Bonferroni-adjusted p-value for the set of categories for the respective predictor.
(3) Selecting the split variable: choose the split-the-predictor variable with the smallest adjusted p-value, i.e., the predictor variable that will yield the most signi-

Table 3. Comparison of experimental results by various algorithms Correction rate comparison of decision tree algorithms

decision-tree algorithm	data type	internal node number	max. tree depth	leaves number[0]	correction rate
C4.5	raw data	33	15	35	83.3%
	weighted data	26	15	25	86.3%
CART	raw data	10	6	11	73.4%
	weighted data	8	6	9	78.4%
CHAID	raw data	9	3	6	79.7%
	weighted data	21	3	12	82.2%
Exhaustive CHAID	raw data	15	3	9	78.1%
	weighted data	17	3	10	82.8%

ficant split; if the smallest (Bonferroni) adjusted p-value for any predictor is greater than some alpha-to-split value, then no further splits will be performed and the respective node is a terminal node. Continue this process until no further splits can be performed.

Exhaustive CHAID, a modification of the basic CHAID algorithm, performs a more thorough merging and testing of predictor variables, and hence requires more computing time. For large data sets, and those with many continuous predictor variables, this modification of the simpler CHAID algorithm may require significant computing time.

4. EXPERIMENT DESIGN

After we digitalized the CHD (coronary heart disease) data (1,723 records of 71 attributes), about 1,400 records were used as training sets; the remaining 323 were considered as the testing data sets. From these, we get two kinds of data: raw and weighted. Attributes in this part of data have equal probability, while the other part of data is weighted by the system reconstruction method so that each attribute has a weight value. We will use these two kinds of data as the experiment data. The analysis methods in the experiment will be C4.5, CART, CHAID, and Exhaustive CHAID. We use every algorithm on the raw and weighted data to compare the decision-tree parameters and correction rate. The results are shown in Table 3.

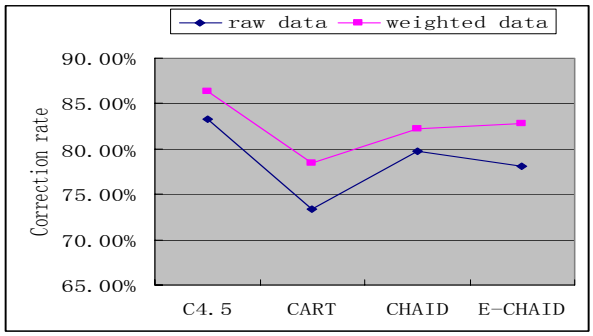

Figure 2. Correction rate comparison of decision tree

Figure 2 clearly reveals that the weighted data in have a higher correction rate than the raw data. A good decision tree usually is judged by the following aspects: minimum leaf number, minimum tree depth, and correction rate. From the experiments we learned that weighted data can get a better correction rate than raw data. From the same data set we can see that CHAID can get minimum leaf number and minimum tree depth, C4.5 in the middle and CART in the last position (see Figure 3). From the figure, we can see that whether data is weighted does not affect the two parameters.

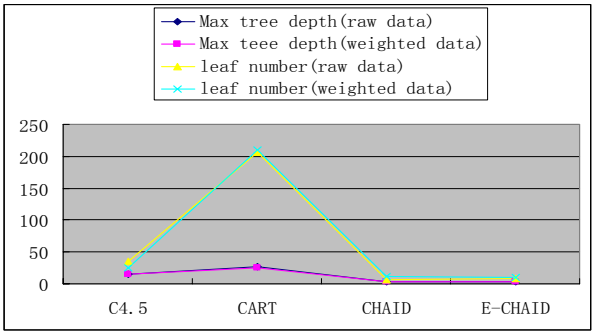

Figure 3. Decision-tree algorithm parameters comparison

Different signs represent different parameters and different data sets: diamonds are maximum tree depth of not-weighted data; empty circles are the maximum tree depth of weighted data; triangles are leaf numbers of not-weighted data; and empty cubes are the leaf number of weighted data.

5. CONCLUSION

The decision-tree algorithm is one of the most effective classification methods. The data we used in the paper were collected directly from clinical diagnoses, and their reliability confirmed by coronary artery radiography. The data will judge the efficiency and correction rate of the algorithm. From the data we get the conclusion that data weighted by the system-reconstruction method can get a higher correction rate but will have little effect on the leaf number and tree depth of the decision tree.

The work will be in several parts. First, we will compare the methods of weighting the data, which will find the best method for weighting data used in decision-tree classification. Second, we will study the different classification methods, such as neural networks and genetic algorithms, based on the weighted data that we studied before. Third, we will study principle component analysis, rough set, feature selection to reduce the attributes of the coronary heart disease data.

References

Ashby, W. R. (1965), Constraint Analysis of Many-dimensional Relations, General Systems Yearbook, 9, 99-105.

Aslandogan, Y. A. and Mahajani, G. A. (2004), Evidence Combination in Medical Data mining, Proceedings of International Conference on Information Technology: Coding and Computing, IEEE.

Berson, A. and Smith, S. J. (1997), Data Warehousing, Data Mining, & OLAP, 365, McGraw-Hil.

Bittencourt, H. R. and Clarke, R. T. (1997), Use of Classification and Regression Trees (CART) to Classify Remotely Sensed Digital Images, IEEE, 3751-3753.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C.J. (1984), *Classification and Decision trees*, Belmont, CA: Wadsworth.

Cavallo, R. E. and Klir, G.. J. (1979), Reconstructability Analysis of Multi-dimensional Relations: A Theoretical Basis for Computer-aided Determination of Acceptable Systems Models, Int. J. of General Systems, 5, 143-171.

Cavallo, R. E. and Klir, G.. J. (1981), Reconstructability Analysis: Evaluation of Reconstruction Hypotheses, Int. J. of General Systems, 7, 7-32.

Jang, J.-S. R. (1994), Structure Determination in Fuzzy Modeling: A fuzzy CART Approach, IEEE, 480-485.

Jones, B. (1998), A Program for Reconstructability Analysis, Int. J. of General Systems, 15, 199-205.

Klir, G.J. (1976), Identification of Generative Structures in Empirical Data, Int. J. of General Systems, 3, No. 2, 89-104.

Martínez-Muñoz, G. and Suárez, A. (2004), Using all Data to Generate Decision Tree Ensemble, IEEE Tran. On Systems, Man and Cybernetics—part C: Applica-tions and Review, 34, No. 4. 393-397.

Prather, J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., and Hammond, W. E. (1997), Medical Data Mining: *Knowledge Discovery in a Clinical Data Warehouse*, Proc AMIA Annual Fall Symposium, 101-5.

Quinlan, J. R. (1983), Learning Efficient Classification Procedures and Their Application to Chess and Games, Machine Learning: An Artificial Intelligence Approach, 1, 463-482.

Quinlan, J. R. (1986), Induction of Decision Trees, Machine Learning, 1, 81-106.

Shu, G. (1997), Reconstruction Analysis Methods for Forecasting, Risks, Design, Dynamical Problems and Applications, Second Workshop of IIGSS, 72-79.

Shu, G. (1998), Reconstructability Analysis with Multiversity Information and Knowledge, Systems Science and its Applications, Third Workshop of IIGSS, 69-74.

Shu, G. (2000), Reconstructability Analysis in China Special Issue, Int. J. of General Systems, 29, No. 3.

Zwick, M. and OCCAM (2000), A Reconstructability Analysis Software Package, World Congress of the Systems Sciences/44th Annual Meeting of ISSS, Toronto, Canada, July, 16-22.