Effective Data Mining of Integrated Data Sets Using Decision Trees

Hyontai Sug

***Abstract*—** CART decision tree algorithm has been considered very good data mining method for data sets in medicine domain. Because CART treats missing values with surrogate variables, it's good for real world data in which some values for attributes are often missing. We are interested in effective data mining of two different liver data sets that are available in the internet having small number of common attributes, while majority of attributes are not common. Experiments using CART for two differently integrated data sets of the two data sets generated successful results. Especially, an overly integrated data set to give each data set almost equal chance to contribute in the final result generated very accurate decision tree with increased tree complexity. Further interactive pruning generated a smaller tree with moderate accuracy. But the accuracy is better than that of the decision tree from conventionally integrated data set.

***Keywords*—**Decision tree induction, CART, data integration, preprocessing.

I. Introduction

DIAGNOSING diseases accurately in medicine domain is very important, even in the situation of limited information is available. Liver is the largest internal organ in the human body, and it is known that the organ is responsible for more than one hundred functions of human body. So, diagnosing the disease as accurate as possible is a high interest to researchers and doctors [1, 2]. There has been much research to diagnose the disease more accurately based on a data set in public domain known as 'liver disorders'. Mostly the research was based on artificial neural network approach for better classification accuracy [3, 4, 5], because the data set is small so that not much information is available for highly accurate classification. But, neural network-based approach has its own limitation that the transformation of trained neural network [6] is not as informative as other data mining tools like decision trees, because the rules from neural network have flat structures, while decision tree have tree structures. So, decision trees can be a good data mining tools, if our interest is on understandability [7]. In addition to 'liver disorders' data set, a data set called 'Indian liver disorder data set' is available in internet public domain since 2012, while 'liver disorders' data set has been available since 1990. So, it'll be interesting to compare the decision trees of the data sets, even though the two data sets have some uncommon attributes.

There are many examples that use decision trees well [8, 9, 10, 11]. Moreover, because we can easily understand the knowledge structures of decision trees, they are considered very good data mining tools in medicine domain so that decision trees are widely accepted in the domain [12, 13]. There are several kinds of decision trees [14]. Among them CART [15] and C4.5 [16] can be two representatives. While CART is often referred in medicine area, C4.5 is often referred in engineering and business area.

C4.5 uses an entropy-based measure to split branches based on attribute values, and the measure selects the most certain split among possible splits of candidate attributes. So, a majority class that has more certain splits in a node is preferred. CART uses a purity-based measure to split branches, and it splits the training data set in a node based on how probably the split makes the child node purer with respect to class values. The algorithm spends more time to generate smaller tree, as a result it produces relatively smaller trees than C4.5.

Therefore, in this paper we want to find some comprehensible decision trees having good accuracy based on CART decision tree generator. Unlikely most other data mining tools, CART was developed by statisticians, Breiman et al., so that it has a very good statistical basis. CART can treat missing values well using surrogate variables and has been known a very good data mining tool for medicine domain [17].

II. MATERIALS AND METHODS

We are interested in finding better decision trees based on CART for two liver disorder data sets, 'liver disorders' [18] and 'Indian liver disorder data set' [19]. Let's see the principle of decision tree algorithm, CART. CART uses GINI index for splitting. GINI index helps to determine how much a node is pure in class value distribution. GINI index can be calculated by equation (1).

$$G = \sum_{i=1\sim r} p_i \ (1 - p_i) = 1 - \sum_{i=1\sim r} p_i^2 \qquad (1)$$

In equation (1) p_i is the probability of class i for the instances of a node in the tree. There are r classes. For possible split in the node each G value for each attribute is calculated. Each attribute is possible candidate for split in the node. CART selects the purist split among possible candidate attributes based on the most diminishing G values, and does binary split. So, the best splitter of a node can be the primary splitter. CART also

This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2012.

H. Sug is with Dongseo University, Busan, 617-716 Korea (corresponding author to provide phone: 51-320-1733; fax: 51-327-8955; e-mail: sht@ gdsu.dongseo.ac.kr).

prepares surrogate splitters that resemble the primary splitter. Surrogate splitters are near equivalent splitters to the primary splitter. So, if the value of an instance for primary splitter is missing, CART considers a surrogate variable for split. If top surrogate variable is missing for the instance, CART uses the second best surrogate, etc. If all surrogates are missing, then CART uses majority rule.

After fully growing a tree, pruning backward is done to generate the minimum cost tree. The cost considers tree complexity and predicted misclassification cost together, and finds the best one. When cross-validation is used, it finds optimal tree with respect to test data [20]. Another consideration in pruning is 1SE tree [21]. SE means the standard error. 1SE tree is the smallest tree of which error rate is not worse than 1 standard error above the optimal tree. So we can set the option to generate some smaller tree than the optimal tree. If we cannot get small enough trees, we may apply more pruning with the expense of error rate. As we can see from the splitting methods of the algorithms, the algorithms fragment a training data set so that instances that are not easily classified by the splitting measure would go down in the lower part of the tree. In this sense, the composition of data set is important for better accuracy and smaller tree. Therefore, in our experiment when we combine the two data sets together, we try to combine the two data sets in a way that affects the result almost equally.

A. Data Sets

Data sets for experiments can be found in UCI machine learning repository [22] named 'liver disorders' [18] and 'Indian liver patient data set' [19]. 'Liver disorders' data set is also known as 'BUPA liver disorders' data set and available since 1990. 'Liver disorders' data set has 345 instances, and there are 145 instances in class 1 (no disease) and 200 instances in class 2 (disease). There are six continuous attributes as independent attributes, and one attribute is dependent attribute that has value of 1 or 2 as class values. There are no missing values in all attributes.

'Indian liver patient data set' is available since 2012. In the data set the number of instances is 583, and there are 167 instances in class 2 (no disease) and 416 instances in class 1 (disease). Therefore, class value has opposite meaning in the two data sets. There are nine continuous attributes as independent attributes, and one independent attribute has gender value. Class attribute has value of 1 or 2. There is small number of missing values. Please see table 1 for details of the attributes of the two data sets. In the table the first column has attribute information of 'liver disorders' data set, while the second column has attribute information of 'Indian liver patient data set'. As we can see there are three common attributes, so we expect that we will have a lot of missing values, if the two data sets are combined.

Table 1. The attributes of the two data sets

Attributes of 'liver disorders' data set	Attributes of 'Indian liver patient data set'	Meaning
mcv		Mean corpuscular volume
Gammagt		Gamma-glutamy transpeptidase
Drinks		Number of half-pint equivalents of alcoholic beverages drunk per day
alkphos	alkphos	Alkaline phosphtase
Sgpt	Sgpt	Alamine aminotransferase
Sgot	Sgot	Aspartate aminotransferase
	Age	Age of patient
	Gender	Gender
	TB	Total bilirubin
	DB	Direct Bilirubin
	TP	Total protains
	ALB	Albumin
	A/G ratio	Albumin and Globulin ratio

So, the two data sets have three common attributes, alkphos, Sgpt, Sgot. Because class value has opposite meaning in the two data sets, the class values of 'liver disorders' was flipped for convenience. Salford system's CART [23] with 10-fold cross validation is used for the experiment. So, a data set is divided into ten equal subsets, then, while nine of ten in data set is used for training and one of ten is used for testing. This alternate process is repeated ten times.

B. Decision Tree for 'liver disorders' Data Set

The following diagram shows decision tree having 10 terminal nodes for 'liver disorders' data set. Overall accuracy in tree learning stage is 77.68%, and overall accuracy in test stage is 70.14%. This accuracy is comparable to the results of other data mining methods. For example, in [24] the accuracy of four different data mining algorithms like Naïve Bayes classifier, C4.5, neural networks, and support vector machines is in the range of 56.52% ~ 71.59% with 10-fold cross validation. Fig. 1 shows the decision tree. In the tree left branch means 'yes' and right branch means 'no'.

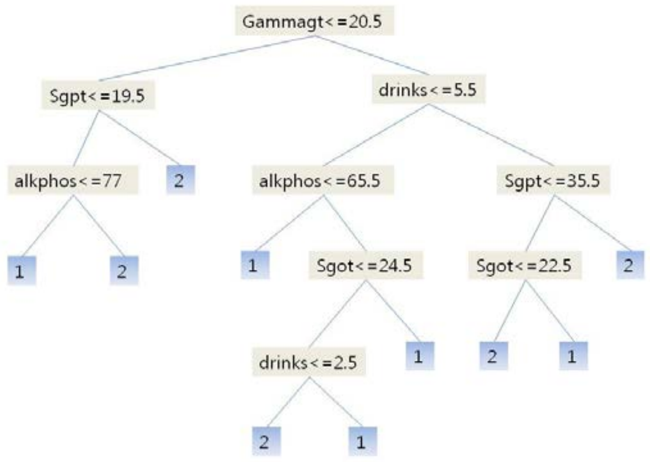

Fig. 1. Decision tree for 'liver disorders' data set

The decision tree can be represented in rule form also. The rules correspond to the terminal nodes of the decision tree from left to right.

1. If (GAMMAGT <= 20.5 & SGPT <= 19.5 & ALKPHOS <= 77) Then class = 1;
2. If (GAMMAGT <= 20.5 & SGPT <= 19.5 & ALKPHOS > 77) Then class = 2;
3. If (GAMMAGT <= 20.5 & SGPT > 19.5) Then class = 2;
4. If (GAMMAGT > 20.5 & DRINKS <= 5.5 & ALKPHOS <= 65.5) Then class = 1;
5. If (GAMMAGT > 20.5 & ALKPHOS > 65.5 & SGOT <= 24.5 & DRINKS <= 2.5) Then class = 2;
6. If (GAMMAGT > 20.5 & ALKPHOS > 65.5 & SGOT <= 24.5 & DRINKS > 2.5 & DRINKS <= 5.5) Then class = 1;
7. If (GAMMAGT > 20.5 & DRINKS <= 5.5 & ALKPHOS > 65.5 & SGOT > 24.5) Then class = 1;
8. If (GAMMAGT > 20.5 & DRINKS > 5.5 & SGPT <= 35.5 & SGOT <= 22.5) Then class = 2;
9. If (GAMMAGT > 20.5 & DRINKS > 5.5 & SGPT <= 35.5 & SGOT > 22.5) Then class = 1;
10. If (GAMMAGT > 20.5 & DRINKS > 5.5 & SGPT > 35.5) Then class = 2;

If 1SE pruning is applied, we have decision tree of 9 terminal nodes. The rightmost subtree that tests 'SGOT <= 22.5' becomes one terminal node. So the 8th and 9th rule are combined to make a rule. The numbering for new rule is adopted for easy distinction of source rules.

89. If (GAMMAGT > 20.5 & DRINKS > 5.5 & SGPT <= 35.5) Then class = 1;

Combining the two terminal nodes generates slight decrease in accuracy. Overall accuracy in tree learning stage becomes 77.10%, and overall accuracy in test stage is 68.99%.

C. Decision Tree for 'Indian liver patient data set'

The following diagram in fig. 2 shows the decision tree of 'Indian liver patient data set' which has only 5 terminal nodes. Overall accuracy in tree learning stage is 69.81%, and overall accuracy in test stage is 64.32%.

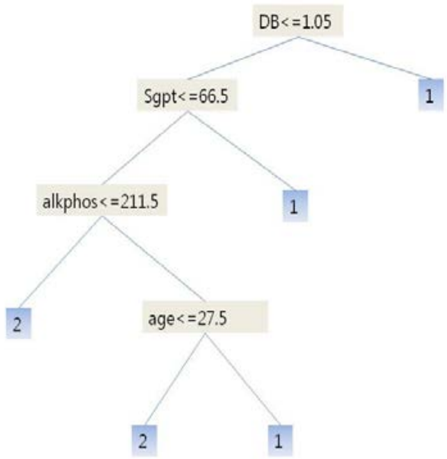

Fig. 2. Decision tree for 'Indian liver patient data set'

The decision tree can be represented in rule form also.

1. If (DB <= 1.05 & SGPT <= 66.5 & ALFPHOS <= 211.5) Then class = 2;
2. If (DB <= 1.05 & SGPT <= 66.5 & ALFPHOS > 211.5 & AGE <= 27.5) Then class = 2;
3. If (DB <= 1.05 & SGPT <= 66.5 & ALFPHOS > 211.5 & AGE > 27.5) Then class = 1;
4. If (DB <= 1.05 & SGPT > 66.5) Then class = 1; 88.5%}
5. If (DB > 1.05) Then class = 1;

If 1SE pruning is applied, we have decision tree of 3 terminal nodes. The left subtree that tests 'ALFPHOS <= 211.5' becomes one terminal node. So the first, second, and third rule are combined as single rule.

123. If (DB <= 1.05 & SGPT <= 66.5) Then class = 2;

Combining three terminal nodes makes some decrease in accuracy. Overall accuracy in tree learning stage becomes 61.58%, and overall accuracy in test stage is 61.41%.

D. Decision Tree of Both Data Sets Combined

The two data sets are integrated to make a decision tree having 7 terminal nodes. The two data sets have common attributes, alkphos, sgpt, and sgot. The vacant values for uncommon attributes are left missing. CART uses surrogate splitters for missing values. So, the best splitter among all non-missing valued splitters related can be a surrogate splitter. Overall accuracy in tree learning stage is 62.82%, and overall accuracy in test stage is 60.24%. Fig. 3 shows the decision tree.

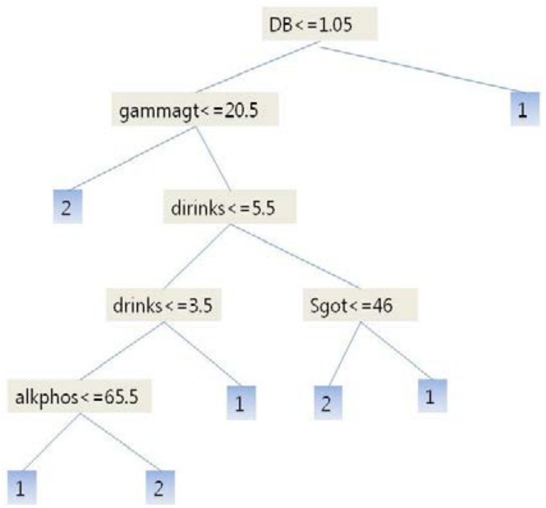

Fig. 3. Decision tree of both data sets combined

The decision tree can be represented in rule form also.

1. If (DB <= 1.05 & GAMMAGT <= 20.5) Then class = 2;
2. If (DB <= 1.05 & GAMMAGT > 20.5 & DRINKS <= 3.5 & ALKPHOS <= 65.5) Then class = 1;
3. If (DB <= 1.05 & GAMMAGT > 20.5 & DRINKS <= 3.5 & ALKPHOS > 65.5) Then class = 2;
4. If (DB <= 1.05 & GAMMAGT > 20.5 & DRINKS > 3.5 & DRINKS <= 5.5) Then class = 1;
5. If (DB <= 1.05 & GAMMAGT > 20.5 & DRINKS > 5.5 & SGOT <= 46) Then class = 2;
6. If (DB <= 1.05 & GAMMAGT > 20.5 & DRINKS > 5.5 & SGOT > 46) Then class = 1;
7. If (DB > 1.05) Then class = 1;

1SE pruning also generates the same tree.

E. Alternate Decision Tree for Both Data Sets Combined

Because 'Indian liver patient data set' has larger number of instances (583) than 'liver disorders' data set (345), an integration based on three times of 'liver disorders' data set plus two times of 'Indian liver patient data set' is made to give each data set almost equal chance to contribute. The resulting decision tree has root node having test of 'DB <= 1.15' and 106 terminal nodes that can be translated into rules directly. The overall accuracy in tree learning stage is 95.71%, and overall accuracy in test stage is 85.76%. The tree can be represented as the following 106 rules.

1. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & SGPT <= 26.5 & MCV <= 87.5 & DRINKS <= 1.25 & ALKPHOS <= 72.5 & SGOT <= 18) Then class = 2;
2. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & SGPT <= 26.5 & MCV <= 87.5 & DRINKS <= 1.25 & ALKPHOS <= 72.5 & SGOT > 18) Then class = 1;
3. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & SGPT <= 26.5 & MCV <= 87.5 & DRINKS <= 1.25 & ALKPHOS > 72.5 & ALKPHOS <= 80.5) Then class = 2;
4. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & SGPT <= 26.5 & MCV <= 87.5 & DRINKS <= 1.25 & ALKPHOS > 80.5) Then class = 1;
5. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & SGPT <= 26.5 & MCV <= 87.5 & DRINKS > 1.25 & DRINKS <= 3.5) Then class = 2;
6. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & SGPT <= 26.5 & MCV <= 87.5 & DRINKS > 3.5) Then class = 1;
7. If (DB <= 1.15 & TB <= 1.5 & MCV > 87.5 & ALKPHOS <= 59.5 & SGPT <= 21 & GAMMAGT <= 14.5) Then class = 1;
8. If (DB <= 1.15 & TB <= 1.5 & MCV > 87.5 & ALKPHOS <= 59.5 & SGPT <= 21 & GAMMAGT > 14.5 & GAMMAGT <= 15.5) Then class = 2;
9. If (DB <= 1.15 & TB <= 1.5 & MCV > 87.5 & ALKPHOS <= 59.5 & SGPT <= 21 & GAMMAGT > 15.5 & GAMMAGT <= 20.5) Then class = 1;
10. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & MCV > 87.5 & ALKPHOS <= 59.5 & SGPT > 21 & SGPT <= 26.5) Then class = 2;
11. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV > 87.5 & MCV <= 90.5 & AGE <= 31.5 & ALKPHOS > 59.5 & ALKPHOS <= 122.5) Then class = 2;
12. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV > 87.5 & MCV <= 90.5 & AGE <= 31.5 & ALKPHOS > 122.5 & ALKPHOS <= 169) Then class = 1;
13. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV > 87.5 & MCV <= 90.5 & ALKPHOS > 59.5 & ALKPHOS <= 169 & AGE > 31.5 & AGE <= 65.5) Then class = 2;
14. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV > 87.5 & MCV <= 90.5 & AGE <= 59 & ALKPHOS > 169 & ALKPHOS <= 185.5) Then class = 1;
15. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV > 87.5 & MCV <= 90.5 & AGE <= 59 & ALKPHOS > 185.5 & ALKPHOS <= 191.5) Then class = 2;
16. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV > 87.5 & MCV <= 90.5 & AGE <= 59 & ALKPHOS > 191.5 & TOTALPROT <= 7.2 & SGPT <= 23.5) Then class = 2;
17. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV > 87.5 & MCV <= 90.5 & AGE <= 59 & ALKPHOS > 191.5 & TOTALPROT <= 7.2 & SGPT > 23.5 & SGPT <= 26.5) Then class = 1;
18. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV >

87.5 & MCV <= 90.5 & AGE <= 59 & ALKPHOS > 191.5 & TOTALPROT > 7.2) Then class = 1;

19. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT <= 12.5 & DRINKS <= 1.25 & MCV > 87.5 & MCV <= 90.5 & ALKPHOS > 169 & AGE > 59 & AGE <= 65.5) Then class = 2;
20. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & ALKPHOS > 59.5 & GAMMAGT <= 12.5 & AGE <= 65.5 & DRINKS <= 1.25 & MCV > 90.5 & SGOT <= 21) Then class = 2;
21. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & ALKPHOS > 59.5 & GAMMAGT <= 12.5 & AGE <= 65.5 & DRINKS <= 1.25 & MCV > 90.5 & SGOT > 21) Then class = 1;
22. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & MCV > 87.5 & ALKPHOS > 59.5 & GAMMAGT <= 12.5 & AGE <= 65.5 & DRINKS > 1.25 & DRINKS <= 2.5) Then class = 1;
23. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & MCV > 87.5 & GAMMAGT <= 12.5 & AGE <= 65.5 & DRINKS > 2.5 & ALKPHOS > 59.5 & ALKPHOS <= 447) Then class = 2;
24. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & MCV > 87.5 & GAMMAGT <= 12.5 & AGE <= 65.5 & DRINKS > 2.5 & ALKPHOS > 447) Then class = 1;
25. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & MCV > 87.5 & GAMMAGT <= 12.5 & AGE > 65.5 & ALKPHOS > 59.5 & ALKPHOS <= 132) Then class = 2;
26. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & MCV > 87.5 & GAMMAGT <= 12.5 & AGE > 65.5 & ALKPHOS > 132) Then class = 1;
27. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT > 12.5 & GAMMAGT <= 20.5 & MCV > 87.5 & MCV <= 90.5 & ALKPHOS > 59.5 & ALKPHOS <= 75.5 & SGPT <= 21) Then class = 1;
28. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT > 12.5 & GAMMAGT <= 20.5 & MCV > 87.5 & MCV <= 90.5 & ALKPHOS > 59.5 & ALKPHOS <= 75.5 & SGPT > 21 & SGPT <= 26.5) Then class = 2;
29. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT > 12.5 & GAMMAGT <= 16.5 & SGOT <= 24.5 & MCV > 87.5 & MCV <= 89.5 & ALKPHOS > 75.5 & ALKPHOS <= 238 & SGPT <= 13) Then class = 2;
30. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT > 12.5 & GAMMAGT <= 16.5 & SGOT <= 24.5 & MCV > 87.5 & MCV <= 89.5 & ALKPHOS > 75.5 & ALKPHOS <= 238 & SGPT > 13 & SGPT <= 26.5 & AGE <= 33) Then class = 2;
31. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT > 12.5 & GAMMAGT <= 16.5 & SGOT <= 24.5 & MCV > 87.5 & MCV <= 89.5 & SGPT > 13 & SGPT <= 26.5 & AGE > 33 & AGE <= 47.5 & ALKPHOS > 75.5 & ALKPHOS <= 102.5) Then class = 2;
32. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT > 12.5 & GAMMAGT <= 16.5 & SGOT <= 24.5 & MCV > 87.5 & MCV <= 89.5 & SGPT > 13 & SGPT <= 26.5 & AGE > 33 & AGE <= 47.5 & ALKPHOS > 102.5 & ALKPHOS <= 238) Then class = 1;
33. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT > 12.5 & GAMMAGT <= 16.5 & SGOT <= 24.5 & MCV > 87.5 & MCV <= 89.5 & ALKPHOS > 75.5 & ALKPHOS <= 238 & AGE > 47.5 & SGPT > 13 & SGPT <= 16) Then class = 1;
34. If (DB <= 1.15 & TB <= 1.5 & GAMMAGT > 12.5 & GAMMAGT <= 16.5 & SGOT <= 24.5 & MCV > 87.5 & MCV <= 89.5 & ALKPHOS > 75.5 & ALKPHOS <= 238 & AGE > 47.5 & SGPT > 16 & SGPT <= 26.5) Then class = 2;
35. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT > 12.5 & GAMMAGT <= 16.5 & SGOT <= 24.5 & MCV > 87.5 & MCV <= 89.5 & ALKPHOS > 238) Then class = 1;
36. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & ALKPHOS > 75.5 & SGOT <= 24.5 & MCV > 89.5 & MCV <= 90.5 & GAMMAGT > 12.5 & GAMMAGT <= 13.5) Then class = 2;
37. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & SGOT <= 24.5 & MCV > 89.5 & MCV <= 90.5 & GAMMAGT > 13.5 & GAMMAGT <= 16.5 & TOTALPROT <= 6.55 & ALKPHOS > 75.5 & ALKPHOS <= 137) Then class = 1;
38. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & SGOT <= 24.5 & MCV > 89.5 & MCV <= 90.5 & GAMMAGT > 13.5 & GAMMAGT <= 16.5 & TOTALPROT <= 6.55 & ALKPHOS > 137) Then class = 2;
39. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & SGOT <= 24.5 & MCV > 89.5 & MCV <= 90.5 & GAMMAGT > 13.5 & GAMMAGT <= 16.5 & TOTALPROT > 6.55 & ALKPHOS > 75.5 & ALKPHOS <= 147.5) Then class = 2;
40. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & SGOT <= 24.5 & MCV > 89.5 & MCV <= 90.5 & GAMMAGT > 13.5 & GAMMAGT <= 16.5 & ALKPHOS > 147.5 & TOTALPROT > 6.55 & TOTALPROT <= 7.25) Then class = 1;
41. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & SGOT <= 24.5 & MCV > 89.5 & MCV <= 90.5 & GAMMAGT > 13.5 & GAMMAGT <= 16.5 & ALKPHOS > 147.5 & TOTALPROT > 7.25) Then class = 2;
42. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & MCV > 87.5 & MCV <= 90.5 & ALKPHOS > 75.5 & GAMMAGT > 12.5 & GAMMAGT <= 16.5 & SGOT > 24.5) Then class = 1;
43. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & MCV > 87.5 & MCV <= 90.5 & ALKPHOS > 75.5 & GAMMAGT > 16.5 & GAMMAGT <= 18) Then class = 1;
44. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & MCV > 87.5 & MCV <= 90.5 & ALKPHOS > 75.5 & GAMMAGT > 18 & GAMMAGT <= 20.5) Then class = 2;

45. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT > 12.5 & GAMMAGT <= 20.5 & MCV > 90.5 & MCV <= 94 & ALKPHOS > 59.5 & ALKPHOS <= 191.5) Then class = 2;
46. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & GAMMAGT > 12.5 & GAMMAGT <= 20.5 & MCV > 90.5 & MCV <= 94 & ALKPHOS > 191.5) Then class = 1;
47. If (DB <= 1.15 & TB <= 1.5 & SGPT <= 26.5 & ALKPHOS > 59.5 & GAMMAGT > 12.5 & GAMMAGT <= 20.5 & MCV > 94) Then class = 1;
48. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & SGPT > 26.5 & ALKPHOS <= 145.5) Then class = 2;
49. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB <= 1.5 & SGPT > 26.5 & ALKPHOS > 145.5) Then class = 1;
50. If (DB <= 1.15 & GAMMAGT <= 20.5 & TB > 1.5) Then class = 1;
51. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO <= 0.965 & SGOT <= 52.5 & AGE <= 32.5) Then class = 2;
52. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO <= 0.965 & SGOT <= 52.5 & AGE > 32.5 & AGE <= 39.5) Then class = 1;
53. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO <= 0.965 & SGOT <= 52.5 & AGE > 39.5 & AGE <= 66 & TOTALPROT <= 4.3) Then class = 1;
54. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO <= 0.965 & SGOT <= 52.5 & AGE > 39.5 & AGE <= 66 & TOTALPROT > 4.3) Then class = 2;
55. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO <= 0.965 & SGOT <= 52.5 & AGE > 66) Then class = 1;
56. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO <= 0.965 & SGOT > 52.5 & TB <= 1.95) Then class = 1;
57. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO <= 0.965 & SGOT > 52.5 & TB > 1.95) Then class = 2;
58. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT <= 24.5 & ALKPHOS <= 65.5 & SGPT <= 22.5) Then class = 1;
59. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT <= 24.5 & ALKPHOS <= 65.5 & SGPT > 22.5 & MCV <= 89.5) Then class = 1;
60. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & ALKPHOS <= 65.5 & MCV > 89.5 & DRINKS <= 1.5 & SGPT > 22.5 & SGPT <= 30.5) Then class = 1;
61. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & ALKPHOS <= 65.5 & MCV > 89.5 & DRINKS <= 1.5 & SGPT > 30.5) Then class = 2;
62. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & SGPT > 22.5 & MCV > 89.5 & DRINKS > 1.5 & DRINKS <= 5.5 & ALKPHOS <= 50) Then class = 1;
63. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & SGPT > 22.5 & MCV > 89.5 & DRINKS > 1.5 & DRINKS <= 5.5 & ALKPHOS > 50 & ALKPHOS <= 65.5) Then class = 2;
64. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & ALKPHOS > 65.5 & DRINKS <= 0.25) Then class = 1;
65. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & DRINKS > 0.25 & DRINKS <= 2.5 & GAMMAGT > 20.5 & GAMMAGT <= 67 & MCV <= 95.5 & ALKPHOS > 65.5 & ALKPHOS <= 78) Then class = 2;
66. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & DRINKS > 0.25 & DRINKS <= 2.5 & GAMMAGT > 20.5 & GAMMAGT <= 67 & MCV <= 95.5 & ALKPHOS > 78 & ALKPHOS <= 79.5) Then class = 1;
67. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 0.25 & DRINKS <= 2.5 & GAMMAGT > 20.5 & GAMMAGT <= 67 & MCV <= 95.5 & ALKPHOS > 79.5 & SGPT <= 37.5 & SGOT <= 21.5) Then class = 2;
68. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 0.25 & DRINKS <= 2.5 & GAMMAGT > 20.5 & GAMMAGT <= 67 & MCV <= 95.5 & ALKPHOS > 79.5 & SGPT <= 37.5 & SGOT > 21.5 & SGOT <= 24.5) Then class = 1;
69. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & DRINKS > 0.25 & DRINKS <= 2.5 & GAMMAGT > 20.5 & GAMMAGT <= 67 & MCV <= 95.5 & ALKPHOS > 79.5 & SGPT > 37.5) Then class = 1;
70. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & ALKPHOS > 65.5 & DRINKS > 0.25 & DRINKS <= 2.5 & GAMMAGT > 20.5 & GAMMAGT <= 67 & MCV > 95.5) Then class = 1;
71. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & ALKPHOS > 65.5 & DRINKS > 0.25 & DRINKS <= 2.5 & GAMMAGT > 67) Then class = 1;
72. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & DRINKS > 2.5 & DRINKS <= 5.5 & ALKPHOS > 65.5 & ALKPHOS <= 69) Then class = 2;
73. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & DRINKS > 2.5 & DRINKS <= 5.5 & ALKPHOS > 69 & SGPT <= 34) Then class = 1;
74. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 24.5 & DRINKS > 2.5 & DRINKS <= 5.5 & ALKPHOS > 69 & SGPT > 34) Then class = 2;

75. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT > 24.5 & MCV <= 85.5 & SGPT <= 32.5) Then class = 1;
76. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT > 24.5 & MCV <= 85.5 & SGPT > 32.5 & SGPT <= 53 & ALKPHOS <= 77) Then class = 1;
77. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT > 24.5 & MCV <= 85.5 & SGPT > 32.5 & SGPT <= 53 & ALKPHOS > 77) Then class = 2;
78. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT > 24.5 & MCV <= 85.5 & SGPT > 53) Then class = 1;
79. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT > 24.5 & MCV > 85.5 & MCV <= 98.5) Then class = 1;
80. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT > 24.5 & MCV > 98.5) Then class = 2;
81. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGPT <= 25.5 & ALKPHOS <= 49) Then class = 2;
82. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & ALKPHOS > 49 & SGPT <= 21.5) Then class = 1;
83. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & ALKPHOS > 49 & SGPT > 21.5 & SGPT <= 23) Then class = 2;
84. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & ALKPHOS > 49 & SGPT > 23 & SGPT <= 25.5) Then class = 1;
85. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGPT > 25.5 & SGOT <= 46 & GAMMAGT > 20.5 & GAMMAGT <= 36.5) Then class = 2;
86. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 46 & GAMMAGT > 36.5 & SGPT > 25.5 & SGPT <= 36.5 & DRINKS > 5.5 & DRINKS <= 7.5) Then class = 1;
87. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 46 & GAMMAGT > 36.5 & SGPT > 25.5 & SGPT <= 36.5 & DRINKS > 7.5 & ALKPHOS <= 64) Then class = 1;
88. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & SGOT <= 46 & GAMMAGT > 36.5 & SGPT > 25.5 & SGPT <= 36.5 & DRINKS > 7.5 & ALKPHOS > 64) Then class = 2;
89. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGOT <= 46 & GAMMAGT > 36.5 & SGPT > 36.5 & MCV <= 89) Then class = 1;
90. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGOT <= 46 & GAMMAGT > 36.5 & MCV > 89 & SGPT > 36.5 & SGPT <= 52) Then class = 2;
91. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGOT <= 46 & GAMMAGT > 36.5 & MCV > 89 & SGPT > 52 & SGPT <= 54) Then class = 1;
92. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & GAMMAGT > 36.5 & MCV > 89 & SGPT > 54 & SGOT <= 28.5) Then class = 1;
93. If (DB <= 1.15 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & GAMMAGT > 36.5 & MCV > 89 & SGPT > 54 & SGOT > 28.5 & SGOT <= 46) Then class = 2;
94. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGPT > 25.5 & SGOT > 46 & ALKPHOS <= 70) Then class = 2;
95. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGPT > 25.5 & SGOT > 46 & ALKPHOS > 70) Then class = 1;
96. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB > 3.45 & SGPT <= 49 & SGOT <= 20.5) Then class = 1;
97. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB > 3.45 & SGPT <= 49 & ALKPHOS <= 169.5 & SGOT > 20.5 & SGOT <= 29.5) Then class = 2;
98. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB > 3.45 & ALKPHOS <= 169.5 & SGOT > 29.5 & SGPT <= 39) Then class = 1;
99. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB > 3.45 & ALKPHOS <= 169.5 & SGOT > 29.5 & SGPT > 39 & SGPT <= 49) Then class = 2;
100. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB > 3.45 & SGOT > 20.5 & ALKPHOS > 169.5 & SGPT <= 47) Then class = 2;
101. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB > 3.45 & SGOT > 20.5 & ALKPHOS > 169.5 & SGPT > 47 & SGPT <= 49) Then class = 1;
102. If (GAMMAGT > 20.5 & ALB > 3.45 & SGPT > 49 & DB <= 0.25) Then class = 1;
103. If (GAMMAGT > 20.5 & ALB > 3.45 & SGPT > 49 & DB > 0.25 & DB <= 1.15 & AGE <= 39.5) Then class = 2;
104. If (GAMMAGT > 20.5 & ALB > 3.45 & SGPT > 49 & DB > 0.25 & DB <= 1.15 & AGE > 39.5) Then class = 1;
105. If (DB > 1.15 & ALB <= 4.1) Then class = 1;
106. If (DB > 1.15 & ALB > 4.1) Then class = 2;

1SE pruning also generates the same tree.

F. Alternate Pruned Decision Tree of Both Data Sets Combined

Even though the good accuracy of the alternate decision tree of both data combined, because the tree is large, its understandability may be limited. So, 25 steps of interactive pruning were performed, and generated the decision tree of 10 terminal nodes. Overall accuracy in tree learning stage becomes 68.44%, and overall accuracy in test stage becomes 64.67%. Comparing this tree with the tree that came from conventional data combination, we get 5.58% and 4.43% more accuracy in learning and test stage respectively with the increase of three terminal nodes. If we need more accuracy, we can always go back. Fig. 4 shows the decision tree.

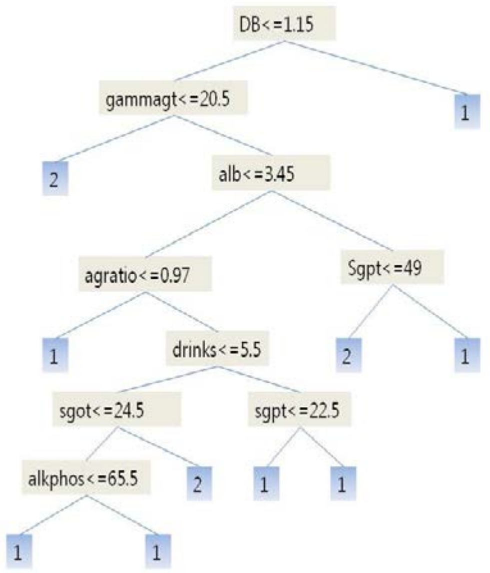

Fig. 4. Alternate pruned decision tree of both data sets

The decision tree can be represented in rule form also.

1. If (DB <= 1.15 & GAMMAGT <= 20.5) Then class = 2;
2. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO <= 0.965) Then class = 1;
3. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT <= 24.5 & ALKPHOS <= 65.5) Then class = 1;
4. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT <= 24.5 & ALKPHOS > 65.5) Then class = 2;
5. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS <= 5.5 & SGOT > 24.5) Then class = 1;
6. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGPT <= 25.5) Then class = 1;
7. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB <= 3.45 & AGRATIO > 0.965 & DRINKS > 5.5 & SGPT > 25.5) Then class = 2;
8. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB > 3.45 & SGPT <= 49) Then class = 2;
9. If (DB <= 1.15 & GAMMAGT > 20.5 & ALB > 3.45 & SGPT > 49) Then class = 1;
10. If (DB > 1.15) Then class = 1;

1SE pruning also generates the same tree.

III. CONCLUSIONS

Liver is a very important organ that is responsible for more than one hundred functions of human body. The complexity of this organ makes it easily affected by disease of disorder. So diagnosing liver disorder disease accurately is a high interest in medicine domain, and decision trees can be a useful data mining tool to diagnose the disease, because decision trees have good properties in understandability and transformability.

There are several popular decision tree algorithms. Among them CART decision tree algorithm has been considered a very good data mining method for data sets in medicine domain. CART treats missing values with surrogate variables, and this property of CART makes it a unique data miner for real world data in which some values of attributes are often missing. The two liver data sets that are available in the internet have three common attributes, while other ten attributes are not common. The uncommon attributes leave missing values, if the two data sets are integrated. Experiments using CART for the integrated data sets of the two liver data sets generated successful results. Especially, overly integrated data set to give each data set almost equal chance to contribute in the final result generated very accurate decision tree with increased tree complexity. Further interactive pruning generated a smaller tree with moderate accuracy. But this accuracy is better than that of the decision tree from conventionally integrated data.

REFERENCES

[1] V. Kichura, Liver Disorders and Diseases. Available: http://www.ehow.com/about_5048281_liver-disorders-diseases.html

[2] R. Ribeiro, R. Marinho, J. Velosa, F. Ramalho, and J.M. Sanches, "Chronic liver disease staging classification based on ultrasound, clinical and laboratorial data," in *Proceedings of 2011 IEEE International Symposium on Biomedical Imaging from Nano to Macro*, 2011, pp. 707-710

[3] M. Neshat, M. Yaghobi, and M. Naghibi, “Designing expert system of liver disorders by using neural network and comparing it with parametric and nonparametric system,” in *Proceedings of 5th International Multi-Conference on Systems, Signals and Devices*, 2008, pp. 1-6.

[4] H. Kahramanli, N. Allahverdi, “Mining Classification Rules for Liver Disorders,” *International Journal of Mathematics and Computers in Simulation*, vol. 3, issue 1, pp. 9-19, 2009.

[5] C. Dendek and J. Mańdziuk, “Improving Performance of a Binary Classifier by Training Set Selection,” in *Proceedings of 18th International Conference on Artificial Neural Networks*, LNCS 5163, pp. 128-135, 2008.

[6] Z. Zhou, “Rule Extraction: Using Neural Networks or For Neural Networks?” *Journal of Computer Science and Technology*, vol.19, no.2, pp. 249-253, 2004.

[7] R. Lin, "An intelligent model for liver disease diagnosis," *Artificial Intelligence in Medicine*, vol. 47, issue 1, pp.53-62, 2009.

[8] Y. Hui, Z. Longqun, and L. Xianwen, "Classification of Wetland from TM imageries based on Decision Tree", *WSEAS Transactions on Information Science and Applications*, issue 7, vol. 6, pp. 1155-1164, July 2009.

[9] S. Segrera and M.N. Moreno, "An Experimental Comparative Study of Web Mining Methods for Recommender Systems," in *Proceedings of the 6th WSEAS International Conference on Distance Learning and Web Engineering*, Lisbon, Portugal, September 22-24, 2006, pp. 56-61.

[10] V. Podgorelec, "Improved Mining of Software Complexity Data on Evolutionary Filtered Training Sets," *WSEAS Transactions on Information Science and Applications*, issue 11, vol. 6, pp. 1751-1760, November 2009.

[11] C. Huang, Y. Lin, and C. Lin, "Implementation of classifiers for choosing insurance policy using decision trees: a case study," *WSEAS Transactions on Computers*, issue 10, vol. 7, pp. 1679-1689, October 2008.

[12] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: an overview and their use in medicine," *Journal of Medical Systems*, Kluwer Academic/Plenum Press, vol. 26, no. 5, pp. 445-463, 2002.

[13] Y.C. Lin, “Design and Implementation of an Ontology-Based Psychiatric Disorder Detection System,” *WSEAS Transactions on Information Sciences and Applications*, issue 1, vol. 7, pp. 56-69, January 2010.

[14] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, World Scientific Publishing Company, 2008.

[15] L. Breiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Wadsworth International Group, 1984.

[16] Quinlan, J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers, Inc., 1993.

[17] R.J. Lewis, An Introduction to Classification and Regression Tree (CART) Analysis. 2000. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf

[18] Liver Disorders Data Set. Available : http://archive.ics.uci.edu/ml/datasets/Liver+Disorders

[19] B. V. Ramana, M.S.P. Babu, and N.B. Venkateswarlu, “A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis”, International Journal of Computer Science Issues, ISSN :1694-0784, May 2012 pp.506-516.

[20] H. Shi, Best-first Decision Tree Learning. PhD thesis, Department of Computer Science, The University of Waikato, pp. 9-14, 2006.

[21] H. Shi, Best-first Decision Tree Learning. PhD thesis, Department of Computer Science, The University of Waikato, pp. 14-15, 2006.

[22] A. Frank and A. Suncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Sciences, 2010

[23] CART® Classification and Regression Trees. Available: http://www.salford-systems.com/en/products/cart

[24] B. V. Ramana, M.S.P. Babu, and N.B. Venkateswarlu, “A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis,” *International Journal of Database Management Systems*, vol. 3, no. 2, pp. 101-114, 2011.

Hyontai Sug received the B.S. degree in Computer Science and Statistics from Busan National University, Busan, Korea, in 1983, the M.S. degree in Computer Science from Hankuk University of Foreign Studies, Seoul, Korea, in 1986, and the Ph.D. degree in Computer and Information Science & Engineering from University of Florida, Gainesville, FL, USA in 1998. He is an associate professor of the Division of Computer and Information Engineering of Dongseo University, Busan, Korea from 2001. From 1999 to 2001, he was a full time lecturer of Pusan University of Foreign Studies, Busan, Korea. He was a researcher of Agency for Defense Development, Korea from 1986 to 1992. His areas of research include data mining and database applications.