Performance Comparison of Decision Tree Algorithms for Medical Data Sets

Hyontai Sug

***Abstract*—** Decision trees have been favored much for the task of data mining in medicine domain, because understandability of found knowledge from the data mining is important. There are two representative decision tree algorithms that have been widely used; C4.5 and CART. While C4.5 has been used some wide range of areas, CART has been favored mostly in medicine domain. Even though the two algorithms have been used in different domains, this fact does not guarantee that CART will be always good data mining tool for all the data sets in medicine domain. In order to compare the performance of the two decision tree algorithms 13 different data sets in medicine domain were used for experiment, and the experiment showed that C4.5 can be a better choice for more cases and CART can be better tool for the cases of lots of missing values in the data sets.

***Keywords*—**Data integration, data mining medicine domain, decision trees, C4.5, CART.

I. Introduction

DECISION trees are one of the mostly used data mining algorithms, and there are many examples that use decision trees well [1, 2, 3, 4]. Moreover, they have been used in medicine domain also as a good tool for diagnosing disease [5, 6], because we can easily understand the structure of trained decision trees, so that we can see how the decision is made.

There are several decision tree algorithms available. According to a survey that was done in the IEEE International Conference on Data Mining (ICDM 2006), two decision tree algorithms were elected among other data mining algorithms as the most popular data mining algorithms. The two decision tree algorithms are C4.5 [7] and CART [8], and each was ranked number one and number ten respectively among other data mining algorithms [9].

C4.5 and CART algorithms have very different tree construction procedures so that their performance can be different for the same data sets. The two algorithms use two different splitting methods in generating each subtree. C4.5 is based on the concept of entropy, and by the tree building process it wants to get smaller and smaller entropies as it generates its children nodes. In C4.5 the number of branches for nominal attributes is dependent on the number of values in the splitting node, while the number of branches for continuous attributes is binary split. On the other hand, CART is based on relatively simpler strategy that uses Gini index. As CART generates its children nodes in binary split only, it wants to get purer nodes than their parents. A node becomes purer, if the node consists of more identical classes of instances. The two algorithms treat missing values differently. While C4.5 treats missing value as a new value in tree building process, CART prepares surrogates for the cases of missing values in each split. As a result, the two decision tree algorithms have attracted different application domains. According to literature survey C4.5 has been used in some wide range of areas [10] like financial areas [11] and engineering areas [12], but CART has been favored mostly in medicine domain, because most researchers in medicine domain reported good performance of CART in their data mining tasks [13].

Because decision tree algorithms fragment a training data set during tree construction process, the resulting decision trees are heavily dependent on the available training data set, and this tendency is severer when the size of the training data set is relatively small. Therefore, we cannot say that CART will be always good for data sets in medicine domain, because data sets in medicine domain are usually relatively small. So, we need some objective research that compares the performance of the two algorithms empirically for wide variety of data sets in medicine domain.

II. Materials and Methods

Thirteen different data sets in medicine domain for experiments can be found in UCI machine learning repository [14]. Table 1 has the summary of the data sets. All data sets are prepared for classification problem. The values in the column 'number of attributes' count for class attribute also.

Table 1. The data sets of medicine domain in UCI machine learning repository

Data set	No. of instances	No. of attributes	No. of classes
Breast tissue	106	10	6
Bupa	345	7	2
Cardiotocography	2126	42	3
Cleveland heart disease	303	14	5
Hungarian heart disease	294	14	5
Switzerland heart disease	123	14	5
VA Long heart disease	200	14	5

This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2013.

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).

Dermatology	366	35	6
Fertility	100	10	2
ILPD	583	11	2
Mammographic mass	961	6	2
Parkinson's	195	24	2
Vertebral column	310	7	2

The experiments were performed in two different ways. The first experiment is based on the original 13 data sets, and the second experiment is based on integrated data sets, where the integration is possible among the data sets.

A. Decision Trees for the Original Data Sets

Table 2 shows the accuracy of C4.5 and CART for each data set in 10-fold cross validation. The number in parentheses represents the tree size. Salford system's CART 7.0 is used for the experiment [15].

Table 2. The accuracy of C4.5 and CART for each data set in 10-fold cross validation

Data set	Accuracy (%)	
	C4.5	CART
Breast tissue	66.04 (29)	**70.75** (17)
Bupa	68.7 (51)	**70.14** (19)
Cardiotocography	**98.78** (27)	98.1 (27)
Cleveland heart disease	**55.78** (63)	53.14 (57)
Hungarian heart disease *	**68.71** (67)	52.04 (55)
Switzerland heart disease	**29.27** (41)	21.95 (5)
VA Long heart disease	**34.0** (75)	30.0 (17)
Dermatology	93.99 (41)	**94.81** (13)
Fertility *	**87.0** (1)	55.0 (3)
ILPD	**68.78** (73)	64.32 (9)
Mammographic mass	82.31 (20)	**83.56** (17)
Parkinson's	80.51 (23)	**84.62** (9)
Vertebral column-2classes	**81.61** (19)	77.42 (3)
Vertebral column-3classes	81.61 (23)	**84.52** (19)
No. of wins	8	6
Average tree size	39.5	19.29
Average of accuracy difference when the algorithm is better	9.10	2.54

From the result of experiment in table 2, we can see that CART may be somewhat inferior to C4.5. Especially for 'Hungarian heart disease' and 'Fertility' data set that are indicated by '*' in the table, CART's accuracy is very bad compared to that of C4.5. On the other hand, for the cases of CART's accuracy are better, the accuracy difference is relatively small, as we can see the fact in the last line of table2. As a conclusion, we can say that C4.5 algorithm may generate more dependable results than CART

Because CART has the tendency of generating smaller trees, more severe pruning was applied to C4.5 to generate similar sized trees of C4.5. The tree size of CART ± 10 or so was generated. Different certainty factor(CF) values of C4.5 may generate differently sized trees. Note that the result of experiment in table 2 is based on default CF value of C4.5 which is 25%. Note that the default CF value was set based on the experience of the creator of C4.5 algorithm. Table 3 has the result of experiment.

Table 3. The accuracy of C4.5 having similar tree size with CART for each data set in 10-fold cross validation

Data set	Accuracy (%)	
	C4.5	CART
Breast tissue	67.92 (19, CF=2%)	**70.75** (17)
Bupa	67.54 (15, CF=5%)	**70.14** (19)
Cardiotocography	**98.78** (27, CF=default)	98.1 (27)
Cleveland heart disease	**56.11** (49, CF=15%)	53.14 (57)
Hungarian heart disease *	**67.35** (63, CF=9%)	52.04 (55)
Switzerland heart disease	**37.4** (5, CF=2%)	21.95 (5)
VA Long heart disease	**32.0** (11, CF=0.7%)	30.0 (17)
Dermatology	93.17 (29, CF=0.05)	**94.81** (13)
Fertility *	**87.0** (1, CF=default)	55.0 (3)
ILPD	**70.33** (1, 1%)	64.32 (9)
Mammographic mass	82.31 (20, CF=default)	**83.56** (17)
Parkinson's	79.49 (11, CF=1%)	**84.62** **(9)**
Vertebral column-2classes	**82.58** (13, CF=5%)	77.42 (3)
Vertebral column-3classes	82.58 (17, CF=1%)	**84.52** (19)
No. of wins	8	6
Average tree size	20.07	19.29

Average of accuracy difference when the algorithm is better	9.95	2.57

Note that for the data set of dermatology C4.5 could not generate further smaller trees, even for smaller CF values than CF of 5%, so the tree size of 29 is the minimum tree to generate. Fig.1 ~ fig. 20 shows the corresponding decision trees. In the figures decision trees for the data set of Cardiotocography are omitted, because the two algorithms generated the same sized trees. Among the four heart disease data sets the decision trees of Cleveland heart disease data set were drown only for comparison in fig. 5 and fig. 6. The decision trees of CART were represented in similar format to that of C4.5 for easy comparison. Details in the nodes of CART were omitted for simplicity.

```
IO <= 551.879287
|  area <= 1664.674076
|  |  DA <= 53.5996
|  |  |  MaxIP <= 18.131014: fad (4.0)
|  |  |  MaxIP > 18.131014
|  |  |  |  PA500 <= 0.165806: gla (19.0/4.0)
|  |  |  |  PA500 > 0.165806: fad (4.0/2.0)
|  |  DA > 53.5996
|  |  |  IO <= 355
|  |  |  |  area <= 346.091312: mas (3.0)
|  |  |  |  area > 346.091312
|  |  |  |  |  PA500 <= 0.127409: fad (8.0)
|  |  |  |  |  PA500 > 0.127409: mas (5.0/2.0)
|  |  |  IO > 355: mas (4.0)
|  area > 1664.674076: car (23.0/3.0)
IO > 551.879287
|  P <= 1524.609204: con (15.0/1.0)
|  P > 1524.609204: adi (21.0)
```

Fig. 1. The decision tree of C4.5 having accuracy of 67.92% for breast tissue data set

```
IO <= 600.62
|  area <= 1710.45
|  |  DA <= 36.50
|  |  |  MaxIP <= 18.18: fad
|  |  |  MaxIP > 18.18: gla
|  |  DA > 36.50
|  |  |  ADA <= 5.17: mas
|  |  |  ADA > 5.17
|  |  |  |  MaxIP <= 29.23
|  |  |  |  |  P <= 216.70: gla
|  |  |  |  |  P > 216.70: fad
|  |  |  |  MaxIP > 29.23: mas
|  area > 1710.45: car
IO > 600.62
|  P <= 1563.84: con
|  P > 1563.84: adi
```

Fig. 2. The decision tree of CART having accuracy of 70.75% for for breast tissue data set (win)

```
gammagt <= 20
|  sgpt <= 19
|  |  gammagt <= 7: 1 (4.0)
|  |  gammagt > 7
|  |  |  alkphos <= 77: 2 (42.0/6.0)
|  |  |  alkphos > 77: 1 (14.0/5.0)
|  sgpt > 19: 1 (80.0/20.0)
gammagt > 20
|  drinks <= 5: 2 (136.0/33.0)
|  drinks > 5
|  |  drinks <= 12
|  |  |  sgot <= 22: 1 (15.0/4.0)
|  |  |  sgot > 22: 2 (50.0/18.0)
|  |  drinks > 12: 1 (4.0)
```

Fig. 3. The decision tree of C4.5 having accuracy of 67.54% for Bupa data set

```
gammagt <= 20.50
|  sgpt <= 19.50
|  |  alkphos <= 77.00: 1
|  |  alkphos > 77.00: 2
|  sgpt > 19.50: 2
gammagt > 20.50
|  drinks <= 5.50
|  |  alkphos <= 65.50: 1
|  |  alkphos > 65.50
|  |  |  sgot <= 24.50
|  |  |  |  drinks <= 2.50: 2
|  |  |  |  drinks > 2.50: 1
|  |  |  sgot > 24.50: 1
|  drinks > 5.50
|  |  sgpt <= 35.50
|  |  |  sgot <= 22.50: 2
|  |  |  sgot > 22.50: 1
|  |  sgpt > 35.50: 2
```

Fig. 4. The decision tree of CART having accuracy of 70.14% for Bupa data set (win)

The following fig. 5 and fig. 6 shows decision trees of Cleveland heart disease data set. Question mark in fig. 5 represents the value of attribute which has missing or unknown value.

```
thal = 3: 0 (167.1/37.55)
thal = 6
|  exang = 0: 0 (10.06/5.0)
|  exang = 1: 2 (8.06/3.0)
|  exang = ?: 0 (0.0)
thal = 7
|  cp = 1: 0 (8.0/3.0)
|  cp = 2: 0 (9.0/4.0)
|  cp = 3
|  |  oldpeak <= 1.9: 0 (17.39/6.0)
|  |  oldpeak > 1.9: 3 (5.0/2.0)
|  cp = 4
|  |  oldpeak <= 0.6
| | | ca = 0
| | | | age <= 42: 1 (3.0)
| | | | age > 42
| | | | | chol <= 237: 0 (5.0)
| | | | | chol > 237: 1 (2.5/1.0)
| | | ca = 1
| | | | chol <= 240: 2 (3.0)
| | | | chol > 240: 1 (2.25/1.0)
| | | ca = 2: 3 (3.15/1.15)
| | | ca = 3: 0 (2.1/1.1)
| | | ca = ?: 0 (0.0)
| | oldpeak > 0.6
| | | trestbps <= 148
| | | | thalach <= 134
| | | | | exang = 0: 2 (3.0/1.0)
| | | | | exang = 1: 3 (21.0/9.0)
| | | | | exang = ?: 3 (0.0)
| | | | thalach > 134
| | | | | slope = 0: 2 (0.0)
| | | | | slope = 1: 1 (7.0/3.0)
| | | | | slope = 2: 2 (13.39/3.0)
| | | | | slope = 3: 1 (1.0)
| | | | | slope = ?: 2 (0.0)
| | | trestbps > 148
| | | | restecg = 0: 1 (2.0/1.0)
| | | | restecg = 1: 3 (0.0)
| | | | restecg = 2
| | | | | slope = 0: 4 (0.0)
| | | | | slope = 1: 3 (1.0)
| | | | | slope = 2: 4 (6.0/1.0)
| | | | | slope = 3: 3 (3.0/1.0)
| | | | | slope = ?: 4 (0.0)
| | | | restecg = ?: 3 (0.0)
| cp = ?: 1 (0.0)
thal = ?: 0 (0.0)
```

Fig. 5. The decision tree of C4.5 having accuracy of 56.11% for Cleveland heart disease data set (win)

Because CART performs binary split only, attributes of nominal values may have several values and represented in parenthesis like 'thal = (6, 7)' as in figure 6.

```
thal =(6, 7)
| trestbps <= 144.50
| | oldpeak <= 2.30
| | | age <= 55.5
| | | | ca = (0,1)
| | | | | slope = (1, 3): 1
| | | | | slope = 2
| | | | | | restecg = (1, 2): 3
| | | | | | restecg = 0
| | | | | | | cp = (2, 4): 2
| | | | | | | cp = (1, 3): 1
| | | | ca = (2, 3): 3
| | | age > 55.5
| | | | ca = (1, 2, 3)
| | | | | thalach <= 133.00: 3
| | | | | thalach > 133.00: 2
| | oldpeak > 2.30
| | | thalach <= 131.00: 3
| | | thalach > 131.00
| | | | trestbps <= 131.00: 4
| | | | trestbps > 131.00: 2
| trestbps > 144.50
| | oldpeak <= 0.70:1
| | oldpeak > 0.70
| | | trestbps <= 167.50
| | | | thalach <= 113.00: 3
| | | | thalach > 113.00
| | | | | thalach <= 135.00: 4
| | | | | thalach > 135.00
| | | | | | chol <= 285.50: 3
| | | | | | chol > 285.00: 4
| | | trestbps > 167.50: 3
thal = 3
| ca = (1, 2, 3)
| | age <= 75.50
| | | chol <= 180.50: 4
| | | chol > 180.50
| | | | cp = (1, 2, 4)
| | | | | thalach <= 161.00
| | | | | | ca = (1, 3): 2
| | | | | | ca = 2: 3
| | | | | thalach > 161.00: 1
| | | | cp = 3: 0
| | age > 75.50: 4
| ca = 0
| | age <= 57.50: 0
| | age > 57.50
| | | thalach <= 83.50: 2
| | | thalach > 83.50
| | | | age <= 64.5
| | | | | fbs = 0: 1
| | | | | fbs = 1:
| | | | age > 64.5: 0
```

Fig. 6. The decision tree of CART having accuracy of 53.14% for Cleveland heart disease data set

For dermatology data set the attribute names were named Ai where i is the sequence number of attributes in fig. 7 and fig. 8, because the original attribute description is somewhat lengthy [16].

```
A27 = 0
| A15 = 0
| | A31 = 0
| | | A28 = 0: 1 (117.0/6.0)
| | | A28 = 1: 4 (9.0/2.0)
| | | A28 = 2
| | | | A5 = 0
| | | | | A26 = 0: 2 (35.0/2.0)
| | | | | A26 = 1: 4 (3.0)
| | | | | A26 = 2: 2 (0.0)
| | | | | A26 = 3: 2 (0.0)
| | | | A5 = 1: 4 (17.0)
| | | | A5 = 2: 4 (8.0/1.0)
| | | | A5 = 3: 4 (2.0)
| | | A28 = 3
| | | | A5 = 0: 2 (21.0/1.0)
| | | | A5 = 1: 4 (5.0)
| | | | A5 = 2: 4 (3.0)
| | | | A5 = 3: 2 (0.0)
| | A31 = 1: 6 (4.0/1.0)
| | A31 = 2: 6 (13.0)
| | A31 = 3: 6 (4.0)
| A15 = 1: 5 (8.0)
| A15 = 2: 5 (22.0/1.0)
| A15 = 3: 5 (23.0)
A27 = 1: 3 (3.0/1.0)
A27 = 2: 3 (43.0)
A27 = 3: 3 (26.0)
```

Fig. 7. The decision tree of C4.5 having accuracy of 93.17% for dermatology data set

```
A31 = (1, 2, 3): 6
A31 = 0
| A27 = (1, 2, 3): 3
| A27 = 0
| | A15 = (1, 2, 3): 5
| | A15 = 0
| | | A20 = (1, 2, 3): 1
| | | A20 = 0
| | | | A5 = (1, 2, 3): 4
| | | | A5 = 0
| | | | | A26 = (1, 2, 3): 4
| | | | | A26 = 0: 2
```

Fig. 8. The decision tree of CART having accuracy of 94.81% for dermatology data set (win)

C4.5 generates a very simple decision tree consisting of only one node. The main reason is that the data set does not have sufficient number of instances in the other class 'O'.

```
: N (100.0/12.0)
```

Fig. 9. The decision tree of C4.5 having accuracy of 87% for fertility data set (win)

CART also generates very simple tree, but tries classification for the other class 'O', but it has poorer accuracy than that of C4.5.

```
age <= 0.65: N
age > 0.65: O
```

Fig. 10. The decision tree of CART having accuracy of 55% for fertility data set

C4.5 experiences similar situation for ILPD data set with fertility data set, when severer pruning of CF 1% was performed.

```
: 1 (583.0/167.0)
```

Fig. 11. The decision tree of C4.5 having accuracy of 70.33% for ILPD data set (win)

```
DB <= 1.05
| SGPT <= 66.50
| | ALFPHOS <=211.50: 2
| | ALFPHOS > 211.50
| | | AGE <= 27.50: 2
| | | AGE > 27.50: 1
| SGPT > 66.50: 1
DB > 1.05: 1
```

Fig. 12. The decision tree of CART having accuracy of 64.32% for ILPD data set

The following fig. 13 and fig. 14 shows decision trees for mammographic mass data set. '?' mark means missing value in fig. 13. '(x, y, …)' means nominal values in fig. 14. Note that nominal values are treated differently in C4.5 and CART as we can see from the attributes, BI_RADS, shape, mad margin in each decision tree.

```
BI_RADS = 1: 0 (0.0)
BI_RADS = 2: 0 (14.27/1.18)
BI_RADS = 3: 0 (36.69/6.46)
BI_RADS = 4
| shape = 1: 0 (189.39/18.41)
| shape = 2: 0 (178.7/18.33)
| shape = 3: 0 (58.22/16.59)
| shape = 4
| | age <= 69
| | | margin = 1: 0 (4.16/1.12)
| | | margin = 2: 0 (2.08/0.06)
| | | margin = 3: 0 (16.33/6.19)
| | | margin = 4: 1 (61.46/29.87)
| | | margin = 5: 1 (24.57/10.21)
| | | margin = ?: 0 (0.0)
| | age > 69: 1 (22.54/2.24)
| shape = ?: 0 (0.0)
BI_RADS = 5: 1 (352.6/42.2)
BI_RADS = ?: 0 (0.0)
```

Fig. 13. The decision tree of C4.5 having accuracy of 82.31% for mammographic mass data set

```
BI_RADS = 5: 1
BI_RADS = (2, 3, 4)
| shape = 4
| | age <= 69.50
| | | margin = (4, 5)
| | | | age <= 35.50: 0
| | | | age > 35.50: 1
| shape = (1, 2, 3)
| | age <= 57.50: 0
| | age > 57.50
| | | margin = (2, 4, 5): 1
| | | margin = (1, 3): 0
```

Fig. 14. The decision tree of CART having accuracy of 83.56% for mammographic mass data set (win)

For Parkinson’s data set the attribute names were named Ai where i is the sequence number of attributes in fig. 15 and fig. 16, because the original attribute description is somewhat lengthy [17].

```
a3 <= 189.621
| a2 <= 234.619
| | a1 <= 129.336
| | | a6 <= 0.00196: 0 (22.0/7.0)
| | | a6 > 0.00196: 1 (52.0/1.0)
| | a1 > 129.336: 1 (76.0/2.0)
| a2 > 234.619
| | a21 <= 0.213353: 0 (12.0)
| | a21 > 0.213353: 1 (14.0)
a3 > 189.621: 0 (19.0/1.0)
```

Fig. 15. The decision tree of C4.5 having accuracy of 79.49% for Parkinson’s data set

```
a23 <= 0.13: 0
a23 > 0.13
| a12 <= 0.01
| | a1 <= 117.99: 0
| | a1 > 117.99
| | | a11 <= 0.01: 1
| | | a11 > 0.01: 0
| a12 > 0.01: 1
```

Fig. 16. The decision tree of CART having accuracy of 84.62% for Parkinson’s data set (win)

Fig. 17 ~ fig. 20 shows decision trees of vertebral column data sets. The data sets have two different class definitions – 2 classes and 3 classes.

```
degree_spondylolisthesis <= 19.854759
| pelvic_radius <= 125.212716
| | sacral_slope <= 40.475232
| | | pelvic_tilt <= 9.976664
| | | | pelvic_radius <= 115.877017: Abnormal (5.0)
| | | | pelvic_radius > 115.877017: Normal (9.0)
| | | pelvic_tilt > 9.976664: Abnormal (62.0/16.0)
| | sacral_slope > 40.475232
| | | degree_spondylolisthesis <= 9.064582: Normal
                                              (31.0/4.0)
| | | degree_spondylolisthesis > 9.064582: Abnormal
                                              (6.0/1.0)
| pelvic_radius > 125.212716: Normal (52.0/7.0)
degree_spondylolisthesis > 19.854759: Abnormal
                                              (145.0/2.0)
```

Fig. 17. The decision tree of C4.5 having accuracy of 82.58% for vertebral column-2 classes data set (win)

```
degree_spondylolisthesis <= 20.09: Normal
degree_spondylolisthesis > 20.09: Abnormal
```

Fig. 18. The decision tree of CART having accuracy of 77.42% for vertebral column-2 classes data set

```
degree_spondylolisthesis <= 15.779697
| sacral_slope <= 46.636577
| | pelvic_radius <= 117.422259: Hernia (46.0/12.0)
| | pelvic_radius > 117.422259
| | | sacral_slope <= 28.131342
| | | | pelvic_tilt <= 17.114312
| | | | | pelvic_tilt <= 14.930725
| | | | | | degree_spondylolisthesis <= 0.75702:
                                          Normal (4.0/1.0)
| | | | | | degree_spondylolisthesis > 0.75702: Hernia
                                                    (5.0)
| | | | | pelvic_tilt > 14.930725: Normal (4.0)
| | | | pelvic_tilt > 17.114312: Hernia (10.0)
| | | sacral_slope > 28.131342: Normal (68.0/10.0)
| sacral_slope > 46.636577
| | degree_spondylolisthesis <= 8.235294: Normal
                                              (21.0/1.0)
| | degree_spondylolisthesis > 8.235294:
                                Spondylolisthesis (4.0/1.0)
degree_spondylolisthesis > 15.779697:
                              Spondylolisthesis (148.0/3.0)
```

Fig. 19. The decision tree of C4.5 having accuracy of 82.58% for vertebral column-3 classes data set

```
degree_spondylolisthesis <= 16.08
| pelvic_radius <= 125.30
| | sacral_slope <= 40.57
| | | pelvic_tilt <= 10.49
| | | | pelvic_radius <= 116.02: Hernia
| | | | pelvic_radius > 116.02: Normal
| | | pelvic_tilt > 10.49: Hernia
| | sacral_slope > 40.57
| | | pelvic_tilt <= 20.17: Normal
| | | pelvic_tilt > 20.17
| | | | lumbar_lordosis_angle <= 56.40: Hernia
| | | | lumbar_lordosis_angle > 56.40: Normal
| pelvic_radius > 125.30
| | sacral_slope <= 29.93
| | | degree_spondylolisthesis <= 1.60: Normal
| | | degree_spondylolisthesis > 1.60: Hernia
| | sacral_slope > 29.93: Normal
degree_spondylolisthesis > 16.08: Spondylolisthesis
```

Fig. 20. The decision tree of CART having accuracy of 84.52% for vertebral column-3 classes data set (win)

B. Decision Trees for Integrated Data Sets

Because CART devote a lot of effort to deal with missing values, next experiment is for the situation that several data sets are integrated. Data integration occurs often in real world situation [18].

Data sets in two different domains were used for integration. As the first domain of integration, the four heart disease data sets are integrated. Because all the four data sets have the same attributes, the integrated data sets have the same missing values as the individual data sets. Table 4 explains the attributes of the data set.

Table 4. The details of attributes of heart disease data sets

Attribute	Meaning
Age	Age
Sex	Gender
Cp	Chest pain type
Trestbps	Resting blood pressure
Chol	Serum cholestoral level
Fbs	Fasting blood sugar
Restecg	Resting electrocardiographic results
Thalach	Maximum heart rate achieved
Exang	Exercise induced angina
Oldpeak	ST depression induced by exercise
Slope	The slope of the peak exercise ST segment
Ca	Number of major vessels colored by flourosopy
Thal	Normal, fixed defect, reversible defect

The number of instances of data set, Cleveland heart disease, Hungarian heart disease, Switzerland heart disease, and VA Long heart disease is 303, 294, 123, and 200 respectively. So, two different integrations were done. First, in simple integration, each data set is combined. Second, because each data set has some different number of instances, two times of Cleveland heart disease data set, and two times of Hungarian heart disease data set, and five times of Switzerland heart disease data set, and three times of VA Long heart disease data set were combined to give each data set almost equal chance to contribute. But, this second leaves more missing values. Table 5 compares the distribution of missing values for each attribute in the two different integration methods.

Table 5. The percentage of missing values in the two different integration methods for heart disease data sets

Attribute	Cleveland×1+Hungarian×1+Switzerland×1+VALong×1	Cleveland×2+Hungarian×2+Switzerland×5+VALong×3
Age	0	0
Sex	0	0
Cp	0	0
Trestbps	6	7
Chol	3	3
Fbs	10	17
Restecg	0	0
Thalach	6	7
Exang	6	7
Oldpeak	7	8
Slope	34	32
Ca	66	74
Thal	53	54
Average	14.69	16.08

Table 6 shows the accuracy of two different decision trees.

Table 6. The accuracy of C4.5 and CART for integrated data sets of heart disease in 10-fold cross validation

Data set	Accuracy of C4.5	Accuracy of CART
Cleveland×1+Hungarian×1+Switzerland×1+VALong×1	**48.59** (117)	46.52 (29)
Cleveland×2+Hungarian×2+Switzerland×5+VALong×3	72.06 (560)	**76.92** (539)
No. of wins	1	1

Table 7 shows the accuracy of two different decision trees when smaller CF value was given for C4.5 to generate similarly sized tree to that of CART.

Table 7. The accuracy of C4.5 having similar tree size with CART for integrated data sets of heart disease in 10-fold cross validation

Data set	Accuracy of C4.5	Accuracy of CART
Cleveland×1+Hungarian×1+Switzerland×1+VALong×1	**49.78** (33, CF=4%)	46.52 (29)
Cleveland×2+Hungarian×2+Switzerland×5+VALong×3	72.06 (560, CF=default)	**76.92** (539)
No. of wins	1	1

As the second domain of integration, the two liver disease data sets were integrated. Bupa liver disorder data set and ILPD(Indian liver patient data set) have common domain of liver disorder disease, and have some common attributes. The class value has opposite meaning in the two data sets. There are some missing values. Please see table 8 for details of the attributes.

Table 8. The details of attributes of Bupa and Indian liver disorder data set

Data set	Attribute	Meaning
ILPD	Age	Age of patient
ILPD	Gender	Gender
ILPD, Bupa	alkphos	Alkaline phosphtase
ILPD, Bupa	Sgpt	Alamine aminotransferase

ILPD, Bupa	Sgot	Aspartate aminotransferase
ILPD	TB	Total bilirubin
ILPD	DB	Direct Bilirubin
ILPD	TP	Total protains
ILPD	ALB	Albumin
ILPD	A/G ratio	Albumin and Globulin ratio
Bupa	mcv	mean corpuscular volume
Bupa	gammagt	gamma-glutamyl transpeptidase
Bupa	drinks	number of half-pint equivalents of alcoholic beverages drunk per day

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 compatibility.

The two data sets were integrated to make a decision tree. The three common attributes, alkpoos, sgpt, and sgot, have no missing values. The vacant values for uncommon attributes are left missing. Because 'Indian liver patient data set' has larger number of records (583) than 'liver disorders' data set (345), 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 in addition to simple combination of the two data sets. Table 9 shows the distribution of missing values for each attribute in the two different integration method.

Table 9. The percentage of missing values for liver disease data sets in two different integration methods

Attribute	bupa x 1 + ILPD x 1	bupa x 3 + ILPD x 2
Age	37	47
Gender	37	47
alkphos	0	0
Sgpt	0	0
Sgot	0	0
TB	37	47
DB	37	47
TP	37	47
ALB	37	47
A/G ratio	38	47
mcv	63	53
gammagt	63	53
drinks	63	53
Average	34.54	37.54

Table 10 has the result of experiment.

Table 10. The accuracy of C4.5 and CART for integrated data sets of liver disease in 10-fold cross validation

Data set	Accuracy	
	C4.5	CART
bupa x 1 + ILPD x 1	**67.24** (29)	60.24 (13)
bupa x 3 + ILPD x 2	84.23 (275)	**90.14** (341)
No. of wins	1	1

Table 11 shows the accuracy of two different decision trees when smaller CF value was given for C4.5 to generate similarly sized tree to that of CART.

Table 11. The accuracy of C4.5 and CART for integrated data sets of liver disease in 10-fold cross validation

Data set	Accuracy	
	C4.5	CART
bupa x 1 + ILPD x 1	**67.24** (15, CF=5%)	60.24 (13)
bupa x 3 + ILPD x 2	86.05 (345, CF=70%)	**90.14** (341)
No. of wins	1	1

The bigger CF value than the default was given to generate larger tree than the tree of default CF value.

From the results in table 6 and table 7of heart disease data sets and table 10 and table 11 of liver disease data sets, we can say that CART may be better when we have lots of missing values, because the second data set has much more missing values that the first data set.

III. CONCLUSIONS

Decision trees can be very useful data mining tools for medicine domain, because the knowledge structure is represented in tree shape so that human expert could interpret the data well for more accurate diagnosis and better understanding on major factors in diagnosing the disease. There are several decision tree algorithms, and among them C4.5 and CART may be the most favored decision tree algorithms, because some survey in ICDM'06 showed that the two algorithms were elected as one of the top 10 algorithms. The two algorithms have been used in different application domains. According to literature survey C4.5 has been used in some wide range of area like engineering and financial domain, while CART has been favored mostly in medicine domain. In this paper, thirteen different data sets in medicine domain were experimented to compare the performance of the two algorithms for data sets in medicine domain. From the result of experiment, we can see that the performance of CART is somewhat inferior to that of C4.5. Moreover, for the cases CART's performance is better, the accuracy difference between the two algorithms is relatively small. So, we can say that C4.5 algorithm may generate better results than CART in more cases.

Integrating several data sets from different sources is a major task in data mining for more significant discovery, but it may leave missing values. In this respect, CART could be a useful tool, because CART has an elaborate technique for the missing values called surrogates. From the result of two different integrated data sets of heart and liver disease, we can see that CART could generate better results when we have lots of missing values.

REFERENCES

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

[2] 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.

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

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

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

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

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

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

[9] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, "Top 10 Algorithms in Data Mining," *Knowledge and Information Systems*, vol. 14, no. 1, 2006, pp.1-37.

[10] S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, 2007, pp. 249-268.

[11] Z. Chang, "The application of C4.5 algorithm based on SMOTE in financial distress prediction model," in *Proceedings of 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC)*, 2011, pp. 5852-5855.

[12] S. Gao, "The Analysis and Application of the C4.5 Algorithm in Decision Tree Technology," *Advanced Materials Research*, vol. 457-458, 2012, pp.754-757.

[13] R.J. Lewis, *An Introduction to Classification and Regression Tree (CART) Analysis*, Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.4103&rep=rep1&type=pdf

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

[15] CART Classification and Regression Tree. Available: http://www.salford-systems.com/products/cart

[16] Dermatology data set. Available: http://archive.ics.uci.edu/ml/datasets/Dermatology

[17] Parkinsons Telemonitoring data set. Available : http://archive.ics.uci.edu/ml/datasets/Parkinsons+Telemonitoring

[18] M. Lenzerini, "Data Integration: A Theoretical Perspective," in *Symposium on Principles of Database Systems*, 2002.

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 a 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.