# Better Decision Tree of Accuracy for the Task of Diagnosis of Hepatitis

Hyontai Sug

***Abstract*—**Hepatitis is a liver disease that can be self-limiting or can cause even death of the patients so that diagnosing the disease correctly is important. As a process of diagnosing the disease understandability of the diagnosed results may be important, because it is related to human life and final decision is attributed to doctors. Decision trees are one of data mining algorithms that can generate understandable knowledge structure which is in tree shape so that the algorithms have been used widely in medicine domain. But, because the algorithms give higher priority to major classes for better accuracy, this may cause poorer results in classification of minor classes. Over-sampling for a minor class has been considered a possible solution for the problem to get better results. But, even though we use the technique, there is innate property in the data and data mining algorithm themselves, which hinders data mining task. If we build a decision tree using a training data set, some data instances are classified wrongly, and these instances may cause lower accuracy. In order to avoid such instances decision tree with higher confidence is used to check each training instances in the minor class, and good ones only are adopted in the later over-sampling process. Experiments using hepatitis data set in various over-sampling rates showed very good results.

***Keywords*—**Decision trees, data mining, over-sampling, hepatitis, minor class.

## I. Introduction

HEPATITIS is a liver disease that can be self-limiting or can progress up to liver cancer, which may cause even death of the patients. Hepatitis viruses are the most common cause of hepatitis in the world. There are 5 types of main hepatitis viruses; type A, B, C, D, and E. Among them type B and C lead to chronic liver disease, and the two types are the most common cause of cirrhosis and liver cancer. Therefore, it's important to diagnose a patient who has hepatitis correctly. As a way to diagnose the disease data mining algorithms are adopted and widely used. Among many data mining algorithms, decision trees are one of the most important data mining algorithms especially in medicine domain [1, 2, 3, 4]. The reason why decision trees are widely adopted in medicine domain is that their structures are easily understandable by human. But, decision tree algorithms have the property of preferring major classes to achieve the best accuracy in classification. Major classes are classes that have more data instances that belong to the classes. But, in real world applications a minor class can be more important than the other classes, because we are often interested in more accurate classification of rare cases [5]. Therefore, increasing the number of instances in the minor class may make the algorithms to give more emphasis on the class. As a way to increase the number of instances in a minor class, over-sampling is common strategy. But, simple over-sampling may have limited effect only, because the same instances are supplied multiple times. On the other hand, we may supply similar and new data instances of the minor class. SMOTE algorithm is one of the representative over-sampling method of such kind. SMOTE stands for Synthetic Minority Over-sampling TEchnique. It selects K-nearest neighbors, and generates a synthetic data instance based on the neighbors. Success was reported for a decision tree algorithm and rule generator [6]. On the other hand, because we are interested in generating a better decision tree that can handle the minor class of hepatitis data well, we may check the appropriateness of each instances in the minor class using a decision tree itself, and we want to get better results by using the select instances for our over-sampling. In section 2 related work is provided, in section 3 we discuss our experiment method, and in section 4 conclusions are provided.

H. Sug is with the Division of Computer Engineering, Dongseo University, Busan, 617-716 Korea (phone: +82-51-320-1733; fax: +82-51-327-8955; e-mail: sht@ gdsu.dongseo.ac.kr). .

## II. RELATED WORK

Decision tree algorithms belong to the class of greedy algorithms, because branches of decision trees are built based on some heuristic functions that are believed to choose the best root of each subtree. There have been a lot of efforts to build best decision trees [7]. Among them C4.5 [8] and CART [9] may be two representatives, because the two algorithms are frequently referred [10]. While C4.5 uses an entropy-based measure to split branches, CART uses a purity-based measure. Either way the splitting measure of the decision tree algorithms prefers the most certain split among possible splits from candidates. Therefore, decision trees prefer major classes to minor classes. In order to avoid such property random forests were suggested [11]. Random forests use many trees to classify new instances. The corresponding many training data sets needed are prepared by random sampling with replacement method. According to the report random forests may generate more accurate classifiers. A weak point of random forests is that they are not easy to understand because of the many number of decision trees in them.

Diagnosing hepatitis accurately has been major concern since

the related data set has been available in public [12]. The researchers in [12] used multiple logistic regressions and bootstrap method, and achieved accuracy of 84%. In [13] principle component analysis was used to select appropriate attributes, and logistic regression was used after the selection achieving accuracy of 89.6%. In [14] a clustering method called CBR-PSO (Case Based Reasoning - Particle Swarm Optimization) was used, achieving the average accuracy of 92.83%. All the previously referred researches were focusing on achieving higher accuracy so that understandability on the final results are limited. On the other hand some other researchers prefer decision tree to analyze the data set. In [15] researchers showed the change of accuracy of decision tree by changing number of attributes and number of training instances as well as pruning confidences. In [16, 17, 18] authors suggested using decision trees to analyze hepatitis data for better understanding of the result.

## III. EMPIRICAL PROCEDURE

### A. Experiment Method

Because we are interested in generating a data mining model of understandability, we use decision tree algorithm C4.5. According to survey performed at ICDM 2006, C4.5 is one of the most frequently used data mining algorithms [10]. For comparison of our method and the other representative over-sampling algorithm, we choose SMOTE. SMOTE generates synthetic data instances of a minor class as a way of over-sampling to build better decision trees of C4.5.

In the following experiments, we first check the appropriateness of each data instance of the minority class by the decision tree of original data. After the checking we select instances that are classified correctly, and do over-sampling in various rates using the select instances. Over-sampling based on SMOTE algorithm will be performed to compare with our method. Finally conventional over-sampling method will be performed to compare with our method as well. Experiments will be performed using a medicine data set called hepatitis in the UCI machine learning repository [19]. The experiment will be based on 10-fold cross validation. The following is the procedure of the over-sampling.

**INPUT**: hepatitis data set
**OUTPUT**: decision trees
**Begin**
For ss = 500 to 2500 step 500
  Do over-sampling rate of ss%;
  Generate decision tree and evaluate;
End for;
Find the decision tree that has the best confusion matrix;
Let the best one's ss be bss;
For ss = bss to bss-400 if bss >= 1000 step -100
  Do over-sampling rate of ss%;
  Generate decision tree and evaluate;
End for;
For ss = bss to bss+400 if bss >= 1000 step 100
  Do over-sampling rate of ss%;
  Generate decision tree and evaluate;
End for;
**End**.

### B. Hepatitis Data Set

Hepatitis data set contains two classes. The first class which means 'die' has 32 instances, and the second class which means 'live' has 123 instances. So, the class having 32 instances is minor class. There are nineteen conditional attributes, and among them six are continuous attributes, and the others are nominal attributes. Table 1 shows the names and domains of the attributes. The data set contains missing values in some attributes.

TABLE I
ATTRIBUTE INFORMATION

| Attribute name | values |
|---|---|
| Class | die, live |
| Age | real |
| Sex | male, female |
| Steroid | no, yes |
| Antivirals | no, yes |
| Fatigue | no, yes |
| Malaise | no, yes |
| Anorexia | no, yes |
| Liver big | no, yes |
| Liver firm | no, yes |
| Spleen palpable | no, yes |
| Spiders | no, yes |
| Ascites | no, yes |
| Varices | no, yes |
| Bilirubin | real |
| Alk phosphate | real |
| SGOT | real |
| Albumin | real |
| Protime | real |
| Histology | no, yes |

Table 2 shows the accuracy of decision tree algorithm, C4.5 for the data set with default parameters. The decision tree shows similar accuracy with the logistic regression based method in [12].

TABEL II
ACCURACY OF THE DECISION TREE

| | | C4.5 |
|---|---|---|
| Accuracy in % | | 83.871 |
| True Positive rate | Class 'DIE' | 0.438 |
| | Class 'LIVE' | 0.943 |

Table 3 shows the corresponding confusion matrix.

TABLE III
CONFUSION MATRIX OF DECISION TREE

| | | Predicted class | |
|---|---|---|---|
| | | Class 'DIE' | Class 'LIVE' |
| Actual class | Class 'DIE' | 14 | 18 |
| | Class 'LIVE' | 7 | 116 |

The following is the generated tree.

```
ASCITES = no
|  ALBUMIN <= 2.8: die (9.19/0.06)
|  ALBUMIN > 2.8
|  |  LIVER_FIRM: = no: live (2.51/0.22)
|  |  LIVER_FIRM: = yes
|  |  |  ALBUMIN <= 2.9: live (2.15)
|  |  |  ALBUMIN > 2.9: die (6.81/2.03)
|  |  LIVER_FIRM: = ?: live (0.0)
ASCITES = yes
|  SPIDERS = no
|  |  SEX = no
|  |  |  LIVER_FIRM: = no
|  |  |  |  SGOT <= 101: live (11.63/0.36)
|  |  |  |  SGOT > 101
|  |  |  |  |  LIVER_BIG = no: die (3.23/0.08)
|  |  |  |  |  LIVER_BIG = yes: live (7.54/2.36)
|  |  |  |  |  LIVER_BIG = ?: die (0.0)
|  |  |  LIVER_FIRM: = yes
|  |  |  |  AGE <= 40: live (4.15/1.0)
|  |  |  |  AGE > 40: die (5.45/0.07)
|  |  |  LIVER_FIRM: = ?: live (0.0)
|  |  SEX = female: live (6.25)
|  |  SEX = ?: live (0.0)
|  SPIDERS = yes: live (96.1/5.62)
|  SPIDERS = ?: live (0.0)
ASCITES = ?: live (0.0)
```

The value '?' in the decision tree indicates missing value, and the values in the parentheses represent the average number of correctly classified instances and the average number of misclassified instances for a leaf. The number of leaves in the tree is 17, and the size of the tree is 27.

*1) Over-sampling in suggested method*

Because the data set may contain some instances that may not be good for better classification using the decision tree algorithm, the minor class instances are checked by decision tree that can be trained by the original data set. The decision tree was trained with the pruning confidence of 50%, while the default confidence is 25%. Larger confidence applies less pruning so that the resulting tree reflects the property of data set more. Among 32 instances in the minor class, four instances are classified as having wrong class value, so 28 of them are selected as candidates for over-sampling. Over-sampling using the select instances only for various rates is performed to find the best decision tree. Table 4 shows the results. The over-sampled instances and the original data set are mixed together to make the final training sets. 10-fold cross validation is used also to test.

TABLE IV
ACCURACY OF DECISION TABLE ON VARIOUS OVER-SAMPLING RATE

| Over-sampling rate of select instances in the minor class in % | Accuracy in % | True positive rate | Confusion matrix | |
|---|---|---|---|---|
| 500 | 91.1864 | 0.988 | 170 | 2 |
| | | 0.805 | 24 | 99 |
| 600 | 93.808 | 0.985 | 197 | 3 |
| | | 0.862 | 17 | 106 |
| 700 | 93.7322 | 0.987 | 225 | 3 |
| | | 0.846 | 19 | 103 |
| 800 | 94.7230 | 0.988 | 253 | 3 |
| | | 0.862 | 17 | 106 |
| **900** | 95.5774 | 0.993 | 282 | 2 |
| | | 0.870 | 16 | 107 |
| 1000 | 96.0920 | 0.990 | 309 | 3 |
| | | 0.886 | 14 | 109 |
| 1100 | 95.8963 | 0.991 | 337 | 3 |
| | | 0.870 | 16 | 107 |
| 1200 | 96.1303 | 0.992 | 365 | 3 |
| | | 0.870 | 16 | 107 |
| **1300** | 97.3025 | 0.992 | 393 | 3 |
| | | 0.911 | 11 | 112 |
| 1400 | 96.8921 | 0.993 | 421 | 3 |
| | | 0.886 | 14 | 109 |
| 1500 | 95.8261 | 0.993 | 449 | 3 |
| | | 0.829 | 21 | 102 |
| 2000 | 97.2028 | 0.995 | 559 | 3 |
| | | 0.862 | 17 | 106 |
| 2500 | 97.6608 | 0.996 | 729 | 3 |
| | | 0.862 | 17 | 106 |

According to the experiment, over-sampling rate of 1300% shows the best result in the confusion matrix. The following is the resulting decision tree. If we use the 14 misclassified instances to estimate error rate using the original 155 data instances, it's about 9%. The error rate is comparable to the best result yet achieved using a special clustering method [14]. But, our decision tree is more understandable.

```
AGE <= 28: live (25.0)
AGE > 28
|  FATIGUE = no
|  |  SEX = male
|  |  |  ALBUMIN <= 3.8
|  |  |  |  HISTOLOGY = no
|  |  |  |  |  BILIRUBIN <= 1.8
|  |  |  |  |  |  AGE <= 50: live (6.99)
|  |  |  |  |  |  AGE > 50: die (3.56/1.0)
|  |  |  |  |  BILIRUBIN > 1.8
|  |  |  |  |  |  AGE <= 38
|  |  |  |  |  |  |  ANOREXIA = no: live (2.09)
```

```
| | | | | | | ANOREXIA = yes: die (12.99/0.59)
| | | | | | | ANOREXIA = ?: die (0.0)
| | | | | | AGE > 38: die (52.73)
| | | | HISTOLOGY = yes
| | | | | LIVER_BIG = no: live (4.74/0.86)
| | | | | LIVER_BIG = yes
| | | | | | PROTIME <= 48: die (269.84/1.71)
| | | | | | PROTIME > 48
| | | | | | | SPIDERS = no
| | | | | | | | BILIRUBIN <= 1.1: live (2.09/0.09)
| | | | | | | | BILIRUBIN > 1.1: die (22.73/0.09)
| | | | | | | SPIDERS = yes: live (2.17/0.09)
| | | | | | | SPIDERS = ?: die (0.0)
| | | | | LIVER_BIG = ?: die (0.0)
| | | | HISTOLOGY = ?: die (0.0)
| | | ALBUMIN > 3.8
| | | | BILIRUBIN <= 2.5
| | | | | SPLEEN_PALPABLE = no: die (3.82/1.2)
| | | | | SPLEEN_PALPABLE = yes: live (27.29/2.13)
| | | | | SPLEEN_PALPABLE = ?: live (0.0)
| | | | BILIRUBIN > 2.5: die (17.84/1.08)
| | SEX = female: live (9.0)
| | SEX = ?: die (0.0)
| FATIGUE = yes
| | VARICES = no: die (15.03/1.03)
| | VARICES = yes: live (41.08/1.0)
| | VARICES = ?: live (0.0)
| FATIGUE = ?: die (0.0)
```

The number of leaves in the tree is 25, and the size of the tree is 41. Note that the number of leaves is 17, and the size of the tree is 27 in the decision tree from the original data set. So, the two trees have similar sizes, while the number of misclassified instances of 'die' reduces from 18 to 3, and the number of misclassified instances of 'live' increases from 7 to 11. Therefore, we have much gain in predicting 'die' while small loss in predicting 'live'.

Another decision tree that is based on the sampling rate of 900 is as follows. The decision tree has the lowest number of misclassification on 'die' which is 2.

```
AGE <= 28: live (25.0)
AGE > 28
| FATIGUE = no
| | SEX = male
| | | ALBUMIN <= 3.8
| | | | HISTOLOGY = no
| | | | | BILIRUBIN <= 1.8
| | | | | | AGE <= 50: live (6.99)
| | | | | | AGE > 50: die (3.39/1.0)
| | | | | BILIRUBIN > 1.8
| | | | | | AGE <= 38
| | | | | | | ANOREXIA = no: live (2.12)
| | | | | | | ANOREXIA = yes: die (9.12/0.5)
| | | | | | | ANOREXIA = ?: die (0.0)
| | | | | | AGE > 38: die (37.1)
| | | | HISTOLOGY = yes
| | | | | LIVER_BIG = no: live (4.69/0.83)
| | | | | LIVER_BIG = yes
| | | | | | PROTIME <= 48: die (192.28/1.68)
| | | | | | PROTIME > 48
| | | | | | SPIDERS = no
| | | | | | | BILIRUBIN <= no.1: live (2.09/0.09)
| | | | | | | BILIRUBIN > 1.1: die (16.72/0.1)
| | | | | | SPIDERS = yes: live (2.18/0.1)
| | | | | | SPIDERS = ?: die (0.0)
| | | | | LIVER_BIG = ?: die (0.0)
| | | | HISTOLOGY = ?: die (0.0)
| | | ALBUMIN > 3.8
| | | | BILIRUBIN <= 2.5
| | | | | SPLEEN_PALPABLE = no: die (3.56/1.22)
| | | | | SPLEEN_PALPABLE = yes: live (27.33/2.1)
| | | | | SPLEEN_PALPABLE = ?: live (0.0)
| | | | BILIRUBIN > 2.5: die (13.28/1.08)
| | SEX = female: live (9.0)
| | SEX = ?: die (0.0)
| FATIGUE = yes
| | VARICES = no: die (11.03/1.03)
| | VARICES = yes: live (41.11/1.0)
| | VARICES = ?: live (0.0)
| FATIGUE = ?: die (0.0)
```

The number of leaves in the tree is 25, and the size of the tree is 41. Note that the number of leaves is 17, and the size of the tree is 27 in the decision tree from the original data set. So, the two trees have similar size, while the number of misclassified instances of 'die' reduces from 18 to 2, and the number of misclassified instances of 'live' increases from 7 to 16.

*2) Over-sampling in SMOTE algorithm*

Another set of experiment based on well-known over-sampling algorithm, SMOTE, for comparison. In order to generate equal number of training instances with the experiment in table 4, appropriate over-sampling rates using SMOTE and default parameters are used. Table 5 shows the result of experiment for various over-sampling rates. The numbers in parentheses indicate the corresponding over-sampling rate label in table 4. Note that the number of instances in the minor class is 32 in the original data set, while the number of select instances is 28 that is the base of the over-sampling in table 4.

TABLE V
ACCURACY OF DECISION TABLE ON VARIOUS OVER-SAMPLING RATE WITH SMOTE

| Over-sampling rate in the minor class in % | Accuracy in % | True positive rate | Confusion matrix | |
|---|---|---|---|---|
| 437.5 (500) | 88.1396 | 0.913 | 157 | 15 |
| | | 0.837 | 20 | 103 |
| 525 (600) | 90.4025 | 0.925 | 185 | 15 |
| | | 0.870 | 16 | 107 |
| 612.5 (700) | 89.7436 | 0.930 | 212 | 16 |
| | | 0.837 | 20 | 103 |
| 700 (800) | 90.7652 | 0.938 | 240 | 16 |
| | | 0.846 | 19 | 104 |
| 787.5 (900) | 90.9091 | 0.944 | 268 | 16 |

| | | | | |
|---|---|---|---|---|
| | | 0.829 | 21 | 102 |
| 875 (1000) | 92.4138 | 0.952 | 297 | 15 |
| | | 0.854 | 18 | 105 |
| 962.5 (1100) | 92.8726 | 0.953 | 324 | 16 |
| | | 0.862 | 17 | 106 |
| 1050 (1200) | 93.2790 | 0.957 | 352 | 16 |
| | | 0.862 | 17 | 106 |
| **1137.5** (1300) | 93.8343 | 0.965 | 382 | 14 |
| | | 0.854 | 18 | 105 |
| 1225 (1400) | 94.1449 | 0.967 | 410 | 14 |
| | | 0.854 | 18 | 105 |
| **1312.5** (1500) | 94.4348 | 0.971 | 439 | 13 |
| | | 0.846 | 19 | 104 |
| 1750 (2000) | 95.1049 | 0.975 | 577 | 15 |
| | | 0.837 | 20 | 103 |
| 2187.5 (2500) | 95.7895 | 0.978 | 716 | 16 |
| | | 0.837 | 20 | 103 |

According to the experiment, over-sampling rate of 1137.5% that corresponds to over-sampling rate of 1300% in our method shows good result in the confusion matrix. The following is the resulting decision tree.

```
MALAISE = no
|  LIVER_BIG = no
|  |  SGOT <= 125.319394: live (9.12/0.07)
|  |  SGOT > 125.319394: die (2.06/0.03)
|  LIVER_BIG = yes
|  |  SPIDERS = no
|  |  |  STEROID = no: die (333.44/5.6)
|  |  |  STEROID = yes
|  |  |  |  ASCITES = no
|  |  |  |  |  BILIRUBIN <= 1.004862: live (3.74/1.0)
|  |  |  |  |  BILIRUBIN > 1.004862: die (27.0)
|  |  |  |  ASCITES = yes: live (4.1)
|  |  |  |  ASCITES = ?: die (0.0)
|  |  |  STEROID = ?: die (0.0)
|  |  SPIDERS = yes
|  |  |  HISTOLOGY = no: live (17.28/1.97)
|  |  |  HISTOLOGY = yes: die (29.08/2.0)
|  |  |  HISTOLOGY = ?: die (0.0)
|  |  SPIDERS = ?: die (0.0)
|  LIVER_BIG = ?: die (0.0)
MALAISE = yes
|  ASCITES = no: die (4.05/1.05)
|  ASCITES = yes
|  |  SPIDERS = no
|  |  |  BILIRUBIN <= 1.8: live (13.61/1.79)
|  |  |  BILIRUBIN > 1.8: die (3.61/0.4)
|  |  SPIDERS = yes: live (71.91/1.0)
|  |  SPIDERS = ?: live (0.0)
|  ASCITES = ?: live (0.0)
MALAISE = ?: die (0.0)
```

The number of leaves in the tree is 20, and the size of the tree is 31. Note that the number of leaves is 17, and the size of the tree is 27 in the decision tree from the original data set. So, the two trees have similar size, while the number of misclassified instances of 'die' reduces from 18 to 14, and the number of misclassified instances of 'live' increases from 7 to 18. Note also that in our method with over-sampling rate 1300% on the select instances in minor class the number of leaves in the tree is 25, and the size of the tree is 41. So, the trees with our method and SMOTE have similar size, but the number of misclassified instances of 'die' reduces from 14 to 3, and the number of misclassified instances of 'live' reduces from 18 to 11 compared to the decision tree of the original data. Therefore, our method generated much better result in predicting 'die' while smaller loss in predicting 'live'.

Another decision tree of SMOTE that is based on the sampling rate of 1312.5% that corresponds to over-sampling rate of 1500% in our method is as follows. The decision tree has the lowest number of misclassification on 'die' which is 13.

```
MALAISE = no
|  SEX = male
|  |  ANTIVIRALS = no
|  |  |  PROTIME <= 46: die (2.29/0.29)
|  |  |  PROTIME > 46: live (5.71)
|  |  ANTIVIRALS = yes
|  |  |  LIVER_BIG = no
|  |  |  |  SGOT <= 125.151899: live (5.07/0.04)
|  |  |  |  SGOT > 125.151899: die (2.04/0.02)
|  |  |  LIVER_BIG = yes
|  |  |  |  AGE <= 34.500558
|  |  |  |  |  AGE <= 28: live (6.98)
|  |  |  |  |  AGE > 28: die (16.83/2.83)
|  |  |  |  AGE > 34.500558
|  |  |  |  |  SPIDERS = no: die (398.79/3.9)
|  |  |  |  |  SPIDERS = yes
|  |  |  |  |  |  ANOREXIA = no: live (4.17/0.08)
|  |  |  |  |  |  ANOREXIA = yes
|  |  |  |  |  |  |  LIVER_FIRM: = no: die (28.72)
|  |  |  |  |  |  |  LIVER_FIRM: = yes: live (4.25/1.25)
|  |  |  |  |  |  |  LIVER_FIRM: = ?: die (0.0)
|  |  |  |  |  |  ANOREXIA = ?: die (0.0)
|  |  |  |  |  SPIDERS = ?: die (0.0)
|  |  |  LIVER_BIG = ?: die (0.0)
|  |  ANTIVIRALS = ?: die (0.0)
|  SEX = female: live (7.0)
|  SEX = ?: die (0.0)
MALAISE = yes
|  ASCITES = no: die (4.05/1.05)
|  ASCITES = yes
|  |  SPIDERS = no
|  |  |  BILIRUBIN <= 1.8: live (13.6/1.79)
|  |  |  BILIRUBIN > 1.8: die (3.61/0.4)
|  |  SPIDERS = yes: live (71.9/1.0)
|  |  SPIDERS = ?: live (0.0)
|  ASCITES = ?: live (0.0)
```

MALAISE = ?: die (0.0)

The number of leaves in the tree is 24, and the size of the tree is 38. Note that the number of leaves is 17, and the size of the tree is 27 in the decision tree from the original data set. So, the two trees have similar size, while the number of misclassified instances of 'die' reduces from 18 to 13, and the number of misclassified instances of 'live' increases from 7 to 19. Note also that in our method with over-sampling rate 900% on the select instances of minor class that has the lowest number of misclassified instances of 'die' which is 2 the number of leaves in the tree is 25, and the size of the tree is 41. So, the trees with our method and SMOTE have similar size, but the number of misclassified instances of 'die' reduces from 13 to 2, and the number of misclassified instances of 'live' reduces from 19 to 16. Therefore, our method generated much better result in predicting 'die' while smaller loss in predicting 'live'. Table 6 and 7 contain the summery of the comparisons.

TABLE VI
COMPARISON OF OUR METHOD AND SMOTE FOR THE CASE OF THE SAME SAMPLE SIZE IN OVER-SAMPLING RATE OF 1300%

| | Our method | SMOTE | Original data |
|---|---|---|---|
| The number of 'die' instances | 396 | 396 | 32 |
| The number of 'live' instances | 123 | 123 | 123 |
| Total number of training instances | 519 | 519 | 155 |
| Accuracy in % | 97.3025 | 93.8343 | 83.8710 |
| The total number of misclassified 'die' instances | 3 | 14 | 18 |
| The total number of misclassified 'live' instances | 11 | 18 | 7 |
| Total number of misclassified instances | 14 | 32 | 25 |

TABLE VII
COMPARISON OF OUR METHOD AND SMOTE FOR THE CASE OF THE SMALLEST NUMBER OF MISCLASSIFICATION IN 'DIE' CLASS

| | Our method | SMOTE | Original data |
|---|---|---|---|
| The number of 'die' instances | 284 | 452 | 32 |
| The number of 'live' instances | 123 | 123 | 123 |
| Total number of training instances | 407 | 575 | 155 |
| Accuracy in % | 95.5774 | 94.4348 | 83.8710 |
| The total number of misclassified 'die' instances | 2 | 13 | 18 |
| The total number of misclassified 'live' instances | 16 | 19 | 7 |
| Total number of misclassified instances | 18 | 32 | 25 |

*3) Over-sampling in conventional method*

Another set of experiments based on conventional over-sampling method were performed for comparison. So, there is no select procedure before over-sampling. In order to generate equal number of training instances with the experiment in table 4, appropriate over-sampling rates are applied like in SMOTE. Table 7 shows the result of experiment for various over-sampling rates. The numbers in parentheses indicates corresponding sampling rate label in table 4. Note that the number of instances in the minor class is 32 in the original data set.

TABLE VII
ACCURACY OF DECISION TREE ON CONVENTIONAL OVER-SAMPLING METHOD WITH VARIOUS RATES

| Over-sampling rate in the minor class in % | Accuracy in % | True positive rate | Confusion matrix | |
|---|---|---|---|---|
| 437.5 (500) | 92.5424 | 0.994 | 171 | 1 |
| | | 0.829 | 21 | 102 |
| **525** (600) | 93.4395 | 1.0 | 200 | 0 |
| | | 0.829 | 21 | 102 |
| 612.5 (700) | 94.0171 | 1.0 | 228 | 0 |
| | | 0.829 | 21 | 102 |
| 700 (800) | 93.1398 | 1.0 | 256 | 0 |
| | | 0.789 | 26 | 97 |
| 787.5 (900) | 93.8575 | 1.0 | 284 | 0 |
| | | 0.797 | 25 | 98 |
| 875 (1000) | 93.7931 | 1.0 | 312 | 0 |
| | | 0.780 | 27 | 96 |
| 962.5 (1100) | 94.3844 | 1.0 | 340 | 0 |
| | | 0.789 | 26 | 97 |
| **1050** (1200) | 95.9267 | 1.0 | 368 | 0 |
| | | 0.837 | 20 | 103 |
| 1137.5 (1300) | 93.6414 | 1.0 | 396 | 0 |
| | | 0.732 | 33 | 90 |
| 1225 (1400) | 94.6984 | 1.0 | 424 | 0 |
| | | 0.764 | 29 | 94 |
| 1312.5 (1500) | 93.5652 | 1.0 | 452 | 0 |
| | | 0.699 | 37 | 86 |
| 1750 (2000) | 95.3846 | 1.0 | 592 | 0 |
| | | 0.732 | 33 | 90 |
| 2187.5 (2500) | 96.0234 | 1.0 | 732 | 0 |
| | | 0.724 | 34 | 89 |

According to the experiment, over-sampling rate of 1050% that corresponds to over-sampling rate of 1200% in our method

shows the best result in the confusion matrix. The following is the resulting decision tree.

```
AGE <= 28: live (25.0)
AGE > 28
|  FATIGUE = no
|  |  SEX = male
|  |  |  PROTIME <= 50
|  |  |  |  ASCITES = no: die (150.27/1.71)
|  |  |  |  ASCITES = yes
|  |  |  |  |  SGOT <= 54
|  |  |  |  |  |  ALK_PHOSPHATE <= 71: die (13.61/1.01)
|  |  |  |  |  |  ALK_PHOSPHATE > 71: live (9.82/1.22)
|  |  |  |  |  SGOT > 54
|  |  |  |  |  |  SPLEEN_PALPABLE = no: die (85.9/1.28)
|  |  |  |  |  |  SPLEEN_PALPABLE = yes
|  |  |  |  |  |  |  ANTIVIRALS = no: die (21.0)
|  |  |  |  |  |  |  ANTIVIRALS = yes
|  |  |  |  |  |  |  |  LIVER_BIG = no: die (43.97/1.99)
|  |  |  |  |  |  |  |  LIVER_BIG = yes
|  |  |  |  |  |  |  |  |  AGE <= 56: live (5.1)
|  |  |  |  |  |  |  |  |  AGE > 56: die (3.69)
|  |  |  |  |  |  |  |  |  LIVER_BIG = ?: die (0.0)
|  |  |  |  |  |  |  ANTIVIRALS = ?: die (0.0)
|  |  |  |  |  |  SPLEEN_PALPABLE = ?: die (0.0)
|  |  |  |  ASCITES = ?: die (0.0)
|  |  |  PROTIME > 50
|  |  |  |  SPIDERS = no
|  |  |  |  |  ANTIVIRALS = no: live (2.16)
|  |  |  |  |  ANTIVIRALS = yes
|  |  |  |  |  |  BILIRUBIN <= 1.1
|  |  |  |  |  |  |  LIVER_FIRM: = no: live (4.57/0.31)
|  |  |  |  |  |  |  LIVER_FIRM: = yes: die (2.82/0.21)
|  |  |  |  |  |  |  LIVER_FIRM: = ?: live (0.0)
|  |  |  |  |  |  BILIRUBIN > 1.1: die (30.56/1.68)
|  |  |  |  |  ANTIVIRALS = ?: die (0.0)
|  |  |  |  SPIDERS = yes
|  |  |  |  |  ALBUMIN <= 3: die (2.27/0.2)
|  |  |  |  |  ALBUMIN > 3: live (21.12/1.44)
|  |  |  |  SPIDERS = ?: die (0.0)
|  |  SEX = female: live (9.0)
|  |  SEX = ?: die (0.0)
|  FATIGUE = yes
|  |  VARICES = no: die (14.03/1.03)
|  |  VARICES = yes
|  |  |  BILIRUBIN <= 0.5: die (6.29/0.29)
|  |  |  BILIRUBIN > 0.5: live (39.81)
|  |  VARICES = ?: live (0.0)
|  FATIGUE = ?: die (0.0)
```

The number of leaves in the tree is 29, and the size of the tree is 47. Note that the number of leaves is 17, and the size of the tree is 27 in the decision tree from the original data set. So, the over-sampled tree is larger in size, while the number of misclassified instances of 'die' reduces from 18 to 0, and the number of misclassified instances of 'live' increases from 7 to 20. Note also that in our method with over-sampling rate 1300% on the select instances of minor class the number of leaves in the tree is 25, and the size of the tree is 41. So, the trees with our method is middle size, but the number of misclassified instances of 'die' reduces from 14 to 3, and the number of misclassified instances of 'live' decreases from 18 to 11 compared to the decision tree from the original data set. Therefore, our method generated much better result with respect to the total number of misclassification (14:20=our method: conventional method). Table 8 summarizes the two result.

TABLE VIII
COMPARISON OF OUR METHOD AND CONVENTIONAL METHOD OF OVER-SAMPLING FOR THE CASE OF THE SMALLEST NUMBER OF TOTAL MISCLASSIFICATION

| | Our method | Conventional method | Original data |
|---|---|---|---|
| The number of 'die' instances | 396 | 368 | 32 |
| The number of 'live' instances | 123 | 123 | 123 |
| Total number of training instances | 519 | 491 | 155 |
| Accuracy in % | 97.3025 | 95.9267 | 83.8710 |
| The total number of misclassified 'die' instances | 3 | 0 | 18 |
| The total number of misclassified 'live' instances | 11 | 20 | 7 |
| Total number of misclassified instances | 14 | 20 | 25 |

Another decision tree that is based on the sampling rate of 525% that corresponds to over-sampling rate of 600% in our method is as follows. The decision tree has smaller over-sampling rate and the lowest number of misclassification on 'die' which is 0.

```
AGE <= 28: live (25.0)
AGE > 28
|  FATIGUE = no
|  |  SEX = male
|  |  |  PROTIME <= 50
|  |  |  |  ASCITES = no: die (79.5/1.62)
|  |  |  |  ASCITES = yes
|  |  |  |  |  SGOT <= 54
|  |  |  |  |  |  ALK_PHOSPHATE <= 71: die (4.98/0.53)
|  |  |  |  |  |  ALK_PHOSPHATE > 71: live (9.6/1.28)
|  |  |  |  |  SGOT > 54
|  |  |  |  |  |  SPLEEN_PALPABLE = no: die (40.62/1.1)
|  |  |  |  |  |  SPLEEN_PALPABLE = yes
|  |  |  |  |  |  |  ANTIVIRALS = no: die (11.0)
|  |  |  |  |  |  |  ANTIVIRALS = yes
```

```
| | | | | | | | AGE <= 53
| | | | | | | | | AGE <= 35: die (6.19/0.14)
| | | | | | | | | AGE > 35: live (6.26)
| | | | | | | | AGE > 53: die (17.91)
| | | | | | | ANTIVIRALS = ?: die (0.0)
| | | | | | | SPLEEN_PALPABLE = ?: die (0.0)
| | | | ASCITES = ?: die (0.0)
| | | PROTIME > 50
| | | | SPIDERS = no
| | | | | BILIRUBIN <= 1.1: live (9.38/2.41)
| | | | | BILIRUBIN > 1.1: die (23.32/1.99)
| | | | SPIDERS = yes: live (24.04/3.17)
| | | | SPIDERS = ?: live (0.0)
| | SEX = female: live (9.0)
| | SEX = ?: die (0.0)
| FATIGUE = yes
| | VARICES = no: die (10.03/1.03)
| | VARICES = yes
| | | BILIRUBIN <= 0.5: die (6.29/0.29)
| | | BILIRUBIN > 0.5: live (39.86)
| | VARICES = ?: live (0.0)
| FATIGUE = ?: die (0.0)
```

The number of leaves in the tree is 23, and the size of the tree is 38. Note that the number of leaves is 17, and the size of the tree is 27 in the decision tree from the original data set. So, the over-sampled tree is larger in size, while the number of misclassified instances of 'die' reduces from 18 to 0, and the number of misclassified instances of 'live' increases from 7 to 21. Note also that in our method with over-sampling rate 900% on the select instances of minor class that has the lowest number of misclassified instances of 'die' which is 2 the number of leaves in the tree is 25, and the size of the tree is 41. So, if we compare the trees in our method and conventional method, the trees with our method of over-sampling rate of 900% and conventional over-sampling method of over-sampling rate of 525% have similar size, but the number of misclassified instances of 'die' increases from 0 to 2, and the number of misclassified instances of 'live' decreases from 21 to 16. Therefore, our method generated better result with respect to the total number of misclassification (18:21=our method: conventional method).

## IV. Conclusion

Decision tree are used for data mining task in medicine domain widely, because we can easily understand the result of data mining. But the algorithms may neglect data instances in minor class, because the minor class often does not have enough data instances for better classification. In order to surmount the problem, over-sampling may be used. Over-sampling technique based on synthetic data generation method like SMOTE or simple random over-sampling method have been considered a good technique for that purpose. But, those well-known techniques may not generate the best result, because each different data set may have different property that may cause results that may need more improvement in data mining. For the case of hepatitis data set, in order to get better decision tree we first select better data instances for our target decision tree. For that purpose we used decision tree to test each data instances in the minor class. The select data instances are used for over-sampling in various rates to find the best decision tree. Comparison experiments with conventional over-sampling as well as SMOTE over-sampling showed that our method is very effective.

## References


[1] L. Rokach, O. Maimon, *Data Mining with Decision Trees: Theories and Applications*, 2nd ed., World Scientific Publishing Company, 2014.

[2] J. Bae, "The clinical decision analysis using decision tree", *Epidemiol Health*, vol. 36, 2014, DOI: 10.4178/epih/e2014025.

[3] K. Madadipouya, "A new decision tree method for data mining in medicine", *Advanced Computational Intelligence: An International Journal*, vol. 2, no. 3, 2015, pp. 31-37.

[4] J. Li, A.W. Fu, H. He, J. Chen, H. Jin, D. McAullay, G. Williams, R. Sparks, C. Kelman, "Mining Risk Patterns in Medical Data," in *Proceedings of KDD 2005*, pp. 770-775.

[5] N.V. Chawla, "Data Mining for Imbalanced data Sets: an Overview", pp. 875-886, in *Data Mining and Knowledge Discovery Handbook*, 2nd ed., O. Maimon, L. Rokach, eds., Springer, 2010.

[6] N.V. Chawla, K.W. Dowyer, L. O. Hall, W. P. Kegelmeyer, , "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 321-357.

[7] L. Rokach, O. Maimon, "Top-down induction of decision trees classifiers – a survey", *IEEE Transactions on Systems, Man, and Cybernetics*, Part C, vol. 35, issue 4, 2005, pp. 476-487.

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

[9] W. Loh, "Classification and Regression Trees", *WIRE's Data Mining and Knowledge Discovery*, John Willey & Sons, Inc., vol. 1, 2011, pp. 14-23.

[10] 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 System*, vol. 14, 2008, pp.1-37.

[11] L. Breiman, "Random Forests", *Machine Learning*, vol. 45, issue 1, 2001, pp. 5-32.

[12] P. Diaconis, B. Efron, *Computer Intensive Methods in Statistics*, Technical Report No. 83, Division of Biostatistics, Stanford University, 1983.

[13] H. Yasin, T.A. Jilani, M. Danish, "Hepatis-C Classification using Data Mining Techniques", *International Journal of Computer Applications*, vol. 24, no. 3, 2011, pp. 1-6.

[14] M. Neshat, M. sargolzaei, A.N. Toosi, A. Masoumi, "Hapatitis Disease Diagnosis Using Hybrid Case Based Reasoning and Particle Swarm Optimization", *ISRN Artificial Intelligence*, vol. 2012, Article ID 609718.

[15] V.S. Sowmien, V. Sugumaran, C.P. Karthikeyan, T.T. Vijayaram, "Diagnosis of Hepatitis using Decision Tree Algorithm", *International Journal of Engineering and technology*, vol. 8, no. 3, 2016, pp. 1414-1419.

[16] K.P. Lokhande, A.P. Wadhe, "Survey on Data Mining Technique Using Decision Tree For Hepatitis Virus", *International Journal of Advanced Research in Computer Science*, vol. 4, no. 6, 2013, pp. 28-31.

[17] S. Hashem, G. Esmat, W. Elakel, S. Habashy, S.A. Raouf, S. Daweesh, M. Soliman, M. Elhefnawi, M. el-Adawy, M. Elhefnawi, "Accurate Prediction of Advanced Liver Fibrosis Using Decision Tree Learning Algorithm in Chronic Hepatitis C Egyptian Patients", *Gastroenterology Research and Practice*, 2016. DOI: 10.1155/2016/2636390.

[18] E. Audureau, V. Bourcier, R. Layese, C. Cagnot, P. Marcellin, D. Guyader, S. Pol, D. Larrey, F. Roudot-Thoraval, P. Nahon, "Identifying residual risk of hepatocellular carcinoma following hepatitis C virus eradication in compensated cirrhosis: decision-tree and random forest models developed in the French multicenter prospective ANRS CO12CirVir cohost", *Journal of Hepatology*, vol. 66, issue 1, 2017, pp. S21-S22.

[19] 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.

**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 Engineering of Dongseo University, Busan, Korea since 2001. From 1999 to 2001, he was a full time lecturer of Pusan University of Foreign Studies, Busan, Korea. He was also a researcher of Agency for Defense Development, Korea from 1986 to 1992. His areas of research include data mining and database applications.