Towards More Accurate Classification of Instances in Minor Classes

Hyontai Sug

***Abstract*—** In the task of data mining using decision trees, the classification accuracy for minor classes is usually poorer than that of major classes, because decision trees are built to optimize accuracy throughout the available data set and the number of instances belonging to minor classes is relatively rare. So the instances in minor classes are treated less importantly in classification. This paper suggests a method based on progressive over-sampling with respect to minor classes to generate more accurate decision trees for the minor classes for the case that we need more accurate classification for the minor classes. Experiments were done with two representative decision tree algorithms, C4.5 and CART, and two data sets, 'adult' and 'internet ads', and showed the validity of the method.

***Keywords*—** Biased sampling, minor classes, data mining.

I. Introduction

It is known that decision trees are a good tool for data mining so that the tool have used in many applications [1][2][3][4]. Understandability in found knowledge and scalability that enables us to deal with large data sets are two ingredients for decision tree to be a good tool. But the problem of disdaining minor classes is some weak point of decision trees. Because of the greedy property of decision tree generation algorithms, as a decision tree is being built, each branch in the decision tree becomes to have less and less training examples as the result of branching, and the instances of major classes are considered more importantly than the instances of minor classes. Therefore, the reliability of lower branches becomes worse than upper branches due to the smaller size of training examples. So, the classification accuracy for minor classes is less accurate than that of major classes.

If the size of data sets for data mining is very large, we usually resort to sampling. Some fact in sampling is that the trained knowledge model based on the samples is likely dependent on the samples. It is known that decision tree algorithms are more dependent upon training data sets than other data mining algorithms, because decision tree algorithms divide the data sets decisively, while some other data mining methods like artificial neural networks [5] supply all training instances simultaneously to all of their networks.

In order to overcome the problem of disdaining minority classes in decision tree generation algorithms, we need some technique so that the minor classes are treated more importantly in the decision tree algorithms. In this paper we investigate some progressive method of over-sampling that allows for decision tree algorithms to consider minor classes more importantly.

In section 2, we provide the related work to our research, and in sections 3 we present our method. Experiments were run to see the effect of the method in section 4. Finally section 5 provides some conclusions.

This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2010.

II. Related Work

Decision tree algorithms use some greedy search methods to split branches so that generated decision trees may not be optimal. There have been a lot of efforts to build better decision trees and splitting measure is a major concern. For example, C4.5 algorithm [6] that is often referred in literature uses an entropy-based measure, and the measure prefers the most certain split among possible splits from candidate features. Other mostly used decision tree algorithm like CART [7] which uses purity-based measure for split does similar process. So, major classes are preferred, because there are more instances of major classes in the data set, and usually more certain in splitting.

Scalability in decision trees was also good issue for research. Some representatives are like SLIQ, SPRINT, PUBLIC, and SURPASS. SLIQ [8] saves some computing time when the data set consists of many continuous attributes by using a pre-sorting technique in tree-growth phase, and SPRINT [9] is an improved version of SLIQ to solve the scalability problem by building trees with parallel processing algorithm. PUBLIC [10] tries to save some computing time by integrating the tasks of pruning and generating branches together. SURPASS [11] solves the problem of large data set size by bringing the portion of data set into main memory that are needed to grow branches at the moment. However, even though these methods may treat large data sets, the problem of neglecting minor classes still may occur.

Because training of decision trees is a kind of induction, and the data is fragmented in the training process, the performance of trained decision tree is dependent on the training data set a lot. So, we can infer that the resulting decision trees may be dependent on the composition of data in the data set. SMOTE method [12] used synthetic data generation method for minor

classes, and showed that it is effective for decision trees. In [13] the authors showed that class imbalance has different effect in neural networks for medical domain data. In [14] the authors suggested a new decision tree algorithm to treat class imbalance problem.

III. THE METHOD

Because decision tree algorithms do not give high priority to minor classes when they split branches, it is highly possible that instances in minor classes are treated in the lower part of the tree, and this treat may increase misclassification rate for the minor classes. So we want decision tree algorithms to treat the instances of minor classes more importantly. In order to do this, we increase the number of instances of minor classes by duplication. Moreover, in order to decide a good duplication rate, we increase the percentage of duplication progressively. The following is a brief description of the procedure of the method.

INPUT: a data set for data mining,
K: the percentage of over-sampling,
X: sample size,
Y: the number of times to do sampling.
OUTPUT: better decision trees with respect to minor class.
Begin
Do random sampling of size of X, and Y times.
For each sample data set **Do**
Generate a decision tree for original sample data**;**
Make confusion matrix with test data;
Do repeat
Duplicate the instances of minor class by increasing K%;
/* increase K% more*/
Generate a decision tree;
Make a confusion matrix using the test data;
m:=number_of_false_classification_in_minor_class;
Until m converges;
End Do;
End.

In the algorithm we duplicate the instances in minor class until the change in false classification for minor class retches to some convergence. We can also set K percentage of duplication during the iteration in the loop. In the following experiment given K value is 100%, and six for Y, 16,000 and 1,100 for X for two different data sets for the experiment.

IV. EXPERIMENTATION

Experiments were run using a database in UCI machine learning repository [15] called 'adult' [16] and 'internet ads' [17] to see the effect of the method. The number of instances is 48,842. Class probabilities for label '<=50K' and '>50K' are 76.07% and 23.93% respectively, so class '>50K' is the minor class. The database was selected because it is relatively large and contains lots of values. The total number of attributes is 14, and among them six are continuous attributes and eight are nominal attributes. 'Internet ads' data set consists of a set of possible advertisements on web pages. The data set has the encoding of the geometry of the image in the web pages and phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. There are two classes, an advertisement, 'ad', or not, 'nonad'. The total number of attributes is 1,558, and among them three are continuous attributes like height, width, aratio, and all other attributes are nominal attributes having only two values. The number of instances is 3,279. Class probabilities for label 'nonad' and 'ad' are 86% and 14% respectively, so class 'ad' is the minor class.

C4.5 and CART were used to generate decision trees for seven sample sets. Sample sets of size 16,000 and 1,100 were used for data set 'adult' and 'internet ads' respectively. Remaining data are used for test. The following Table 1 to 12 show accuracy and confusion matrix in minor class over-sampling for 'adult' data set.

Table 1. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 1 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.67%	4,696	3,157
	1,550	23,441
200%: 83.49%	5,660	2,191
	3,230	21,761
300%: 81.28%	6,075	1,776
	4,373	20,618
400%: 80.46%	6,077	**1,774**
	4,642	20,349

In table 1, the difference of false '>50K' between 300% and 400% minor class over-sampling is only two, so we stop further over-sampling.

Table 2. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 1 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 86.08%	4,617	3,254
	1,339	23,652
200%: 81.94%	5,522	2,329
	3,601	21,390
300%: 81.53%	5,615	2,236
	3,829	21,162
400%: 81.14%	5,764	2,087
	4,107	20,884
500%: 80.96%	5,819	2,032
	4,221	20,770
600%: 80.81%	5,978	**1,873**

	4,428	20,563
700%: 80.26%	5,923	1,898
	4,586	20,405

In table 2, the difference of false '>50K' between 600% and 700% minor class over-sampling is -25, so we stop further over-sampling.

Table 3. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 2 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.51%	4,462	3,398
	1,362	23,620
200%: 83.04%	5,744	2,116
	3,455	21,527
300%: 81.27%	6,045	1,815
	4,336	20,646
400%: 80.67%	6,081	**1,773**
	4,575	20,407

In table 3, the difference of false '>50K' between 300% and 400% minor class over-sampling is 42, so we stop further over-sampling.

Table 4. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 2 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.87%	4,823	3,037
	1,603	23,379
200%: 82.63%	5,558	2,302
	3,402	21,580
300%: 81.37%	5,649	2,211
	3,906	21,076
400%: 80.86%	5,847	2,013
	4,273	20,709
500%: 80.51%	6,113	1,747
	4,654	20,328
600%: 80.13%	6,136	**1,724**
	4,803	20,179

In table 4, the difference of false '>50K' between 500% and 600% minor class over-sampling is 23, so we stop further over-sampling.

Table 5. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 3 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.63%	4,874	2,976
	1,745	23,249
200%: 83.73%	5,679	2,171
	3,173	21,819
300%: 81.95%	5,904	1,946
	3,981	21,011
400%: 80.58%	6,097	1,753
	4,626	20,366
500%: 80.07%	6,115	**1,735**
	4,809	20,183

In table 5, the difference of false '>50K' between 400% and 500% minor class over-sampling is only 18, so we stop further over-sampling.

Table 6. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 3 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 86.22%	4,489	3,361
	1,166	23,826
200%: 81.36%	5,519	2,331
	3,790	21,202
300%: 80.86%	5,604	2,246
	4,041	20,951
400%: 80.77%	5,779	2,071
	4,243	20,749
500%: 80.52%	5,861	**1,989**
	4,407	20,585
600%: 80.12%	5,539	1,991
	4,618	20,374

In table 6, the difference of false '>50K' between 500% and 600% minor class over-sampling is -2, so we stop further over-sampling.

Table 7. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 4 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.62%	4,600	3,195
	1,528	23,519
200%: 83.65%	5,611	2,184
	3,185	21,862
300%: 81.93%	5,971	1,824
	4,112	20,935
400%: 80.91%	6,091	**1,704**
	4,565	20,482
500%: 80.52%	6,045	1,750
	4,647	20,400

In table 7, the difference of false '>50K' between 400% and 500% minor class over-sampling is -46, so we stop further over-sampling.

Table 8. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 4 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.93%	4,647	3,148
	1,472	23,575
200%: 82.51%	5,515	2,280
	3,464	21,583
300%: 81.14%	5,523	2,272
	3,921	21,126
400%: 80.75%	5,695	2,100
	4,221	20,826
500%: 80.45%	5,731	**2,064**
	4,538	20,689

In table 8, the difference of false '>50K' between 400% and 500% minor class over-sampling is 36, so we stop further over-sampling.

Table 9. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 5 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.53%	4,563	3,343
	1,408	23,528
200%: 83.79%	5,543	2,363
	2,960	21,976
300%: 81.71%	6,123	1,783
	4,225	20,711
400%: 80.64%	6,194	1,712
	4,647	20,289
500%: 80.43%	6,158	**1,748**
	4,679	20,257

In table 9, the difference of false '>50K' between 400% and 500% minor class over-sampling is only -36, so we stop further over-sampling.

Table 10. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 5 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.79%	4,579	3,327
	1,340	23,596
200%: 81.69%	5,588	2,318
	3,697	21,239
300%: 81.63%	5,763	2,143
	3,891	21,045
400%: 81.49%	5,692	2,214
	3,866	21,070
500%: 81.04%	5,787	2,119
	4,109	20,827
600%: 81.21%	5,928	**1,948**
	4,223	20,713
700%: 80.89%	5,939	1,967
	4,313	20,623

In table 10, the difference of false '>50K' between 600% and 700% minor class over-sampling is -19, so we stop further over-sampling.

Table 11. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 6 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.68%	4,691	3,156
	1,548	20,447
200%: 83.46%	5,714	2,133
	3,299	21,696
300%: 81.47%	6,109	1,738
	4,348	20,647
400%: 80.84%	6,226	**1,621**
	4,672	20,323
500%: 80.26%	6,165	1,682
	4,802	20,193

In table 11, the difference of false '>50K' between 400% and 500% minor class over-sampling is only -51, so we stop further over-sampling.

Table 12. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 6 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 86.30%	4,758	3,089
	1,411	23,584
200%: 82.36%	5,545	2,302
	3,491	21,504
300%: 81.27%	5,677	2,170
	3,981	21,014
400%: 81.02%	5,921	1,926
	4,308	20,687
500%: 80.92%	5,981	1,866
	4,401	20,594
600%: 80.24%	6,066	**1,781**
	4,680	20,315
700%: 80.07%	6,041	1,806
	4,740	20,255

In table 12, the difference of false '>50K' between 600% and 700% minor class over-sampling is -25, so we stop further over-sampling.

The following Table 13 to 24 show accuracy and confusion

matrix in minor class over-sampling for 'internet ads' data set.

Table 13. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 1 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.69%	239	60
	34	1,846
200%: 96.01%	247	52
	35	1,845
300%: 95.73%	251	48
	45	1,835
400%: 95.78%	259	40
	52	1,828
500%: 95.50%	260	39
	59	1,821
600%: 95.55%	260	39
	58	1,822
700%: 95.55%	267	**32**
	65	1,815
800%: 95.55%	267	32
	65	1,815
900%: 95.55%	267	32
	65	1,815
1,000%: 95.41%	264	35
	65	1,815

In table 13, there is no improvement after 700% minor class over-sampling, so we stop further over-sampling.

Table 14. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 1 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.82%	236	63
	28	1,852
200%: 95.27%	236	63
	40	1,840
300%: 94.63%	243	56
	61	1,819
400%: 95.14%	247	52
	54	1,826
500%: 95.14%	247	52
	54	1,826
600%: 94.98%	247	52
	57	1,823
700%: 95.27%	251	**48**
	55	1,825
800%: 95.14%	251	48
	58	1,822
900%: 95.23%	250	49
	55	1,825
1,000%: 95.23%	250	49
	55	1,825

In table 14, there is no improvement after 800% minor class over-sampling, so we stop further over-sampling.

Table 15. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 2 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.82%	235	73
	18	1,853
200%: 95.78%	247	61
	31	1,840
300%: 95.36%	249	59
	42	1,829
400%: 95.32%	257	**51**
	51	1,820
500%: 95.32%	257	51
	51	1,820
600%: 95.32%	257	51
	51	1,820
700%: 95.32%	257	51
	51	1,820

In table 15, there is no improvement after 400% minor class over-sampling, so we stop further over-sampling.

Table 16. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 2 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 96.65%	257	51
	22	1,849
200%: 96.19%	249	59
	24	1,847
300%: 96.24%	261	**47**
	35	1,836
400%: 95.23%	245	63
	41	1,830
500%: 95.23%	245	63
	41	1,830
600%: 95.23%	245	63
	41	1,830

In table 16, there is no improvement after 300% minor class over-sampling, so we stop further over-sampling.

Table 17. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 3 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'

Original: 95.46%	233	76
	23	1,847
200%: 94.13%	245	64
	64	1,806
300%: 95.73%	247	**62**
	74	1,796
400%: 93.76%	247	62
	75	1,795
500%: 93.71%	247	62
	74	1,796
600%: 93.76%	245	64
	71	1,799

In table 17, there is no improvement after 300% minor class over-sampling, so we stop further over-sampling.

Table 18. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 3 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.27%	231	78
	25	1,845
200%: 95.87%	248	61
	29	1,841
300%: 94.08%	267	42
	87	1,783
400%: 94.13%	267	42
	86	1,784
500%: 94.22%	276	33
	93	1,777
600%: 94.98%	277	**32**
	95	1,775
700%: 94.17%	277	32
	95	1,775
800%: 94.13%	276	33
	95	1,775
900%: 94.13%	276	33
	95	1,775

In table 18, there is no improvement after 600% minor class over-sampling, so we stop further over-sampling.

Table 19. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 4 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.59%	225	81
	15	1,858
200%: 96.19%	246	60
	23	1,850
300%: 95.87%	248	58
	32	1,841
400%: 95.09%	260	**46**
	61	1,812
500%: 95.09%	260	46
	61	1,812
600%: 95.09%	260	46
	61	1,812
700%: 95.04%	260	46
	62	1,811

In table 19, there is no improvement after 400% minor class over-sampling, so we stop further over-sampling.

Table 20. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 4 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.32%	222	84
	18	1,855
200%: 96.33%	254	52
	28	1,845
300%: 96.87%	261	45
	45	1,828
400%: 95.96%	261	45
	43	1,830
500%: 96.01%	262	44
	43	1,830
600%: 96.01%	262	44
	43	1,830
700%: 95.96%	263	**43**
	45	1,828
800%: 95.69%	263	43
	51	1,822
900%: 95.55%	261	45
	52	1,821
1,000%: 95.64%	258	48
	47	1,826

In table 20, there is no improvement after 700% minor class over-sampling, so we stop further over-sampling.

Table 21. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 5 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 96.19%	219	72
	11	1,878
200%: 96.88%	241	50
	18	1,871
300%: 97.06%	250	41
	23	1,866
400%: 96.97%	253	**38**
	28	1,861
500%: 96.83%	252	39
	30	1,859

600%: 96.93%	252	39
	28	1,861
700%: 96.83%	249	42
	27	1,862

In table 21, there is no improvement after 400% minor class over-sampling, so we stop further over-sampling.

Table 22. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 5 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 96.06%	217	74
	12	1,877
200%: 96.47%	241	50
	27	1,862
300%: 96.79%	244	47
	23	1,866
400%: 96.79%	243	48
	22	1,867
500%: 96.56%	241	50
	25	1,864
600%: 96.65%	243	48
	25	1,864
700%: 96.61%	242	49
	25	1,864
800%: 96.61%	248	**43**
	31	1,858
900%: 96.47%	243	48
	29	1,860
1,000%: 96.56%	245	46
	29	1,860
1,100%: 96.56%	245	46
	29	1,860

In table 22, there is no improvement after 800% minor class over-sampling, so we stop further over-sampling.

Table 23. Confusion matrix of decision tree by C4.5 with various percentages of over-sampling for minor class for sample set 6 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 96.28%	250	57
	24	1,8749
200%: 96.38%	267	**40**
	39	1,834
300%: 96.15%	265	42
	42	1,831
400%: 96.10%	265	42
	43	1,830
500%: 96.15%	265	42
	42	1,831

In table 23, there is no improvement after 200% minor class over-sampling, so we stop further over-sampling.

Table 24. Confusion matrix of decision tree by CART with various percentages of over-sampling for minor class for sample set 6 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.32%	227	80
	22	1,851
200%: 96.56%	261	46
	29	1,844
300%: 96.51%	271	36
	40	1,833
400%: 96.33%	269	38
	42	1,831
500%: 96.28%	270	37
	44	1,829
600%: 96.24%	272	**35**
	47	1,826
700%: 96.24%	272	35
	47	1,826
800%: 95.87%	271	36
	54	1,819
900%: 96.33%	270	37
	43	1,830

In table 24, there is no improvement after 600% minor class over-sampling, so we stop further over-sampling.

Let's think of how we can use the trees, and assume that DT1 is a decision tree generated from original sample data set, and DT2 is the best decision tree with respect to the number of false classification for minor class from over-sampling. According to the result of experiment DT1 has good accuracy for the major class. On the other hand, DT2 is good for the minor class. But the confidence of each terminal node in DT2 is originated from the over-sampled data set, so that it is somewhat exaggerated. So, we need to modify the confidence of each terminal node of DT2 with test data set. In addition, the confidence of each terminal node of DT2 had better be modified with test data set to provide more accurate confidence for each terminal node. Note that the size of test data sets is larger than the size of training data sets, or we may use the whole data set.

In order to classify class-unknown instances, we try to classify them using both trees. If the two decision trees classify an instance as it belongs to the same class, we decide it is in the class. If it is classified differently, we trace the branches of the two trees, and select a class that has higher confidence. In the above experiment, we may use the decision tree in table 25 to 28 for tie break, if we use voting.

Table 25. Decision tree by C4.5 for sample set 7 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'

Original: 85.39%	4,603	3,304
	1,495	23,440

Table 26. Decision tree by CART for minor class for sample set 7 of adult data set

Over-sampling Ratio: accuracy	True '>50K'	False '>50K'
	False '<=50K'	True '<=50K'
Original: 85.93%	4,645	3,262
	1,360	23,575

Table 27. Decision tree by C4.5 for sample set 7 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.92%	230	70
	19	1,861

Table 28. Decision tree by CART for sample set 7 of internet ads data set

Over-sampling Ratio: accuracy	True 'ad'	False 'ad'
	False 'nonad'	True 'nonad'
Original: 95.69%	233	67
	27	1,853

V. CONCLUSIONs

For the task of data mining decision trees are one of good data mining tools because of their understandability and scalability. Even though the good points, there is some weak point of disdaining minor classes due to the fact that their branching criteria give higher priority for major classes. So, the classification for minor classes that occurs scarcely in the data set is less accurate than that of major classes.

If target databases for data mining are very large, we may resort to sampling. An important fact in decision tree algorithms is that the trained decision trees are highly dependent on the training data set. So, in order to overcome the problem of disdaining minority classes in decision tree algorithms, we resort to a technique of progressive over-sampling for minor classes with duplication. The generated decision trees can be used with voting method to predict minor classes for unseen instances. Experiments were done with two very different real world data sets and two decision tree algorithms, C4.5 and CART, and the experiments showed good results.

REFERENCES

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

[2] S. Segrera, 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*, Issue 11, Volume 6, November 2009, pp. 1751-1760.

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

[5] P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining*, Addison-Wesley, 2006.

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

[7] M. Mehta, R. Agrawal, and J. Rissanen, "SLIQ : A Fast Scalable Classifier for Data Mining," *EDBT'96*, Avignon, France , 1996.

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

[9] J. Shafer, R. Agrawal, and M. Mehta., "SPRINT : A Scalable Parallel Classifier for Data Mining," in *Proc. 1996 Int. Conf. Very Large Data Bases*, Bombay, India, Sept. 1996, pp. 544-555.

[10] R. Rastogi, K. Shim, "PUBLIC : A Decision Tree Classifier that Integrates Building and Pruning," *Data Mining and Knowledge Discovery*, vol. 4, no. 4, Kluwer International, 2002, pp. 315-344.

[11] X. Li, "A Scalable Decision Tree System and Its Applications in Pattern Recognition and Intrusion Detection," *Decision Support Systems*, Vol. 41, issue 1, 2005, pp. 112-130.

[12] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, 2002, pp. 341-378.

[13] M.A. Mazuro, P.A. Habas, J.M. Zurada, J.Y. Lo, J.A. Baker, G.D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, Issues 2-3, 2008, pp. 427-436.

[14] C. Lee, C. Tsai, C. Chen, "A Hierarchical Shrinking Decision Tree for Imbalanced Datasets, " in *Proceedings of the 5th WSEAS Int. Conf. on DATA NETWORKS, COMMUNICATIONS & COMPUTERS*, Bucharest, Romania, October 16-17, 2006, pp. 178-183.

[15] A. Suncion, D.J. Newman, UCI Machine Learning Repository [http://www.ics.uci.edu/$\sim$mlearn/{MLR}epository.html]. Irvine, CA: University of California, School of Information and Computer Sciences, 2007.

[16] R. Kohavi, "Scaling up the accuracy of Naive-Bayes classifiers: a decision-tree hybrid," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 202-207.

[17] N. Kushmerick,"Learning to remove Internet advertisements," in *Proceedings of the 3rd International Conference on Autonomous Agents*, 1999, pp.175-181.

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, 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.