More Balanced Decision Tree Generation for Imbalanced Data Sets including the Parkinson's Disease Data

Hyontai Sug

***Abstract*—**The performance of decision tree algorithms for minor classes may be poor, because the algorithms are constructed to achieve the maximum accuracy for a given data set, while the minor classes are often neglected. Over-sampling can be a plausible strategy for better classification in such cases. SMOTE was devised, as an over-sampling method that generates artificial instances. But, the quality of the generated instances may not be as good as desired, even though these instances are based on nearest neighbors. To surmount this problem we suggest a new method that examines the generated instances by using artificial neural networks so that we may achieve better training set for the minor class. The effectiveness of this method is shown by experiments.

***Keywords*—**Decision trees, synthetic data, over-sampling, minority class.

I. Introduction

DECISION trees are important data mining tools because of their good understandability and are used frequently for classifying data [1, 2, 3, 4]. However, the two tools tend to neglect minor data to achieve maximum overall accuracy, making the misclassification of minor data a major concern for effective data mining [5, 6]. Minor classes are classes having a relatively smaller number of instances in the target data sets. The accurate classification of these minor classes is more important than major classes, because we are often more interested in these rare cases. Over-sampling can be a good strategy to overcome this problem, when data collectability is limited. This method is especially applied to imbalanced data sets to find more reliable classifiers for minor classes [7].

There are two kinds of over-sampling method: simple over-sampling and artificially generating minor instances. SMOTE[8] is a representative over-sampling method based on artificial instance generation for a minor class. Because of its importance several slightly modified methods based on SMOTE have been suggested [9, 10]. But, incorrect training instances can easily lead to incorrect classifications. So, generating correct instances is important.

The true class of the artificially generated instances, however, may be questioned, because they are not real data. One possible solution is to rely on the opinion of a domain expert, but, an expert may not be always available. A second possibility is to rely on the data themselves.

There are many data mining algorithms available, and depending on the particular data mining algorithm used, each data set may have a different performance. For example, decision trees and artificial neural networks may have a different performance for the same data set, because decision tree algorithms are based on greedy search methods, while artificial neural networks are based on repeated and gradual training methods. In other words, decision tree algorithms have a stronger tendency to be satisfied with local optima compared to artificial neural networks. As a result, it is known that decision trees have poorer performance than artificial neural networks in many cases [11, 12]. So, we may use them to test the artificially generated instances. In section 2 we discuss our experiment method, and in section 3 conclusions are provided.

II. METHOD AND EXPERIMENT

A. Method

Using artificially generated instances of the minor class, SMOTE attempts to generate better decision trees like those of C4.5[13]. The artificial instances are made based on the K-nearest neighbors algorithm and randomization on continuous values between the nearest neighbors. While the performance of the system has been verified with a10-fold cross validation, there is some possibility that the class of the artificially generated instances may not be correct.

On the other hand, due to the differences of the data mining algorithms used, each data set may have a different accuracy. So, we want to check the class of artificially generated instances by SMOTE using artificial neural networks. In the following experiments, we first check the performance of three different data mining algorithms using data sets generated from SMOTE, and then check the quality of the over-sampled data sets.

B. Principles of Related Algorithms

In this paper we want to find better decision trees. So, let's see the principles of two representative decision tree algorithms.

C4.5 was invented by J. L. Quinlan [13]. It uses entropy-based splitting criterion to grow its branches of the tree. The definition of entropy for information can be

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

expressed H(X). Suppose $i \in \{1, 2, \ldots, n\}$ and let $p(x_i)$ be the fraction of instances of class *i* in the data set.

$$H(X) = -\sum_{i=1\sim n} p(x_i) \log p(x_i) \qquad (1)$$

So, more skewed distributions of X can have smaller H(X) values.

CART stands for Classification And Regression Trees, and invented by Breiman et al. [14]. CART uses purity-based splitting criterion to grow its branches in the tree called Gini index.

$$G(X) = \sum_{i=1\sim n} p(x_i) \{1-p(x_i)\} \qquad (2)$$

So, if all instances are in the same class, the Gini index value is 0. Because of the property of the equation, the values generated from Gini index is more uniform than those from entropy.

C. Performance Measures

This paper uses two performance measures to evaluate the effect of the suggesting method.

Accuracy is very common measure to explain the overall performance. Assume that we have a confusion matrix like table 1.

Table 1. Confusion matrix

	Actual class		
		Positive(p)	Negative(n)
Predicted class	Positive (p)	True positives(TP)	False positives(FP)
	Negative (n)	False negatives(FN)	True negatives(TN)

The accuracy can be calculated

$$\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN) \qquad (3)$$

The TP rate for class p can be calculated

$$TPp = TP / (TP + FN) \qquad (4)$$

The TP rate for class n can be calculated

$$TPn = TN / (TN + FP) \qquad (5)$$

Because different data sets have different performance with respect to TP rate for each class, geometric mean can be a good measure to compare whole TP rates for each used data mining algorithm. The geometric mean Gm can be calculated

$$Gm = SQRT(TPp \times TPn) \qquad (6)$$

If we have more than two classes, say N, true positives of each class will be in each respective diagonal position of N×N confusion matrix, and the same notation as the 2×2 confusion matrix can be applied.

D. Related Work

Developing better data mining models for imbalanced data sets attracted a lot of attentions, because many real world data sets have such property [15, 16]. Imbalanced data sets may suffer neglecting minor classes, because minor classes occupy only small portion of the whole population of training data set so that the true positive rates for the minor classes may be poor even the overall accuracy is good. As a means to mitigate the class imbalance problem in [17] SVM was used. Because SVM uses data repeatedly to train its model that usually found in the training method of artificial neural networks, better results in experiment was reported. In [18] the effect of over-sampling and under-sampling was investigated with several data mining algorithms of accuracy like SVM, rough sets, cost sensitive classifiers, and compared the effect of the algorithms.

We used two very different data sets for experiment; the annealing data and the Parkinson's data. So, let's see related work that used the data sets to find better data mining models.

Several research results were reported for better performance in the classification of the annealing data. In [19] neural network ensemble is used. The accuracy is about 92.81% with 10-fold cross validation. In [20] cost sensitive learning method including BP was suggested to find better machine learning algorithms. Several public data sets including the annealing data were used for experiment. Around 2.3% error rate was reported for the annealing data in the experiment.

Parkinson's disease also attracted research interests a lot. Because artificial neural networks are known to have higher accuracy than other machine learning algorithms in many data sets, many researchers tried to use related algorithms to gain data mining model of accuracy for the data set. In [21] probabilistic neural network approach was used, and achieved accuracy around 81% with 70% of data for training and 30% of data for testing. In [22] MLP and SVM were used to train the Parkinson's disease data, and achieved accuracy of 92%~93%. In [23] six different machine learning algorithms were used to find the best one for the data set. Using leave-one-out cross validation, they reported accuracy of 78.1% ~ 81.2% with their models, and among them naive Bayesian and fuzzy rule-based system achieved the accuracy of 81.2%. In [24] genetic algorithm was used to select subset of attributes, and after the selection SVM was used. SVM achieved accuracy of 96.06%, 93.58%, and 93.61% when the number of selected attributes is 4, 7, and 9 respectively with 75% of the data for training and 25% of the data for testing. Anyway, because most used algorithms do not generate knowledge models of comprehension, the understandability of the models is more limited than that of decision trees.

E. Experiment

An experiment was performed using two very different data sets in the UCI machine learning repository [25].

The annealing data set has 798 instances and 38 attributes consisting of 6 continuously-valued, 3 integer-valued, and 29 nominal-valued attributes. Because the 3 integer-valued attributes have a few different values only, they are considered nominal in the experiment. The data set contains many missing values, and the instances consist of class values 1, 2, 3, 4, 5, and U.

The Parkinson's disease data set has biomedical voice measurement data from 31 people among them 23 have Parkinson's disease. Each column of data record has information of voice measure, and each record corresponds voice recording from each individual. The data set contains 195 records. The main aim of the data is to discriminate healthy people from those with Parkinson's disease. Table 2 and 3 shows the property of the data sets. Each data set has nominal class values. So, each distinct class value for each data set is represented in numbers for convenience.

Table 2. The property of data sets

Data set	No. of instances	No. of attributes	
		continuous	nominal
Annealing data	798	6	32
Parkinson's disease	195	23	0

Table 3. The number of instances for each data

Data set	Number of instances per class					
	1	2	3	4	5	6
Annealing data	**8**	88	608	0	60	34
Parkinson's disease	**48**	147				

The class having the least number of instances is considered a minor class for each data set as indicated by bold characters in table 3. The class having 0 instances is not considered for a minor class, because SMOTE cannot generate artificial instances for such class. The annealing data consist of both training and test data sets, but because the test data set does not have any instances of class 1 of our interest, only the training data set is used for this experiment.

Three different data mining algorithms were used for the experiment: C4.5, CART and multilayer perceptron (MLP) as artificial neural network [26]. C4.5 and CART were chosen because they can be representative decision tree algorithms [1]. The weka data mining package was used [27] with default parameters for C4.5 and CART. The experiment was performed with the original data set, and over-sampled data sets for the minor class with over-sampling rates of 100%, 200%, 300%, 400%, and was based on 10-fold cross validation. In addition, two other over-sampled training data sets were made based on all the previously over-sampled data sets. One that was classified as 'correct' by MLP(true positive), and one that was classified as 'incorrect' by MLP(false positive). Table 4 shows the change of number of instances as the over-sampling rate changes from 100% to 400%.

Table 4. The change in the number of instances of a minor class as over-sampling rate changes

Data set	No. of instance of the minor class for each over-sampling rate			
	100 %	200 %	300 %	400 %
Annealing data	16	24	32	40
Parkinsons disease	96	144	192	240

1) The Annealing Data

Table 5 shows the result for annealing data with C4.5, CART, and MLP. The minor class of the data set is class 1 that is indicated in bold characters. The training time of MLP is 2000. There are no instances of class 4, so its true positive rate is 0 in the table. The last row shows geometric mean of TP rate. The TP rate of class 4 which is 0 is omitted in the calculation of the geometric mean.

Table 5. The accuracy of three different data mining algorithms for the annealing data

		C4.5	CART	MLP
Accuracy(%)		92.6065	91.8546	98.7649
TP rate	**Class 1**	0.5	0.625	0.625
	Class 2	0.716	0.761	0.625
	Class 3	0.969	0.952	0.993
	Class 4	0	0	0
	Class 5	0.883	0.867	1
	Class 6	0.882	0.882	0.912
Gm		0.5198	0.5884	0.5948

Table 6 and table 7 show the change of accuracy and true positive rate of each class as the over-sampling rate for the minor class 1 increases from 100% to 400%.

Table 6. The accuracy of three different data mining algorithms with over-sampling rate of 100% and 200% for class 1 of the annealing data

		Over-sampling rate	
		100%	200%
Accuracy (%)	C4.5	92.1836	92.7518
	CART	92.1836	92.2604
	MLP	97.8908	98.0344

Table 7. The accuracy of three different data mining algorithms with over-sampling rate of 300% and 400% for class 1 of the annealing data

		Over-sampling rate	
		300%	400%
Accuracy (%)	C4.5	92.2141	91.6867
	CART	92.8224	91.9277
	MLP	95.7421	97.1084

Because class 1 occupies relatively smaller number of instances in the training data set, increasing the number of instances by over-sampling does not improve accuracy. Fig. 1 shows the resulting graph of change of accuracy.

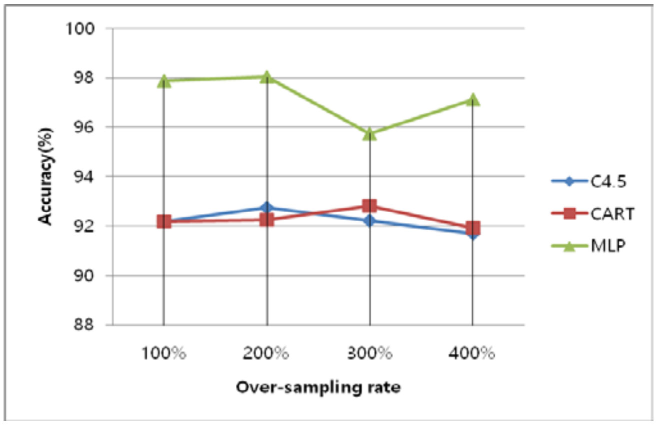

Fig. 1 Change of accuracy as over-sampling rate changes

To check the quality of artificial data instances generated by SMOTE we performed more experiments: the first using all the artificial instances that are true positives with the MLP of the original data set, and the second using all the artificial instances that are false positives with the MLP of the original data set.

There are 59 and 9 distinct instances in TP and FP groups respectively. Table 8 and table 9 show the result of experiment with data set that is made with the original data set and all the over-sampled instances by SMOTE that was classified as 'correct' by MLP(true positive), all the over-sampled instances by SMOTE that was classified as 'incorrect' by MLP(false positive), and part of TP instances.

Table 8. The accuracy of three different data mining algorithms for the annealing data with over-sampled true positive and false positive instances with respect to MLP

			Over-sampled sets	
			Original plus instances of true positive(59)	Original plus instances of false positive(9)
Accuracy (%)		C4.5	92.7655	91.0781
		CART	92.2987	91.5737
		MLP	96.266	97.3978
TP rate for each class	1	C4.5	0.985	0.765
		CART	0.955	0.706
		MLP	0.985	0.882
	2	C4.5	0.705	0.67
		CART	0.727	0.75
		MLP	1	1
	3	C4.5	0.959	0.962
		CART	0.952	0.946
		MLP	1	0.997
	4	C4.5	0	0
		CART	0	0
		MLP	0	0
	5	C4.5	0.9	0.8
		CART	0.9	0.933
		MLP	1	1
	6	C4.5	0.882	0.882
		CART	0.882	0.882
		MLP	0.088	0.5
Gm		C4.5	0.7271	0.6241
		CART	0.7243	0.6420
		MLP	0.2944	0.6631

In addition, because the performance of data mining algorithms may depend on available training instances for each class, a third experiment was done with an equal number of over-sampled instances of true and false positives. Because the number of instances in FP is nine, ten random samples of size nine were made from the 59 TP instances. The numbers in the left column of table 9 are the average of the ten samples.

Table 9. The accuracy of three different data mining algorithms for the annealing data with equal number of over-sampled true positive and false positive instances with respect to MLP

			Over-sampled sets	
			Original plus 9 random instances of true positive	Original plus instances of false positive(9)
Accuracy (%)		C4.5	**92.1066**	91.0781
		CART	91.5613	**91.5737**
		MLP	97.4102	97.3978
TP rate for each class	1	C4.5	0.7119	0.765
		CART	0.8063	0.706
		MLP	0.9	0.882
	2	C4.5	0.6738	0.67
		CART	0.7093	0.75
		MLP	1	1
	3	C4.5	0.9719	0.962
		CART	0.9511	0.946
		MLP	0.9956	0.997
	4	C4.5	0	0
		CART	0	0
		MLP	0	0
	5	C4.5	0.8534	0.8
		CART	0.91	0.933
		MLP	1	1
	6	C4.5	0.88	0.882
		CART	0.882	0.882
		MLP	0.5	0.5
Gm		C4.5	0.5917	**0.6241**
		CART	**0.6607**	0.6420
		MLP	0.6693	0.6631

Comparing the values in table 9, true positive instances with respect to MLP show slightly better results, because C4.5 shows better accuracy of 1.0285%, while CART shows almost identical accuracy with equal number of TP instances.

More experimentation was done with the artificial data instances that are true positive with respect to the MLP. Four different percentages of over-sampling like the ones in table 6 and table 7 were performed to compare the effect of true positive instances for C4.5 and CART with over-sampling rate of 100%, 200%, 300%, and 400%. The results are summarized in table 10, table 11, table 12, and table 13 respectively.

In table 10 ~ table 13 each right sub-column named 'from TP' in the column of over-sampling sets shows the accuracy of each algorithm and true positive rate of each class. The accuracy and TP rate in the column labeled ‘From TP’ are the average of ten random samples from the 59 TP instances.

Table 10. The accuracy of three different data mining algorithms for the annealing data with 100% over-sampling rate

			Over-sampled sets	
			Conventional	From TP
Accuracy (%)		C4.5	**92.1836**	91.9727
		CART	92.1836	**92.9295**
		MLP	97.8908	97.8806
TP rate for each class	1	C4.5	0.688	0.7065
		CART	0.75	0.7878
		MLP	0.813	0.813
	2	C4.5	0.659	0.6534
		CART	0.818	0.8265
		MLP	1	1
	3	C4.5	0.974	0.9725
		CART	0.959	0.9566
		MLP	0.997	0.9968
	4	C4.5	0	0
		CART	0	0
		MLP	0	0
	5	C4.5	0.883	0.8699
		CART	1	0.8436
		MLP	1	1
	6	C4.5	0.853	0.853
		CART	0.912	0.882
		MLP	0.647	0.647
Gm		C4.5	0.5767	**0.5772**
		CART	0.6309	**0.6808**
		MLP	0.7243	0.7241

Table 11. The accuracy of three different data mining algorithms for the annealing data with 200% over-sampling rate

			Over-sampled sets	
			Conventional	From TP
Accuracy (%)		C4.5	**92.7518**	92.5921
		CART	**92.2604**	92.2481
		MLP	98.0344	97.6413
TP rate for each class	1	C4.5	0.83	0.82
		CART	0.875	0.871
		MLP	0.875	0.875
	2	C4.5	0.75	0.709
		CART	0.773	0.781
		MLP	1	1
	3	C4.5	0.964	0.967
		CART	0.956	0.969
		MLP	0.997	0.997
	4	C4.5	0	0
		CART	0	0
		MLP	0	0
	5	C4.5	0.9	0.9
		CART	0.933	1
		MLP	1	1
	6	C4.5	0.853	0.853
		CART	0.882	0.879
		MLP	0.676	0.582
Gm		C4.5	**0.6683**	0.657
		CART	**0.7117**	0.692
		MLP	0.7679	0.7125

Table 12. The accuracy of three different data mining algorithms for the annealing data with 300% over-sampling rate

			Over-sampled sets	
			Conventional	From TP
Accuracy (%)		C4.5	**92.2141**	91.9708
		CART	**92.8224**	92.7007
		MLP	95.7421	95.8637
TP rate for each class	1	C4.5	0.875	0.866
		CART	0.906	0.9029
		MLP	0.969	0.969
	2	C4.5	0.682	0.66
		CART	0.739	0.7274
		MLP	1	1
	3	C4.5	0.965	0.965
		CART	0.961	0.9606
		MLP	0.995	0.995
	4	C4.5	0	0
		CART	0	0
		MLP	0	0
	5	C4.5	0.9	0.9
		CART	0.917	0.9218
		MLP	1	1
	6	C4.5	0.853	0.853
		CART	0.882	0.882
		MLP	0.088	0.118
Gm		C4.5	**0.6649**	0.6507
		CART	**0.7214**	0.7162
		MLP	0.2933	0.3733

Table 13. The accuracy of three different data mining algorithms for the annealing data with 400% over-sampling rate

			Over-sampled sets	
			Conventional	From TP
Accuracy (%)		C4.5	91.6867	91.6867
		CART	91.9277	**92.0482**
		MLP	97.1084	97.1084
	1	C4.5	0.975	0.975
		CART	0.9	0.925

TP rate for each class		MLP	0.95	0.95
	2	C4.5	0.614	0.614
		CART	0.75	0.75
		MLP	1	1
	3	C4.5	0.959	0.959
		CART	0.954	0.954
		MLP	0.995	0.995
	4	C4.5	0	0
		CART	0	0
		MLP	0	0
	5	C4.5	0.917	0.917
		CART	0.85	0.85
		MLP	1	1
	6	C4.5	0.882	0.882
		CART	0.882	0.882
		MLP	0.441	0.441
Gm		C4.5	0.6814	0.6814
		CART	0.6948	**0.7044**
		MLP	0.6456	0.6456

The better accuracy between 'Conventional' and 'From FP' in each over-sampling is indicated by bold numbers in the table. From table 9 ~ table 12, if we consider the cases of an equal number of instances in the minor class, the ratio of our method being superior versus inferior with respect to the two target algorithms, C4.5 and CART, is 3:6, and that of Gm values is 4:5. Among them the method generated slightly better result for CART. The reason why CART generated better results is because of the splitting measure that CART uses. Gini index values are more uniform than entropy-based method.

2) *The Parkinson's Disease Data*

Table 14 shows the result for the Parkinson’s disease data with C4.5, CART, and the third algorithm MLP. The minor class of the data set is class 1. The training time of MLP is 2000. Table 15 and table 16 shows the change of accuracy as the over-sampling rate for the minor class 1 increases.

Table 14. The accuracy of three different data mining algorithms for the Parkinson's disease data

		C4.5	CART	MLP
Accuracy(%)		80.5128	85.641	93.3333
TP rate	**Class 1**	0.583	0.708	0.896
	Class 2	0.878	0.905	0.946
Gm		0.7155	0.8005	0.9207

Table 15. The accuracy of three different data mining algorithms with over-sampling rate of 100% and 200% for class 1 of the Parkinson's disease data

		Over-sampling rate	
		100%	200%
Accuracy (%)	C4.5	88.0658	86.9416
	CART	88.8889	85.9107
	MLP	94.6502	94.5017

Table 16. The accuracy of three different data mining algorithms with over-sampling rate of 300% and 400% for class 1 of the Parkinson's disease data

		Over-sampling rate	
		300%	400%
Accuracy (%)	C4.5	89.0855	90.1809
	CART	86.7259	91.9897
	MLP	93.5103	96.124

Table 17 shows the result of experimentation with a data set that is composed of the original data set and all the over-sampled instances by SMOTE that were classified as 'correct' by MLP (TP), and all the over-sampled instances by SMOTE that were classified as 'incorrect' by MLP (FP). There are 437 and 33 distinct instances in the TP and FP groups respectively.

Table 17. The accuracy of three different data mining algorithms for the Parkinson's disease data with over-sampled true positive and false positive instances with respect to MLP

			Over-sampled sets	
			Original plus instances of true positive(437)	Original plus instances of false positive(33)
Accuracy (%)		C4.5	93.6709	85.0877
		CART	93.6709	87.7193
		MLP	93.8324	90.3509
TP rate for each class	**1**	C4.5	0.965	0.778
		CART	0.975	0.765
		MLP	0.994	0.889
	2	C4.5	0.944	0.891
		CART	0.81	0.939
		MLP	0.884	0.912
Gm		C4.5	0.9544	0.8326
		CART	0.8887	0.8475
		MLP	0.9374	0.9004

Table 18 shows the result when the number of over-sampled instances in true and false positives are equal. The result of rightmost column is average of ten random samples.

Table 18. The accuracy of three different data mining algorithms for the Parkinson's disease data with equal number of over-sampled true positive and false positive instances with respect to MLP

		Over-sampled sets	
		Original plus instances of false positive(33)	Original plus 33 instances of true positive
Accuracy (%)	C4.5	85.0877	**86.4463**
	CART	87.7193	**88.4649**

		MLP	90.3509	93.8596
TP rate for each class	**1**	C4.5	0.778	0.8111
		CART	0.901	0.8334
		MLP	0.889	0.9368
	2	C4.5	0.891	0.894
		CART	0.918	0.9185
		MLP	0.912	0.9423
Gm		C4.5	0.8326	**0.8515**
		CART	0.8475	**0.867**
		MLP	0.9004	0.9395

Four different percentages of over-sampling like the ones in table 15 and table 16 were performed to compare the effect of the true positive instances for C4.5 and CART. Table 19 ~ table 22 has the result. The column labeled 'From TP' has average of ten random samples from the 437 TP instances.

Table 19. The accuracy of three different data mining algorithms for the Parkinson's disease data with 100% over-sampling rate

			Over-sampled sets	
			Conventional	From TP
Accuracy (%)		C4.5	88.0658	**88.3539**
		CART	**88.8889**	88.6831
		MLP	94.6502	93.4156
TP rate for each class	**1**	C4.5	0.833	0.8548
		CART	0.865	0.8634
		MLP	0.958	0.9417
	2	C4.5	0.912	0.9028
		CART	0.905	0.9021
		MLP	0.939	0.9203
Gm		C4.5	0.8716	**0.8785**
		CART	**0.8848**	0.8825
		MLP	0.9485	0.9309

Table 20. The accuracy of three different data mining algorithms for the Parkinson's disease data with 200% over-sampling rate

			Over-sampled sets	
			Conventional	From TP
Accuracy (%)		C4.5	86.9416	**89.3137**
		CART	85.9107	**90.0**
		MLP	94.5017	94.1924
TP rate for each class	**1**	C4.5	0.882	0.9098
		CART	0.882	0.8634
		MLP	0.965	0.9644
	2	C4.5	0.857	0.8769
		CART	0.837	0.9021
		MLP	0.925	0.9177
Gm		C4.5	0.8694	**0.8932**
		CART	0.8592	**0.9001**
		MLP	0.9448	0.9309

Table 21. The accuracy of three different data mining algorithms for the Parkinson's disease data with 300% over-sampling rate

			Over-sampled sets	
			Conventional	From TP
Accuracy (%)		C4.5	89.0855	**90.9735**
		CART	86.7259	**92.0059**
		MLP	93.5103	95.7522
TP rate for each class	**1**	C4.5	0.906	0.9408
		CART	0.906	0.9522
		MLP	0.974	0.9811
	2	C4.5	0.871	0.8694
		CART	0.816	0.8783
		MLP	0.884	0.9265
Gm		C4.5	0.8883	**0.9044**
		CART	0.8598	**0.9145**
		MLP	0.9279	0.9534

Table 22. The accuracy of three different data mining algorithms for the Parkinson's disease data with 400% over-sampling rate

			Over-sampled sets	
			Conventional	From TP
Accuracy (%)		C4.5	90.1809	**91.2145**
		CART	91.9897	**93.4189**
		MLP	96.124	96.3824
TP rate for each class	**1**	C4.5	0.95	0.933
		CART	0.942	0.9635
		MLP	0.975	0.992
	2	C4.5	0.823	0.878
		CART	0.884	0.8863
		MLP	0.939	0.918
Gm		C4.5	0.8842	**0.9051**
		CART	0.9125	**0.9241**
		MLP	0.9568	0.9543

Comparing the two groups of values, we can see that the quality of the artificial instances that generated the right column is generally better than the other for the Parkinson's data.

Table 23 shows the summary of the result of experiments. The first and second row shows the ratio of our method being superior versus inferior in accuracy and geometric mean of TP rates with respect to the two target algorithms, C4.5 and CART.

Table 23. Summary of the result of the experiment

Data set	Parkinson's disease	Annealing data
Number of superiority vs. inferiority in accuracy	9 : 1	3 : 6
Number of superiority vs. inferiority in Gm	9 : 1	4 : 5
Number of FP vs. TP	33 : 437	9 : 59

Accuracy of MLP for the original data set(%)	9 3.3	98.7
Number of minor instances	48	8
Number of major instances	147	790

From the summary result, we can find that the number of minor instances is a major factor for the success of our method. We can infer that this result is due to the SMOTE's generation method for artificial instances. The smaller number of minor instances may contribute to generate artificial instances in a not good quality, so that the final result may not be as good as we expect.

III. Conclusion

Because data mining algorithms like decision trees are made to achieve the maximum accuracy for a given data set, minor classes are often neglected, so that the performance of the decision trees for these classes may not be good. Over-sampling is a common strategy to cope with the situation of insufficient data especially for these minor classes. SMOTE has been considered a good methodology of over-sampling the minor classes for decision trees. But, there is some possibility that the quality of the artificially generated instances by SMOTE may not be as good as we expected, even though they are made based on the nearest neighbors. In this paper we suggested a new method to surmount the problem by resorting to MLP, which is a more reliable data mining algorithm. In other words, by examining the artificially generated instances with the MLP, and supplying true positive instances only to the target algorithms of decision tree algorithm like C4.5 or CART, we may obtain better trees. Experiments using the two very different data sets showed the property and utility of the method.

Acknowledgment

H. Sug thanks professor Douglas D. Dankel II who has worked for the department of computer and information science and engineering, University of Florida, USA for his invaluable comments about the idea and draft of the paper.

References

[1] 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.

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

[3] A. Kumar, S. Kumar, "Decision Tree based Learning Approach for Identification of Operating System Processes," *WSEAS Transactions on Computers*, vol. 13, 2014, pp. 277-288.

[4] M.M. Mazid, A.B.M.S. Ali, K.S. Tickle, "Input space reduction for Rule Based Classification", *WSEAS Transactions on Information Science and Applications*, issue 6, vol. 7, June2010, pp. 749-759.

[5] Q. Yang, X. Wu, "10 Challenging Problems in Data Mining Research," *International Journal of Information Technology & Decision Making*, vol. 5, no. 4, 2006, pp. 597-604.

[6] Q. Wu, H. Liu, K. Liu, “Mixed-sampling Approach to Unbalanced Data Distribution: A Case Study involving Leukemia Document Profiling,” *WSEAS Transactions on Information Science and Applications*, vol. 8, issue 9, 2011, pp. 356-379.

[7] H. He, “Learning from Imbalanced Data,” *IEEE Trans. Knowledge and Data Engineering*, vol. 21, issue 9, Sep. 2009, pp. 1263-1284.

[8] 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.

[9] H. Han, W. Wang, B. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” *LNCS*, vol. 3644, 2005, pp. 878-887.

[10] D. Zhang, W. Liu, X. Gong, H. Jin,, “A Novel Improved SMOTE Resampling Algorithm Based on Fractal,” *Journal of Computational Information Systems*, vol. 7, no. 6, 2011, pp. 2204-2211.

[11] Y. Kim, "Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size," *Expert Systems with Applications: An International Journal*, vol. 34, issue 2, February 2008, pp. 1227-1234.

[12] L.O. Hall, X. Liu, K.W. Bowyer, R. Banfield, "Why are neural networks sometimes much more accurate than decision trees: an analysis on bio-informatics problem," in *Proc. 2003 IEEE International Conference on Systems, Man and Cybernetics*, vol. 3, 2003, pp. 2851-2856.

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

[14] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*. CRC Press; 1984.

[15] H. He, "Learning from Imbalanced data," *IEEE Transaction on Knowledge Engineering*, vol.21, issue 9, 2009, pp. 1263-1284.

[16] N.V. Chawla, "Data mining for imbalanced datasets: an overview," *Data Mining and Knowledge Discovery Handbook*, O. Maimon, L. Rokach, eds. Springer, 2005, pp. 853-867.

[17] S. Zhang, S. Sadaoui, M. Mouhoub, "An empirical analysis of imbalanced data classification," *Computer and Information Science*, vol. 8, no. 1, 2015, pp. 151-162.

[18] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, issue 4, 2012, pp. 42-47.

[19] Y. Jiang, Z. Zhou, "Editing training data for kNN classifiers with neural network ensemble," *Lecture Notes in Computer Science*, vol. 3173, 2004, pp. 356-361.

[20] Z. Zhou, X. Liu, "Training cost-sensitive neural networks with methods addressing the class imbalance problem," *IEEE Transactions on Data and Knowledge Engineering*, 2006, pp. 1-14.

[21] M. Ene, "Neural Network-Based Approach to Discriminate Healthy People from Those with Parkinson’s Disease," *Mathematics and Computer Science Series*, vol. 35, 2008, pp. 112-116.

[22] D. Gil, M. Johnson, "Diagnosing Parkinson by Using Artificial Neural Networks and Support Vector Machines," *Global Journal of Compute Science and Technology*, vol. 9, 2009, pp. 63-71.

[23] W. Froelich, K. Wrobel, P. Porwik, "Diagnosing Parkinson's Disease Using The Classification of Speech Signals," *Journal of Medical Informatics & Technologies*, vol.23, 2014, pp. 187-193.

[24] M. Shahbakhi, D.T. Far, E. Tahami, "Spech Analysis for Diagnosis of Parkinson's Disease Using Genetic Algorithm and Support Vector Machine," *Journal of Biomedical Science and Engineering*, vol. 7, 2014, pp. 147-156.

[25] 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

[26] M.T. Hagan, H.B. Demuth, *Neural Network Design*, 2nd ed., Martin Hagan, 2014.

[27] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, issue 1, 2009.

Hyontai Sug received BS degree in computer science and statistics from Busan national university, Korea, in 1983, and MS degree in computer science from Hankuk university of foreign studies, Korea, in 1986, and Ph.D. degree in computer and information science and engineering from university of Florida, USA in 1998. He was a researcher of Agency for Defense Development, Korea from 1986 to 1992, and a full time lecturer of Pusan university of foreign studies, Korea from 1999 to 2001. Currently, he is a professor of Dongseo university, Korea since 2001. His research interests include data mining and database applications.