Flexible Data Mining for Medicine Data

HYONTAI SUG
Division of Computer Engineering
Dongseo University
47 Jurye-ro, Sasang-gu, Busan 47011
KOREA
sht@gdsu.dongseo.ac.kr

Abstract: - Hepatitis is a liver disease characterized by inflammatory cells in the tissues of the liver. It causes mild to serious effect to the liver so that patient may even die of it. On the other hand, decision trees are important data mining tools in medicine, because doctors can easily understand the final result of data mining so that they can be used to diagnose the disease. Decision tree algorithms give priority to the classes having more training instances for better classification, so that over-sampling for a minor can be a plausible technique for better classification of the minor class, if we are more interested in the better classification of the minor class. In our hepatitis data the ratio of minor versus major is 32 vs. 123. As a way to build better decision tree for a minority class without sacrificing overall accuracy much, we select good synthetic over-sampled data instances for our decision tree. By selecting good synthetic data instances, we may achieve our goal. Experiments with various levels of over-sampling proved our assertion.

Key-Words: - Data mining, decision tree, sampling, data preparation.

1 Introduction

Hepatitis is a liver disease characterized by inflammatory cells in the tissues of the liver. It causes mild to serious effect to the liver so that patient may even die of it. On the other hand, decision trees are important data mining tools in medicine, because their structure is in tree shape, doctors can easily understand the final result of data mining so that they can be used to diagnose the liver disease to aid doctors. There are many good examples [1, 2].

But, decision tree algorithms have a weak point of preferring major classes. Decision tree algorithms try to achieve better accuracy by focusing on the better classification of major classes, because majority of instances belong to a major class. But, in real world applications we may have more interest in more accurate classification of a minor class [3]. Therefore, as a way for decision tree algorithms to pay more attention to a minor class, we may increase the number of instances in a minor class. But, if data set is limited, over-sampling may be applied. As a way of over-sampling, instead of supplying identical instances randomly, we may supply some very similar synthetic data instances of the minor class. SMOTE [4] is one of well-known over-sampling method that generates synthetic instances of a minor class. But, some of the synthetic instances may not be as good as we expected. In other words, because the instances are not real data, some of them might be not helpful for our target data mining models. So, we want to eliminate some bad instances for our new decision tree for minority.

In section 2 we discuss our experiment method, and in section 3 conclusions are provided.

2 Problem and Method

We want to generate better decision trees for a given data set of hepatitis. Hepatitis data set contains two classes. Class 1 represents ‘die’, and 32 data instances belong to the class 1. Class 2 represents ‘live’, and 123 data instances belong to the class 1. So, the class having 32 instances is a minor class. The data set has nineteen attributes and among them six are continuous attributes, and the others are nominal attributes. The data set can be found in the UCI machine learning repository [5]

There are several well-known decision tree algorithms; C4.5 [6], CART (Classification And Regression Tree) [7], CHAID (Chi-squared Automatic Interaction Detector) [8], etc. Among them, we use C4.5, because it’s one of the mostly used data mining algorithm [9], and easily available. The synthetic data instances are made based on K-

nearest neighbors algorithm and randomization on continues values of the related attribute values of the neighbors. There is some possibility that the quality of the generated instances may not be as good as expected. Because we want to build better decision tree for a minor class, we will check the class of the synthetic instances using the decision tree made from the original data set

Experiments will be performed using a data mining tool called Weka [10], and based on 10-fold cross validation and.

3 Experiment and Results

The data set contains two classes. Class 1 represents 'die' having 32 instances, and class 2 represents 'live' having 123 instances. So, the class 1 is the minor class we are interested in.

Table 1 shows the accuracy of decision tree algorithm, C4.5 for the data set with default parameters. Because the true positive rate (TP) of class 1 which predicts 'die' is very low, it is uncertain that a patient will die or not unless he is classified class 2, 'live', with the original data.

Table 1. The accuracy of C4.5 decision tree for the original data

Accuracy (%)		83.871
TP rate	Class 1	0.438
	Class 2	0.943

Table 2 shows the corresponding confusion matrix.

Table 2. The confusion matrix of C4.5 decision tree for the original data

		Predicted class	
		Class 1	Class 2
Actual class	Class 1	14	18
	Class 2	7	116

After generating the decision tree, over-sampling rate of 900% is applied for the minor class of class 1 using SMOTE. So, additional instances ($32 \times 9 = 288$) are added to the original data set. Table 3 shows the result of the experiment for the new data set.

Table 3. The accuracy of C4.5 decision tree for the over-sampling rate of 900% for class 1

Accuracy (%)		91.6479
TP rate	Class 1	0.95
	Class 2	0.829

We can see positive effect of the over-sampling from table 3. Table 4 shows the corresponding confusion matrix.

Table 4. The confusion matrix of C4.5 decision tree for the over-sampling rate of 900% for class 1

		Predicted class	
		Class 1	Class 2
Actual class	Class 1	304	16
	Class 2	21	102

In order to check the quality of the over-sampled instances by SMOTE, all the over-sampled instances of SMOTE are checked by the decision tree of the original data set. While 118 distinct instances are checked to belong to true positive, the other 170 distinct instances are checked belong to false positive. Using these two groups of over-sampled instances and the original data set, two more experiment were run. Table 5 shows the result of the experiment using over-sampled instances of true positive plus the original data set.

Table 5. The accuracy of C4.5 decision tree for over-sampled instances of true positive plus the original data

Accuracy (%)		88.6447
TP rate	Class 1	0.907
	Class 2	0.862

Table 6 shows the corresponding confusion matrix.

Table 6. The confusion matrix of C4.5 decision tree for over-sampled instances of true positive plus the original data

		Predicted class	
		Class 1	Class 2
Actual class	Class 1	136	14
	Class 2	17	106

Table 7 shows the result of the experiment using over-sampled instances of false positive plus the original data set.

Table 7. The accuracy of C4.5 decision tree for over-sampled instances of false positive plus the original data

Accuracy (%)		90.4615
TP rate	Class 1	0.936
	Class 2	0.854

Table 8 shows the corresponding confusion matrix.

Table 8. The confusion matrix of C4.5 decision tree for over-sampled instances of false positive plus the original data

	Predicted class		
Actual class		Class 1	Class 2
	Class 1	189	13
	Class 2	18	105

We may wonder at the results in table 5 and table 7. So, more experiments were done. Because the total number instances of false positive instances is 25% more than the true positive instances in the over-sampled data, we did sample 118 instances randomly among 170 instances seven times for more objectivity. Table 9 shows the average of the experiment.

Table 9. The average accuracy of C4.5 decision tree for 118 over-sampled instances of false positive plus the original data

Accuracy (%)		87.7028
TP rate	Class 1	0.899
	Class 2	0.843

Table 10 shows the corresponding confusion matrix.

Table 10. The average of the confusion matrix of C4.5 decision tree for 118 over-sampled instances of false positive plus the original data

	Predicted class		
Actual class		Class 1	Class 2
	Class 1	134.9	15.1
	Class 2	18.4	104.6

If we compare table 5 and 9, we can find that the true positive instances by decision tree of the original data is doing better than the false positive instances. Note that adding over-sampled instances of class 1 may affect some decrease in true positive rate of class 2.

We also want to compare the naive application of SMOTE with our method. Because the data set generated by the naive application of SMOTE plus the original data has 170/118= 140% more data instances, we sampled only 118 instances from the SMOTED data of 900% over-sampling, and the sampled data were added to the original data. Table 11 and table 12 show the result of experiment.

Table 11. The accuracy of C4.5 decision tree for 118 over-sampled instances from the SMOTED instances of 900% plus the original data

Accuracy (%)		87.9121
TP rate	Class 1	0.913
	Class 2	0.837

Table 12 shows the corresponding confusion matrix.

Table 12. The confusion matrix of C4.5 decision tree for 118 over-sampled instances from the SMOTED instances of 900% plus the original data

	Predicted class		
Actual class		Class 1	Class 2
	Class 1	137	13
	Class 2	20	103

If we compare table 5 and 11, we can find that the true positive instances checked by decision tree of the original data is doing better than naive application of SMOTE.

If we look at the confusion matrix in table 2 that contains confusion matrix of the original data and table 6 that contains the confusion matrix of the over-sampled true positive instances plus original data, while the number of wrongly classified instances of class 1 is decreased by 4 (18 →14), the number of wrongly classified instances of class 2 is increased by 10 (7→17). But, because we are more interested in more accurate classification of class 1, this fact may not discourage our method.

The naive application of SMOTE also does not do better than our suggested method. That is, if we compare the confusion matrix in table 2 that contains the confusion matrix of the original data and table 12 that contains the confusion matrix of the 118 over-sampled SMOTED instances plus original data, while the number of wrongly classified instances of class 1 is decreased by 5 (18→13), the number of wrongly classified instances of class 2 is increased by 13 (7→20). So, the naive application of SMOTE is worse than our method. Remember that in our method the number of wrongly classified instances of class 1 is decreased by 4, while the number of wrongly classified instances of class 2 is increased by 10.

4 Conclusion

Hepatitis is a disease characterized by inflammatory cells in the tissues of the liver. It causes mild to serious effect to the liver so that patient may even die of it. On the other hand, decision trees are important data mining tools in medicine, because their structure is in tree shape, doctors can easily understand the final result of data mining so that they can be a good tool to diagnose the disease. But, decision trees have tendency to neglect minor classes. On the other hand, we are often interested in the better classification of the minor classes. Hepatitis data set contains 2 classes, 'die' and 'live'. Each class has 32 and 123 instances. So, over-sampling can be a solution, if we are more interested in better classification of the class, 'die'. Over-sampling technique based on synthetic data generation method like SMOTE has been considered a good technique. But, it may be possible that some of the artificially generated instances may not be proper for more accurate classification of the minor class. In this paper we showed how we may surmount the problem by resorting to the decision tree algorithm itself to select better artificial instances for target decision tree for the hepatitis data set. Experiments with various over-sampling rates showed a promising result.

References:

[1] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman, Decision Trees: An Overview and Their Use in Medicine, *Journal of Medical Systems*, Vol.26, No.5, 2002, pp. 445-463.

[2] J. Bae, The clinical decision analysis using decision tree, *Epidemiology and Health*, Vol.36, 2014.

[3] 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.

[4] 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.

[5] 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

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

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

[8] J. Magidson, The chaid approach to segmentation modeling: Chi-squared Automatic Interaction Detection, *Advanced Methods of Marketing Research*, 1994, pp. 118-159.

[9] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 algorithms in data mining, *Knowledge and Information System*, Vol. 14, 2008, pp.1-37.

[10] 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.