Better Decision Tree Induction for Limited Data Sets of Liver Disease

Hyontai Sug

Division of Computer & Information Engineering, Dongseo University,
47 Jurye-ro, Sa-sang-gu, Busan 617-716, Korea
sht@gdsu.dongseo.ac.kr

Abstract. Decision trees can be very useful data mining tools for human experts to diagnose the disease, because the knowledge structure is represented in tree shape. But we may not get satisfactory decision tree, if we do not have enough number of consistent instances in the data sets. Recently two kinds of relatively small data sets of liver disorder from America and India are available, so in order to generate more accurate and useful decision trees for the disease this paper suggests appropriate sampling for the data instances that are in the class of higher error rate. Experiments with the two public domain data sets and a representative decision tree algorithm, C4.5, shows very successful results.

Keywords: Decision trees, C4.5, liver disorder, sampling.

1 Introduction

Data mining tools have been being adapted more and more in the domain of medicine to diagnose disease more accurately based on clinical test data. Liver is one of the most important internal organs in the human body, and it is known that the organ is responsible for more than one hundred functions of human body. The complexity of this organ makes it not easy to diagnose the disease of disorder in the organ [1]. A lot of attention has been given to build more accurate models based on public domain data of liver disorder since the data set is available in 1990 [2], and most recent research is based on some artificial neural network based approach for better accuracy. But, even though trained neural networks can be transformed into some rule forms, because the rules are in flat structure, identifying major factors in classification can be difficult [3]. On the other hand, tree structures can give information on what are major factors of the disease. But, unfortunately the accuracy of trained decision trees may not be as good as that of neural networks. This is especially true when we do not have enough number of instances for training. But, because we can easily understand the knowledge structures of decision trees, they have been considered very good data mining tools in medicine domain [4, 5].

The training algorithms of decision trees have the tendency of neglecting minor class in tree building process to achieve maximum accuracy with respect to the whole

T.-h. Kim et al. (Eds.): MulGraB/BSBT/IUrC 2012, CCIS 353, pp. 88–93, 2012.

data set. Minor class is the class that has smaller number of instances and relatively higher error rates than the other classes. Therefore, we may need some means to compensate the property of decision trees for better utilization.

Related issues in generating decision trees are random sampling. Because we usually do not have a perfect data set for data mining, and we don't have exact knowledge about the property of data sets, we use random sampling [6]. Each random sampling could generate different training and test data sets, so each random samples could generate slightly different results. Moreover, since our target data sets have limited information, random sampling like over-sampling could generate more different results. More recently, another data set called ‘Indian liver patient data set’ becomes available since 2012 [17], so it’ll be interesting to compare the results of the two data sets. In order to see over-random sampling is effective, we want to do experiment based on the two different data sets of liver disorder disease and 10-fold cross validation for better objectivity in the experiment.

2 Related Work

A decision tree generation method, C4.5 [8], can be a representative decision tree algorithm, because the algorithm has been referred frequently and ranked number one by a survey in ICDM’06 [9]. When we build a decision tree, as each subtree in the tree is being built, each node will have smaller number of instances, so the reliability of lower part of the tree becomes worse than upper part of the tree. This problem is called fragmentation problem. Fragmentation problem can affect the training of decision tree, especially the data set has class imbalance in data composition. Class imbalance has different effect on over-sampling and under-sampling. Over-sampling means more duplicate instances are sampled from minor classes, while under-sampling means less number of instances are sampled from major classes than normal. In [10] five data sets that are mostly in large size are experimented using C4.5, and preferred under-sampling. In [11] four data sets consisting of 208~840 instances are experimented, and preferred under-sampling because it produced better sensitivity in misclassification cost. On the other hand, SMOTE [12] used synthetic data generation method as a way of over-sampling for minor classes, and showed that it is effective for nine different data sets in small to very large size. A weak point of the approach is that we need to understand the characteristics of data sets to synthesize the data. In [13] over-sampling was preferred for more accurate classification. In [14] an ensemble of neural networks are used to create new instances having different class values with original instances, and C4.5 with the 100% new instances of liver disorder data set showed worse accuracy of 67.8%, while C4.5 with the original data set showed the accuracy of 68.1%. From the reports of different results, we can conclude that for some data sets under-sampling can be more desirable, but for some other data sets, over-sampling can be more desirable.

For the liver disorder data set many researchers reported their results of experiments. In [15] sparse gird based approach achieved the accuracy of 72.5% for test data, but the approach does not consider symbolic representation of found models like artificial neural networks. In [16] artificial neural network based approach called artificial immune algorithm is used to find more accurate model, and achieved the accuracy of 94.8% for training data, and generated rules. In [17] four different data mining algorithms like Naïve Bayes classifier, C4.5, neural networks, and support vector machines were tried, and the accuracy of the algorithms ranges 56.52% ~ 71.59% in 10-fold cross-validation. Even though some artificial neural network based approach achieved high accuracy on training data, and we may drive rules from trained artificial neural networks, the rules are in flat structure so that determining major factors can be difficult, and moreover, we might confront with over-fitting problem.

3 Experimentation

We want to find better decision trees for the two liver disorder data sets of America and India. The size of data sets is relatively small. In this sense over-sampling the instances in the class of higher error rate or minor class to generate decision trees could improve our decision trees. In other words, because the splitting criterion of decision tree is heavily dependent on the number of instances, we increase the number of instances in the class of higher error rate by duplication to find decision trees of better accuracy. The following is the process.

```
Begin
   Do random sampling of 10-fold cross validation;
   Determine the class of higher error rate by generating decision tree of C4.5;
   For each fold Do
     R := 10%;
     Repeat
       Do R% more sampling for class of higher error rate;
       Generate decision trees of C4.5;
       Increase R by 10%;
     Until R = 200%;
   End For;
End.
```

The two data sets in UCI machine learning repository [18] called 'liver disorder' [2] and ‘Indian liver patient data set’ [7] were used. The liver disorder data set has the following properties: The number of instances is 345. There are 145 instances in class 1 and 200 instances in class 2 (disorder). Class 1 is the class of higher error rate, because its error rate is 50.1%, while the error rate of class 2 is 20.1% based on 10-fold cross-validation with C4.5. There are six continuous attributes as independent attributes. There are no missing values in all attributes. Table 1 shows the results of experiment. The averages of conventionally sampled data and the best of 20 over-sampled data are presented in 10-fold cross validation. The average accuracy of the liver disorder data set is slightly lower than the accuracy reported in other papers which is 64.6% ~ 68.9% in 10-fold cross validation [19]. Anyway, this difference comes from random sampling effect, so it doesn't matter for our experiment of over-sampling. As we can see in table 1, over-sampling gives 7% better accuracy with increase of 16 leaf nodes in the tree. In the table '±' symbol means standard deviation.

Table 1. Comparison of conventional and over-sampling for the liver disorder data set

Sampling method	Average accuracy	Average no. of leaves
Conventional	66.67%±5.94%	23.9±5.8
Over-sampling	73.35%±2.84%	39.9±15.7

'Indian liver patient data set' has the following properties; the number of instances is 583, and there are 167 instances in class 2 and 416 instances in class 1 (disorder). Note that class value has opposite meaning in the two data sets. There are nine continuous attributes as independent attributes, and one attribute has gender value. Small number of missing values exists in the data set. Class 2 is the class of higher error rate, because its error rate is 52.3%, while the error rate of class 1 is 17.8% based on 10-fold cross-validation with C4.5. As we can see in table 2, oversampling gives 5% better accuracy with increase of 26 leaf nodes in the tree.

Table 2. Comparison of conventional and over-sampling for Indian liver data set

Sampling method	Average accuracy	Average no. of leaves
Conventional	69.64%±7.39%	33.1±8.3
Over-sampling	74.63%±7.0%	58.9±14.3

4 Conclusions and Future Work

Decision trees have been considered one of the best data mining tools of understandability. But, weakness of decision trees arises due to the fact that their branching criteria give higher priority to the classes of majority. Two different data sets related to liver disorder attract our interest for data mining. The data sets are relatively small and have some error rates so that decision trees in conventional means may not generate good results due to the property.

In order to generate more accurate trees we used over-sampling technique for the data instances of the class of higher error rates. Experiments with a decision tree algorithm, C4.5, showed very good results. But, the trees become larger, so we may want to apply severer pruning for better understandability. Because pruning can generate smaller trees, we want to apply appropriate pruning parameters to generate comprehensible and accurate trees for the data sets.

A branch will be pruned, if predicted error rate decreases by pruning the branch. The upper limit of predicted error rate P for a leaf is calculated by $U_{CF}(e, t)$ function that is in binomial distribution. In the function e is the number of incorrectly classified instances and t is the number of training instances in a node. CF is confidence level of the predicted error rate. The number of predicted error for the leaf is t×P. This number is summed for each leaf in a subtree to calculate the number of predicted errors for the subtree. The number of predicted error is calculated for both of pruned state and unprunned state of a subtree, and if pruned state generates smaller number of predicted errors, the subtree will be pruned. Because P value is proportional to CF value, lower CF value can generate smaller predicted error rate. Default CF value in C4.5 is 25% and this values was set based on C4.5 developer's experience [8]. Moreover, as we can see in table 1 and 2, the standard deviations are somewhat large in our trees. Therefore, we have room to find proper CF value for accurate and understandable trees from the results of the experiment.

References

1. Ribeiro, R., Marinho, R., Velosa, J., Ramalho, F., Sanches, J.M.: Chronic liver disease staging classification based on ultrasound, clinical and laboratorial data. In: Proceedings of 2011 IEEE International Symposium on Biomedical Imaging from Nano to Macro, pp. 707–710 (2011)
2. UCI Machine Learning Repository, `http://archive.ics.uci.edu/ml/datasets/Liver+Disorders`
3. Zhou, Z., Jiang, Y., Chen, S.: Extracting symbolic rules from trained neural network ensembles. AI Communications 16(1), 3–15 (2003)
4. Podgorelec, V., Kokol, P., Stiglic, B., Rozman, I.: Decision trees: an overview and their use in medicine. Journal of Medical Systems 26(5), 445–463 (2002)
5. Lin, Y.C.: Design and Implementation of an Ontology-Based Psychiatric Disorder Detection System. WSEAS Transactions on Information Sciences and Applications 7(1), 56–69 (2010)
6. Tryfos, P.: Sampling for Applied Research: Text and Cases, Willy (1996)
7. Ramana, B.V., Babu, M.S.P., Venkateswarlu, N.B.: A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis. International Journal of Computer Science, 506–516 (2012)
8. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, Inc. (1993)
9. Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z., Steinbach, M., Hand, D.J., Steinberg, D.: Top 10 Algorithms in Data Mining. Knowledge Information System 14, 1–37 (2008)
10. Chawla, N.V.: C4.5 and Imbalanced data sets : Investigating the effect of sampling emthod, probalistic estimate, and decision tree structure. In: Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC (2003)
11. Drummond, C., Holte, R.C.: C4.5, Class Imbalance, and Cost Sensitivity: Why Under-sampling beats Over-sampling. In: Workshop on Learning from Imbalanced Datasets II, ICML, Washington DC (2003)
12. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research 16, 341–378 (2002)
13. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis 6(5), 429–449 (2002)

14. Zhou, Z., Jiang, Y.: NeC4.5: Neural Ensemble Based C4.5. IEEE Transactions on Knowledge and Data Engineering 16 (2004)
15. Garcke, J., Griebel, M.: Classification with sparse grids using simplicial basis function. Intelligent Data analysis 6 (2002)
16. Kahramanli, H., Allahverdi, N.: Mining Classification Rules for Liver Disorders. International Journal of Mathematics and Computers in Simulation 3(1), 9–19 (2009)
17. Ramana, B.V., Babu, M.S.P., Venkateswarlu, N.B.: A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis. International Journal of Database Management Systems 3(2), 101–114 (2011)
18. Frank, A., Suncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Sciences, Irvine (2010), `http://archive.ics.uci.edu/ml`
19. Zheng, Z.: Scaling up the Rule Generation of C4.5. In: Wu, X., Kotagiri, R., Korb, K.B. (eds.) PAKDD 1998. LNCS, vol. 1394, pp. 348–359. Springer, Heidelberg (1998)