Improving the Prediction Accuracy of Liver Disorder Disease with Oversampling

HYONTAI SUG
Division of Computer and Information Engineering
Dongseo University
Busan, 617-716
REPUBLIC OF KOREA
hyontai@yahoo.com http://kowon.dongseo.ac.kr/~sht

Abstract: - The complexity of liver makes it easily affected by disease of disorder. So diagnosing liver disorder disease is a high interest to data miners, and decision trees have been useful data mining tools to diagnose the disease, but the accuracy of decision trees has been limited due to insufficient data. In order to generate more accurate decision trees for liver disorder disease this paper suggests a method based on over-sampling in minor classes to compensate the insufficiency of data effectively. Experiments were done with two representative algorithms of decision trees, C4.5 and CART, and a data set, 'BUPA liver disorder', and showed the validity of the method.

Key-Words: - biased sampling, liver disorder disease, classification.

1 Introduction

Liver is the largest internal organ 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 easily affected by disease of disorder. So diagnosing liver disorder disease is a high interest to researchers and doctors [1], and decision trees have been a good data mining method to diagnose the disease [2, 3].

Decision trees have very good property that the structure is easy to understand, unless the size of the trees is not so large. This property of decision tree is important in case that human should understand the knowledge structures fully. This is one of the main reasons why decision trees are widely accepted in medical domain [4]. Another good point of decision trees is that it is very straightforward to transform decision trees into rules so that the rules can be used, for example, to build expert systems [5, 6]. So decision trees have been used in many applications [7, 8, 9, 10]. But, the training algorithms of decision trees have the weak of disdaining minor classes. Decision tree generation algorithms divide the training data set based on their own branching criteria. So, due to the dividing as each subtree is being built, each branch in the decision tree becomes to have less training instances. So, the reliability of lower branches becomes worse than upper branches due to the smaller size of training examples. Therefore, the classification accuracy for minor classes can become less accurate than that of major classes.

Another issue is random sampling. Because we may not have a perfect data set for data mining, and we usually don't have exact knowledge about the property of data sets, we may resort to random sampling [8]. But, the trained knowledge models based on random the samples are likely dependent on the samples. Moreover, due to the data fragmentation, it is known that the decision tree algorithms are more dependent upon the training data sets, while other machine learning algorithms like neural networks [12] that do not divide the data set during training are less dependent on.

We are interested in finding better decision trees for the data set of liver disorder 'BUPA liver disorder' that has relatively small number of instances. So in order to overcome the problem of neglecting minority classes with decision tree generation algorithms, we need a new technique so that the minor classes in the data set are treated more importantly.

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

2 Related Work

It is known that generating optimal decision trees is NP-complete problem [13] so that we rely on greedy algorithms to split branches. As a result, the generated decision trees may not be optimal. There have been a lot of efforts to build better decision trees [14]. Among them C4.5 [15] and CART [16] can be two representative decision tree algorithms, because the two algorithms are frequently referred, and C4.5 is often referred in engineering and business domain, while CART is often referred in medicine domain. C4.5 uses an entropy-based measure to split branches based on feature values, and the measure selects the most certain split among possible splits of candidate features. So classes that have more certain splits with respect to entropy are preferred. CART uses a purity-based measure, and the algorithm splits the training data set based on how probably the subsets become purer for a class, and it spends more time to generate smaller trees, so CART produces relatively smaller trees than C4.5. The splitting measure of the decision tree algorithms prefers the most certain split among possible splits of candidate features. So, major classes are preferred, because there are more instances of major classes in data set, and this fact makes it more certain in splitting.

Because the training process of decision trees is inductive process, and the data are fragmented in the training process, the performance of trained decision tree is heavily dependent on the composition of training data set. In [17, 18] class imbalance has different effect in neural networks for medical domain data so that we can see the importance of data set for the task of data mining. SMOTE [19] used synthetic data generation method for minor classes to cope with the situation of data shortage in minor classes, and showed that it is effective for decision trees. A weak point of the approach is that we need to understand the characteristics of data to synthesize effective data.

There has been much research interest for better prediction models for liver disorder disease. In [20, 21] undesirable features were eliminated to find better prediction models of neural networks, and an expert system was made based on the generated knowledge models. In [22] neural network having hidden layer of adaptive activation function and output layer of fixed sigmoid function were used, and better rules were generated than [21]. A potential problem of the neural network based approach in [22] is data overfitting, because their knowledge models have comparatively high accuracy without providing testing conditions. This fact was shown by other papers like in [23, 24]. In [23] 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 is 56.52% ~ 71.59% with 10-fold cross-validation. The accuracy of C4.5 is 68.69%. In [24] outliers were removed to improve k-NN and neural networks for the same data set.

3 The Method of Experimentation

We are interested in finding better decision trees for BUPA liver disorder data set [25]. Because the data set is relatively small and has somewhat high error rate, we want to compensate the property of disdaining minor classes in decision tree generation algorithms. Since decision tree algorithms do not give high priority to minor classes in splitting branches, it is highly possible that instances of minor classes are treated in the lower part of the tree, and this treatment may increase misclassification rate for 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. The following is a brief description of the procedure of the method.

INPUT: BUPA liver disorder data set,
OUTPUT: decision trees.
Begin
Do random sampling of size of 172, seven times.
For each sample data set **Do**
  Generate a decision tree for the sample data**;**
  Do while the accuracy of decision tree increases;
    Duplicate the instances of minor class;
    Generate a decision tree;
    End while;
  End Do;
  End.

In the algorithm we duplicate the instances of minor class by 100 percents until the accuracy of generated decision tree decreases. The sample size is half of the data set so that we have large enough data set for testing.

4 Experimentation

Experiments were run using a data set in UCI machine learning repository [26] called 'liver disorder' [25] to see the effect of the method. The number of instances is 345. There are 145 instances in class 1 and 200 instances in class 2. Class 1 is the minor class, because

its error rate is 68/145=46.9%, while the error rate of class 2 is 40/200=25% based on 10-fold cross-validation in C4.5. The overall error rate is 31.3043%. There are six continuous attributes as dependent attributes, and one attribute is class attribute that has value of 1 or 2. Please see table 1 for detail of the attributes.

Table 1. The meaning of attributes

Attribute name	Meaning
mcv	mean corpuscular volume
alkphos	alkaline phosphotase
sgpt	alamine aminotransferase
sgot	aspartate aminotransferase
gammagt	gamma-glutamyl transpeptidase
drinks	number of half-pint equivalents of alcoholic beverages drunk per day

C4.5 and CART were used to generate decision trees for seven random sample sets. Sample sets of size 172 were used. Remaining data were used for test. The following Table 2 and 3 show the accuracy for each decision tree algorithm with minor class over-sampling. In the table numbers in bold characters represent the best result in the sample set.

Table 2. The accuracy of decision tree by C4.5 for sample sets

Sampl Set #	Conv.	Over-samp. 100%	Over-samp. 200%	Over-samp. 300%
1	65.32%	**68.21%**	59.54%	NA
2	**64.74%**	62.43%	60.69%	NA
3	53.45%	**56.32%**	56.32%	NA
4	60.12%	61.27%	**64.16%**	60.34%
5	**66.47%**	63.58%	61.27%	NA
6	**63.58%**	57.23%	NA	NA
7	61.27%	**65.90%**	56.07%	NA

If we look at table 2, we notice that we can have better results with over-sampling in 4 out of 7. In the table 'Conv.' means 'Conventional sampling'.

Table 3. The accuracy of decision tree by CART for sample sets

Sample Set #	Conv.	Over-samp. 100%	Over-samp. 200%
1	65.32%	**67.05%**	52.60%
2	61.27%	**63.58%**	60.69%
3	**67.82%**	60.92%	NA
4	58.38%	**61.85%**	53.76%
5	65.90%	**69.94%**	67.63%
6	**60.69%**	57.23%	NA
7	62.43%	**65.32%**	63.58%

If we look at table 3, we notice that we can have better results with over-sampling in 5 out of 7.

All in all, we may say that the oversampling method is effective in decision tree algorithms and the 'liver disorder' data set, and it is more effective in CART decision tree algorithm.

5 Conclusions and future work

Liver is the largest internal organ 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 easily affected by disease of disorder. So diagnosing liver disorder disease is a high interest to researchers of data miners, and decision trees have been a good data mining method to diagnose the disease. Decision trees have been considered one of good data mining tools with respect to understandability and transformability. But, weakness of decision trees arises due to the fact that their branching criteria give higher priority for major classes. BUPA liver disorder data set that is our interest for data mining is relatively small and has high error rate so that it may be vulnerable due to the property of decision trees.

In order to overcome the problem of disdaining

minority classes of the data set in decision tree generation algorithms, we used over-sampling technique for minor classes. Experiments with two representative decision tree algorithms, C4.5 and CART, showed very good results so that we may recommend oversampling for the data set to generate decision trees. Future work is to see the effect of the method with smaller percentage of increase in minor class incrementally.

References:

[1]http://www.ehow.com/about_5048281_liver-disorders-diseases.html

[2] R. Ribeiro, R. Marinho, J. Velosa, F. Ramalho, J.M. Sanches, "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.

[3] R. Lin, "An intelligent model for liver disease diagnosis," *Artificial Intelligence in Medicine*, Vol. 47, issue 1, pp.53-62, 2009.

[4] V. Podgorelec, P. Kokol, B. Stiglic, I. Rozman, "Decision trees: an overview and their use in medicine," *Journal of Medical Systems,* Kluwer Academic/Plenum Press, Vol. 26, Num. 5, pp. 445-463, October 2002

[5] D. Chiang, W. Chen, Y. Wang, L. Hwang, "Rules Generation From the Decision Trees," *Journal of Information Science and Engineering*, Vol. 17, 2001, pp. 325-339.

[6] T. Tamai, M. Fujita, "Development of an expert system for credit card application assessment," *International Journal of Computer Applications in Technology,* Vol. 2, No. 4, 1989, pp. 1-7

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] P. Tryfos, *Sampling for Applied Research: Text and Cases*, Willy, 1996.

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

[13] S. Murthy, S. Salzberg, "Decision Tree Induction: How Effective is the Greedy Heuristic", in *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pp. 222-227, 1995.

[14] L. Rokach, O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, World Scientific Publishing Company, 2008.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] M. Neshat, M. Yaghobi, M. Naghibi, "Designing expert system of liver disorders by using neural network and comparing it with parametric and nonparametric system," in *Proceedings of 5th International Multi-Conference on Systems, Signals and Devices*, pp. 1-6, 2008.

[21] M. Neshat, M. Yaghobi, M. Naghibi, A. Esmaelzadel, "Fuzzy Expert System Design for Diagnosis of Liver Disorders," in 2*008 International Symposium on Knowledge Acquition and Modeling*, pp. 252-256, 2008.

[22] H. Kahramanli, N. Allahverdi, "Mining Classification Rules for Liver Disorders,"

International Journal of Mathematics and Computers in Simulation, Vol. 3, Issue 1, pp. 9-19, 2009.

[23] B. V. Ramana, M.S.P. Babu, N.B. Venkateswarlu, "A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis," *International Journal of Database Management Systems*, Vol. 3, No. 2, pp. 101-114, 2011.

[24] C. Dendek, J. Mań dziuk, "Improving Performance of a Binary Classifier by Training Set Selection," in *Proceedings of 18th International Conference on Artificial Neural Networks*, LNCS 5163, pp. 128-135, 2008.

[25]http://archive.ics.uci.edu/ml/datasets/Liver+Disorders

[26] A. Frank, A. Suncion, UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Sciences, 2010.