Improving Identification of Difficult Small Classes by Balancing Class Distribution

Jorma Laurikkala

Department of Computer and Information Sciences, University of Tampere,
FIN-33014 University of Tampere, Finland
Jorma.Laurikkala@cs.uta.fi

Abstract. We studied three methods to improve identification of difficult small classes by balancing imbalanced class distribution with data reduction. The new method, neighborhood cleaning rule (NCL), outperformed simple random and one-sided selection methods in experiments with ten data sets. All reduction methods improved identification of small classes (20-30%), but the differences were insignificant. However, significant differences in accuracies, true-positive rates and true-negative rates obtained with the 3-nearest neighbor method and C4.5 from the reduced data favored NCL. The results suggest that NCL is a useful method for improving the modeling of difficult small classes, and for building classifiers to identify these classes from the real-world data.

1 Introduction

Real-world data sets often have imbalanced class distribution, because many natural processes produce certain observations infrequently. For example, rare diseases in a population may result in medical data with small diagnostic groups. When some classes are heavily under-represented, statistical and machine learning methods are likely to run into problems. Cases from the rare classes are lost among the other cases during learning. The resulting classifiers misclassify new unseen rare cases, and descriptive models may give an inadequate picture of the data. The learning task is even more problematic, if a small class is difficult to identify because of its other characteristics. A small class may, for example, overlap heavily the other classes. In the following, we refer to a small and difficult class as a *class of interest.*

We balanced class distribution with data reduction before the actual analysis, because we aimed to develop a general-purpose method, whose results may be given directly to statistical and machine learning methods. The most well-known data reduction technique comes from the area of statistics, where sampling [1] is used to allow analyses which would be impractical with large populations. Data reduction has been utilized in the area of machine learning especially to accelerate the instance-based learning methods [2,3]. Recently Kubat *et al.* [4] presented one-sided selection (OSS) which uses instance-based methods to reduce the larger class when class distribution of a two-class problem is imbalanced. In this paper, we describe a new method called *neighborhood cleaning rule* that utilizes the OSS principle, but considers more carefully the quality of the data to be removed.

S. Quaglini, P. Barahona, and S. Andreassen (Eds.): AIME 2001, LNAI 2101, pp. 63–66, 2001.

2 Methods and Materials

Simple random sampling (SRS), which was used as a baseline method, is the most basic one of the sampling methods applied in statistics [1]. In SRS a sample (sub-set) S is selected randomly from the original data T so that each example in T has an equal probability to be selected into S. We applied SRS to classes that were larger than class of interest C and selected a sample with size of $| C |$ from each of these classes. Unfortunately, the within classes SRS (SWC) may produce biased samples, because small samples may have an over-representation of outliers or noisy data.

One-sided selection (OSS) [4] reduces T by keeping all examples of C and by removing examples from the rest of data $O = T - C$. Firstly, Hart's condensed nearest neighbor rule (CNN) [3,4] is applied to select a sub-set A from T which is consistent with T in the sense that A classifies T correctly with the one-nearest neighbor rule (1-NN). CNN starts from S, which contains C and an example from each class in O, and moves examples misclassified by 1-NN from O to S, until a complete pass over O has been done without misclassifications. Secondly, examples that are noisy or lie in the decision border are removed from O. The major drawback of OSS is CNN rule which is extremely sensitive to noise [3]. Since noisy examples are likely to be misclassified, many of them will be added to the training set. Moreover, noisy training data will misclassify several of the subsequent test examples [2,3]. We also argue that data cleaning should be done before the data analysis. For example, in statistical analyses and data mining, data pre-processing is an important step before the actual analysis.

The basic idea of our *neighborhood cleaning rule* (NCL) is the same as in OSS: All examples in C are saved, while O is reduced. In contrast to OSS, NCL emphasizes more data cleaning than data reduction. Our justification for this approach is two-fold. Firstly, the quality of classification results does not necessarily depend on the size of the class. Therefore, we should consider, besides the class distribution, other characteristics of data, such as noise, that may hamper classification. Secondly, studies of data reduction with instance-based techniques [3] have shown that it is difficult to maintain the original classification accuracy while the data is being reduced. This aspect is important, since while improving the identification of small classes, the method should be able to classify the other classes with an acceptable accuracy.

Consequently, we chose to use Wilson's edited nearest neighbor rule (ENN) [3] to identify noisy data A_1 in O. ENN removes examples whose class differs from the majority class of the three nearest neighbors. ENN retains most of the data, while maintaining a good classification accuracy [3]. In addition, we clean neighborhoods of examples in C: The three nearest neighbors that misclassify examples of C and belong to O are inserted into set A_2. To avoid excessive reduction of small classes, only examples from classes larger or equal to $0.5 \cdot | C |$ are considered while forming A_2. Lastly, the union of sets A_1 and A_2 is removed from T to produce reduced data S. Fig. 1 illustrates the NCL rule. To make NCL to suit better for solving real-world problems than OSS, we utilized Heterogeneous value difference metric (HVDM) [3] and designed NCL with multi-class problems in mind.

Experiments were made with ten real-world data sets of which eight were retrieved from UCI machine learning repository [5]. Six of these data sets were medical data which is our primary application area. Female urinary incontinence [6] and vertigo [7] data sets are medical data which we have studied earlier with different methods. The characteristics of the data sets, as well as the classes of interest are reported in [8].

```
1. Split data $T$ into the class of interest $C$ and the rest of data $O$.
2. Identify noisy data $A_1$ in $O$ with edited nearest neighbor rule.
3. For each class $C_i$ in $O$
       if ( $x \in C_i$ in 3-nearest neighbors of misclassified $y \in C$ )
       and ( $| C_i | \quad 0.5 \cdot | C |$ ) then $A_2 = \{ x \} \cup A_2$
4. Reduced data $S = T - ( A_1 \cup A_2 )$
```

Fig. 1. Neighborhood cleaning rule

We applied the three data reduction methods to the whole data as in [4] and to the training sets of 10-fold cross-validation process as in [3]. The data were classified with the three-nearest neighbor (3-NN) method (with HVDM) and C4.5 (release 8, default settings) [9]. Classification measures were accuracy, true-positive (TPR) and true-negative rates (TNR), and the true-positive rate of the class of interest (TPRC). The two-tailed Wilcoxon signed ranks test was used to test the significance of differences in the measures (p<0.05).

3 Results

Mean accuracies, TPRs, TNRs, and TPRCs from the original data were 76, 68, 73, 39 for 3-NN and 76, 69, 72, 40 for C4.5, respectively. Table 1 shows the changes in means of classification measures from the reduced data in comparison with the mean results from the original data. More detailed results are reported in [8]. With reduced training and test sets, the differences in accuracy, TPRs and TNRs were significant and in favor of NCL (NCL>SWC>OSS). With reduced training and original test sets, the following differences were significant in 3-NN measures: Accuracy: NCL>OSS, TPR: SWC>OSS and NCL>OSS, and TNR: NCL>SWC and NCL>OSS. In all the C4.5 measures, except TPRC, SWC>OSS and NCL>OSS were the significant differences. All statistically significant differences were in favor of NCL. There were no significant differences in TPRCs with both types of test data.

Table 1. Changes in means of accuracies (a), true-positive rates (tpr), true-negative rates (tnr) and true-positive rates for the classes of interest (c) in percents from the reduced data in comparison with mean results from the original data. Balanced (*) and original (**) test data.

Method	SWC				OSS				NCL			
	a	tpr	tnr	c	a	tpr	tnr	c	a	tpr	tnr	c
3-NN*	-7	1	-4	22	-19	-23	-17	25	11	16	13	27
C4.5*	-4	4	-1	26	-14	-13	-11	22	5	9	8	25
3-NN**	-7	1	-4	24	-8	-6	-4	25	-5	1	-1	23
C4.5**	-6	4	-3	30	-10	-8	-6	25	-6	-1	-1	20

4 Discussion

The classification results obtained from the reduced original data sets showed that the NCL method was significantly better than SWC and OSS. Only the differences in TPRCs were insignificant. All reduction methods improved clearly (22-27%) these rates in comparison with the results of original data. NCL was the only reduction method which resulted in higher accuracies, TPRs and TNRs than those of the original data sets. NCL was able to overcome the noise related drawbacks of OSS. NCL attempts to avoid the problems caused by noise by applying the ENN algorithm that is designed for noise filtering. NCL also cleans neighborhoods that misclassify examples belonging to the class of interest. The results suggest that NCL is a useful tool to build descriptive models that take better into account difficult small classes.

NCL was also the best method when the test data had the original imbalanced class distribution. All reduction methods improved clearly (20-30%) TPRCs in comparison with the results of original data. Although the other classification measures of NCL method were lower than the original ones, decrease was slight and slightly smaller in comparison with OSS. Our method may also be useful in building better classifiers for new unseen rare examples for the real-world data with imbalanced class distribution. NCL might allow us, for example, to generate classifiers that are able to identify examples of the sensory urge class [6] better than the classifiers built from the original data. This type of classifiers would be very useful in an expert system which we plan to develop to aid physicians in the differential diagnosis of female urinary incontinence [6]. In the preliminary tests with two-fold cross-validation, we have found that 3-NN classifier with NCL reduced training sets have improved on average 20% the TPRs of the difficult sensory urge class.

References

1. Cochran, W.G.: Sampling Techniques. 3rd edn. Wiley, New York (1977)
2. Aha, D.W., Kibler, D., Albert, M.K.: Instance-Based Learning Algorithms. Mach. Learn. **6** (1991) 37-66
3. Wilson, D.R., Martinez, T.R.: Reduction Techniques for Instance-Based Learning Algorithms. Mach. Learn. **38** (2000) 257-286
4. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: Fisher, D.H. (ed.): Proceedings of the Fourteenth International Conference in Machine Learning. Morgan Kaufmann, San Francisco (1997) 179-186
5. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, University of California, Department of Information and Computer Science (1998)
6. Laurikkala, J., Juhola, M., Lammi, S., Penttinen, J., Aukee P.: Analysis of the Imputed Female Urinary Incontinence Data for the Evaluation of Expert System Parameters. Comput. Biol. Med. **31** (2001)
7. Kentala, E.: Characteristics of Six Otologic Diseases Involving Vertigo. Am. J. Otol. **17** (1996) 883-892
8. Laurikkala J.: Improving Identification of Difficult Small Classes by Balancing Class Distribution [ftp://ftp.cs.uta.fi/pub/reports/pdf/A-2001-2.pdf]. Dept. of Computer and Information Sciences, University of Tampere, Tech. Report A-2001-2, April 2001
9. Quinlan, J.R.: C4.5 Programs for Machine Learning. Morgan Kaufman, San Mateo (1993)