IMPROVING IDENTIFICATION OF DIFFICULT SMALL CLASSES BY BALANCING CLASS DISTRIBUTION

Jorma Laurikkala

DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES
UNIVERSITY OF TAMPERE
REPORT A-2001-2

UNIVERSITY OF TAMPERE
DEPARTMENT OF COMPUTER AND INFORMATION SCIENCES
SERIES OF PUBLICATIONS A
A-2001-2, APRIL 2001

IMPROVING IDENTIFICATION OF DIFFICULT SMALL CLASSES BY BALANCING CLASS DISTRIBUTION

Jorma Laurikkala

Department of Computer and Information Sciences
P.O. Box 607
FIN-33014 University of Tampere, Finland

ISBN 951-44-5093-0
ISSN 0783-6910

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 different methods to improve identification of small classes, which are also difficult to classify, by balancing imbalanced class distribution with data reduction. The new method, neighborhood cleaning rule (NCL), outperformed simple random selection within classes and one-sided selection method in experiments with ten real-world data sets. All reduction methods improved clearly identification of small classes (20-30%), but differences between the methods were insignificant. However, the significant differences in accuracies, true-positive rates and true-negative rates that were obtained with the three-nearest neighbor method and C4.5 decision tree generator from the reduced data were in favor of NCL. The results suggest that NCL is a useful method for improving modeling of difficult small classes, as well as for building classifiers that identify these classes from the real-world data which often have an imbalanced class distribution.

1 Introduction

Real-world data sets often have a non-uniform class distribution, because many natural processes produce certain observations infrequently. For example, medical databases in UCI machine learning repository [1] have small diagnostic groups, because some diseases are rare in the population from which the data were collected. We have encountered this problem also with our medical data sets, such as female urinary incontinence [2] and vertigo data [3], which we have analyzed with different methods [2-5]. Acquisition of examples of certain classes may be more difficult than collection of the rest of the data, and insufficient resources may sometimes lead to imbalanced class distribution. Rare classes are especially problematic when data is scarce. For example, the detection of oil spills in satellite-borne radar images [6,7] is troublesome, because of the high cost of such images and limited time of experts. As a result, data is scarce and, moreover, due to imbalanced class distribution, positive examples (true oil spills) are very rare.

A small imbalance in the class distribution is not serious, but when some classes are heavily under-represented, many statistical and machine learning methods are likely to run into problems. Cases belonging to small classes are lost among the more frequent cases during learning, and, consequently, classifiers such as decision rules or trees are unable to classify correctly new unseen cases from the minority classes. The learning task is even more problematic, if the small class is difficult to identify not

only because of its size, but also because of its other characteristics. A small class may, for example, overlap heavily the other classes. Moreover, imbalanced class distribution may hamper descriptive analysis, where models describing the data are constructed. The models may give an inadequate picture of the data, if the information from the small classes is not fully included into them.

One approach to overcome imbalanced class distribution is data reduction before the actual analysis. There are also other approaches, such as generating artificial data [7,8], weighing training cases [6,7], and introducing different misclassification costs [6,7]. We chose to adopt the data reduction approach, because we aimed to develop a general-purpose method, whose results may be given directly to statistical and machine learning methods. Generation of the artificial data has been considered in the neural-network community [7,8], but the characteristics of the data should be considered very carefully in this approach to avoid biasing the data. Other above mentioned approaches would require modification of the analysis method. There is also some research of methods (for example, BRUTE and SHRINK systems) that are insensitive to the underlying class distribution in the training set [6].

The most well-known data reduction technique comes from the area of statistics, where sampling [9] is used to allow analyzes which would otherwise be impossible or impractical to conduct, because of the large size of the population. Another tradition of data reduction exists in the areas of pattern recognition and machine learning, where research has been motivated especially by the need to accelerate instance-based learning methods [10,11] such as the nearest neighbor classification. Recently Kubat *et al.* [7] presented an instance-based data reduction method called *one-sided selection* which utilizes Hart's condensed nearest neighbor rule [12] to reduce the larger class when the class distribution of a two-class problem is imbalanced. In this paper, we describe a new method called *neighborhood cleaning rule* that utilizes the one-sided selection principle, but considers more carefully the quality of the data to be removed.

The paper is organized as follows. In Section 2, we present data reduction methods of this study. Then, the experiments and the ten real-world data sets, which were used as the test material, are described in Section 3. Classification results obtained from the reduced data with different methods with the three-nearest neighbor method and C4.5 decision tree generator are reported in Section 4. Lastly, we discuss the results in Section 5 and draw some conclusions in Section 6.

2 Methods

In this section, we present the one-sided selection method, neighborhood cleaning rule, and *simple random sampling within classes* which was used as the baseline method for the more advanced methods. In addition, we discuss of the shortcomings of the methods and how the drawbacks of one-sided sampling have been addressed in our method. We chose, from each data set, a class that was considerably smaller than the largest class and was also difficult to classify correctly. We will denote this class as the *class of interest*. It is important that the class is difficult to identify, because there are classes which may be identified well, even if they are quite small in

comparison with other classes. For example, in the female urinary incontinence data "normal" class [2], which includes only 3% of the data, can be easily differentiated from the other classes.

2.1 Simple Random Sampling within Classes

Simple random sampling (SRS) is the most basic one of the sampling methods applied in statistics [9]. In SRS a sample (sub-set) S is selected randomly from the population T so that each observation in T has an equal probability to be selected into S. We applied SRS to classes that were larger than the class of interest C and selected a sample with size of $| C |$ from each of these classes. If C was the smallest class, then the class distribution in S was uniform. SRS gives an unbiased sample from the original data with reasonable large sample sizes. Unfortunately, the sample S may be biased, if we balance class distribution by SRS within classes (SWC). Although the large classes may be large enough for sampling, the samples themselves may be too small. Small samples are often statistically unrepresentative, because they may have, for example, an over-representation of outliers or noisy data.

2.2 One-Sided Selection

Instance-based learning methods [10,11] are "lazy" learners which simply store learning instances and postpone processing until they need to classify new instances. Reduction of training instances results in smaller storage requirements and faster classification of unseen cases. Different data reduction methods have been studied widely in the context of nearest neighbor classification. Wilson *et al*. survey several of these methods in [12].

One-sided selection (OSS) [7] reduces the original data T by keeping all examples of the class of interest C and by removing examples from the rest of data $O = T - C$. Firstly, Hart's condensed nearest neighbor rule (CNN) [7,12] is applied to select a sub-set A from the original data T which is consistent with T in the sense that A classifies T correctly with the one-nearest neighbor method (1-NN). CNN starts from S, which contains C and one randomly selected 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 belong to O and participate in Tomek links [7,12] are removed from A. Tomek link is defined as a pair of examples x and y from different classes, that there exists no example z such that $d(x,z) < d(x,y)$ or $d(y,z) < d(x,y)$, where d is the distance between a pair of examples. Examples in Tomek links are noisy or lie in the decision border.

2.3 Neighborhood Cleaning Rule

The major drawback of the OSS method is the CNN rule which is extremely sensitive to noise [12]. Since noisy examples are likely to be misclassified, many of them will be added to the training set. Moreover, noisy training data will classify incorrectly

several of the subsequent testing examples. This type of behavior was apparent in Aha's IB2 algorithm which closely resembles the CNN rule [10,12]. Although OSS applies Tomek links to remove noise, it is clear that the reduced data is not the best possible due to the CNN rule. We also consider that data cleaning (or editing) should be done before the actual data analysis, as usual. For example, identification and possible removal of outliers prior to the linear regression analysis gives more reliable results [13]. Data pre-processing is also an important step before applying a data mining algorithm in knowledge discovery in databases [14]. A recent study [15] lends additional support to our critique: Dasarathy *et al*. found that data editing should be performed before data reduction.

The basic idea of our method is the same as in OSS: All examples in the class of interest C are saved, while the rest O of the original data T 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. There are small classes that identify well and large classes that are difficult to classify. 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 [12,15] 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) [12] to identify noisy data A_1 in O. ENN removes examples whose class label differs from the majority class of the three nearest neighbors. ENN retains most of the data, while maintaining a good classification accuracy [12]. 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 than $0.5 \cdot | C |$ are considered while forming A_2. Lastly, the union of sets A_1 and A_2 is removed from T to produce the reduced data set S. Since our method considers data cleaning in neighborhoods from two viewpoints, it was named as the neighborhood cleaning rule. Fig. 1 illustrates the NCL rule.

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 the edited nearest neighbor rule.
3. For each class C_i in O

   if ($x \in C_i$ in the 3-nearest neighbors of misclassified $y \in C$)

   and ($| C_i | \geq 0.5 \cdot | C |$) then $A_2 = \{ x \} \cup A_2$
4. Reduced data $S = T - (A_1 \cup A_2)$

Fig. 1. Neighborhood cleaning rule

We wanted also to develop our method to suit better for solving real-world problems than OSS. The OSS method uses the Euclidean distance metric which is not the best

possible distance measure for mixed data, i.e. data described with qualitative and quantitative attributes (or variables). Since the real-world data is frequently mixed, we chose to use the heterogeneous value difference metric (HVDM) [16] which treats nominal attributes more appropriately than the Euclidean metric. In addition, Kubat *et al*. studied only two-class problems, while we designed our method with multi-class problems in mind.

3 Materials and Experimental Setup

Experiments were made with ten real-world data sets of which eight were retrieved from UCI machine learning repository [1]. Six of these data sets were medical data which is our primary application area. Female urinary incontinence [2] and vertigo [3] data sets are medical data which we have studied earlier with different methods [2-5]. Missing values in these data sets were filled in with the Expectation-Maximization imputation method [2,17] and the nine missing values in Ljubljana breast cancer data were replaced with modes. The other data sets were complete.

The non-medical flags and glass data sets were included in the study, due to the lack of medical data sets that were both complete and had a large enough class of interest to allow reliable cross-validation. Breast cancer, flags, incontinence and vertigo data sets had mixed attributes, whereas the other data sets were characterized by quantitative attributes only. The other characteristics of the data sets, as well as the classes of interest are shown in Table 1. The classes of interest represented in the multi- and two-class problems 4-16% and 27-42% of the data, respectively.

Table 1. Characteristics of the data sets and classes of interest. N = sizes of data sets and classes of interest, N_C = number of classes, N_A = number of attributes

Data set				Class of interest	
Name	N	N_C	N_A	Name	N
Breast cancer	286	2	9	Recurrence-events	85
Buba	345	2	6	Sick	145
Ecoli	336	8	7	iMu	35
Flags	194	8	28	White	17
Glass	214	6	9	Vehicle-windows-float-processed	17
Haberman	306	2	3	Died	81
Incontinence	529	5	11	Sensory urge	33
New-thyroid	215	3	5	Hyper	35
Pima	768	2	8	Positive	268
Vertigo	564	6	38	Sudden deafness	21

We applied the three data reduction methods to the whole data sets before cross-validation as in [7] and to the training sets of cross-validation process as in [12]. The data sets were classified with the three-nearest neighbor (3-NN) method (with HVDM), and C4.5 decision tree generator (release 8, default settings) [18].

The data were reduced differently, because we wanted to test how the 3-NN and C4.5 methods could classify test data both with the balanced and original imbalanced

class distributions. Analysis with the balanced class distribution gave directions of how descriptive analysis, i.e. analysis of the data without cross-validation, would be affected by data reduction. In addition, this type of analysis revealed clearly the differences in the relative performance of the data reduction methods. On the other hand, runs with test sets having the original class distribution and the reduced training sets, showed how the reduction methods would affect the classification of real-world data. This was an important aspect of our study, because we are, besides modeling of the data, interested in building classifiers for medical data [2,4,5].

Classification ability of the learning methods was measured with accuracy, (the ratio of correctly classified test examples to all the test examples), true-positive rate (the ratio of correctly classified positive test examples to all the positive test examples), and true-negative rate (the ratio of correctly classified negative test examples to all the negative test examples). In addition, we recorded the true-positive rates of the class of interest. Due to small sample sizes (10 pairs in each comparison) the two-tailed Wilcoxon signed ranks test [19] was used instead of the paired *t* test to examine whether differences in the measures were significant ($p<0.05$).

4 Results

Accuracies, true-positive rates (TPR), true-negative rates (TNR), and true-positive rates for the classes of interest (TPRC) were calculated from a contingency table which was created by summing the frequencies of 10 contingency tables from the 10-fold cross-validation process. Since the number of TPRs and TNRs ranged from 2 to 8, medians of these measures were reported. Table 2 shows that the classification results of 3-NN and C4.5 methods with the original data sets were similar, and TPRCs were low in comparison with TPRs.

Tables 3 and 4 report the classification measures of the learning methods for the *reduced original data sets* and the changes of means of measures in comparison with the mean results from the original data (Table 2). Differences in accuracy, TPRs and TNRs were significant and in favor of NCL (NCL>SWC>OSS), while there were no significant differences in TPRCs. NCL reduction produced the best results. Improvements in TPRCs were 22-27% and 22-26%, respectively. SWC, OSS and NCL methods removed on average 51%, 60% and 24% of the original data, respectively.

Tables 5 and 6 show the classification measures of the learning methods for the *reduced training cross-validation sets* and the changes of means of measures in comparison with the mean results from original data (Table 2). The following differences were significant in the 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. Again, there were no significant differences in TPRCs. Improvements in TPRCs were 23-25% and 20-30%, respectively. SWC, OSS and NCL methods removed on average 51%, 60% and 25% of the original data, respectively.

Table 2. Accuracies (a), true-positive rates (tpr), true-negative rates (tnr) and true-positive rates for class of interest (c) in percents from 10-fold cross-validation with 3-NN method and C4.5 decision tree generator with original data sets

Data set	3-NN				C4.5			
	a	tpr	tnr	c	a	tpr	tnr	c
Breast cancer	66	56	56	31	72	58	58	24
Buba	64	62	62	51	67	65	65	51
Ecoli	84	80	84	49	83	65	83	49
Flags	58	25	58	6	55	33	55	6
Glass	72	65	72	35	66	69	66	29
Haberman	70	58	58	33	70	57	57	31
Incontinence	85	80	86	36	88	87	88	55
New-thyroid	94	89	94	80	93	94	93	83
Pima	73	69	69	56	76	72	72	60
Vertigo	90	95	90	14	87	93	86	14
Mean	76	68	73	39	76	69	72	40

Table 3. Accuracies (a), true-positive rates (tpr), true-negative rates (tnr) and true-positive rates for classe s of interest (c) in percents from 10-fold cross-validation with 3-NN method and data sets reduced with different methods. Change of means in comparison with results from original data.

Data set	SWC				OSS				NCL			
	a	tpr	tnr	c	a	tpr	tnr	c	a	tpr	tnr	c
Breast cancer	58	58	58	54	48	47	47	53	78	77	77	68
Buba	64	64	64	64	54	46	46	75	82	78	78	91
Ecoli	78	80	78	74	70	52	71	71	95	92	95	86
Flags	49	29	50	41	40	9	40	18	75	47	76	48
Glass	62	64	62	29	32	18	33	41	80	86	79	24
Haberman	62	62	62	62	46	45	45	58	82	80	80	67
Incontinence	74	82	73	67	64	55	64	76	96	97	96	64
New-thyroid	90	93	88	83	75	38	79	90	98	100	97	83
Pima	70	70	70	69	61	55	55	77	86	86	86	82
Vertigo	82	83	81	67	77	81	76	81	96	100	95	48
Mean	69	69	69	61	57	45	56	64	87	84	86	66
Change	-7	1	-4	22	-19	-23	-17	25	11	16	13	27

Table 4. Accuracies (a), true-positive rates (tpr), true-negative rates (tnr) and true-positive rates for classes of interest (c) in percents from 10-fold cross-validation with C4.5 and data sets reduced with different methods. Change of means in comparison with results from original data.

Data set	SWC				OSS				NCL			
	a	tpr	tnr	c	a	tpr	tnr	c	a	tpr	tnr	c
Breast cancer	63	63	63	55	50	51	51	47	75	74	74	67
Buba	62	62	61	62	68	64	64	78	69	66	66	77
Ecoli	70	70	68	60	62	57	60	57	90	80	90	77
Flags	53	41	53	53	42	34	43	24	63	39	64	35
Glass	75	80	74	29	33	29	32	41	81	83	81	53
Haberman	70	70	70	67	60	60	60	57	74	71	71	56
Incontinence	77	82	76	64	76	76	76	70	94	99	94	64
New-thyroid	91	90	92	90	80	50	84	93	94	91	95	90
Pima	72	72	72	76	71	64	64	88	80	80	80	77
Vertigo	85	98	82	100	74	76	74	67	90	93	89	53
Mean	72	73	71	66	62	56	61	62	81	78	80	65
Change	-4	4	-1	26	-14	-13	-11	22	5	9	8	25

Table 5. Accuracies (a), true-positive rates (tpr), true-negative rates (tnr) and true-positive rates for classes of interest (c) in percents from 10-fold cross-validation with 3-NN method and training data sets reduced with different methods. Change of means in comparison with results from original data.

Data set	SWC				OSS				NCL			
	a	tpr	tnr	c	a	tpr	tnr	c	a	tpr	tnr	c
Breast cancer	59	57	57	52	58	60	60	65	57	59	59	65
Buba	58	58	58	59	59	62	62	80	55	59	59	85
Ecoli	81	74	81	69	82	66	83	79	84	81	84	77
Flags	47	37	48	41	50	17	50	6	53	39	54	41
Glass	58	65	58	24	58	57	57	29	62	59	61	24
Haberman	61	62	62	64	53	57	57	64	62	62	62	63
Incontinence	82	79	82	79	82	65	83	79	85	74	85	58
New-thyroid	92	93	91	80	87	80	89	80	93	86	94	80
Pima	71	72	72	73	66	69	69	77	71	73	73	80
Vertigo	82	90	81	86	85	86	84	76	90	96	89	48
Mean	69	69	69	63	68	62	69	64	71	69	72	62
Change	-7	1	-4	24	-8	-6	-4	25	-5	1	-1	23

Table 6. Accuracies (a), true-positive rates (tpr), true-negative rates (tnr) and true-positive rates for classes of interest (c) in percents from 10-fold cross-validation with C4.5 and training data sets reduced with different methods. Change of means in comparison with results from original data.

Data set	SWC				OSS				NCL			
	a	tpr	tnr	c	a	tpr	tnr	c	a	tpr	tnr	c
Breast cancer	59	60	60	60	58	57	57	55	57	60	60	67
Buba	68	67	67	64	54	56	56	70	55	58	58	77
Ecoli	77	71	77	80	77	60	77	54	82	71	82	71
Flags	53	49	53	47	50	26	48	18	53	46	54	47
Glass	58	73	57	71	51	52	52	82	61	60	61	24
Haberman	66	62	62	52	63	59	59	49	66	61	61	51
Incontinence	82	89	82	61	82	69	84	61	86	78	86	64
New-thyroid	88	91	87	97	80	80	80	87	91	87	91	87
Pima	70	70	70	71	65	69	69	83	68	70	70	76
Vertigo	76	94	75	95	75	84	73	91	85	90	84	33
Mean	70	73	69	70	66	61	66	65	70	68	71	60
Change	-6	4	-3	30	-10	-8	-6	25	-6	-1	-1	20

5 Discussion

We studied three data reduction methods, simple random selection within classes (SWC), one-sided selection (OSS) and neighborhood cleaning rule (NCL), which were applied to balance imbalanced class distributions to allow better identification of small classes that were also difficult to identify (class of interest). NCL is a new method which is based on the OSS idea, but addresses the shortcomings of the OSS method. The effect of data reduction to the classification ability of the actual data analysis methods was studied with the 3-NN and C4.5 classifiers.

The classification results obtained from the reduced original data sets with the 3-NN and C4.5 showed that the NCL method was significantly better than SWC and OSS, while SWC outperformed OSS (Tables 3 and 4). Only the differences in TPRs of the class of interest were insignificant. All reduction methods improved clearly (22-27%) these rates in comparison with the results from the original data. NCL was the only reduction method which resulted in higher accuracies, TPRs and TNRs than those of the original data sets.

The results showed that NCL was able to overcome the drawbacks of OSS method. 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. In addition, NCL removed approximately 25% of the data, while the other methods dropped over 50% of the data. This is partly due to the ENN algorithm which saves the majority of the data [12] and, consequently, manages to keep the accuracies near the original level. Although we did not study descriptive analysis, i.e. modeling of the whole data without cross-

validation, the more objective cross-validation results suggest that the NCL method is a useful tool to build models that take better into account the difficult small classes.

According to statistically significant differences, NCL was also the best method as measured with accuracies, TPRs and TNRs, when the test data had the original imbalanced class distribution, and the training data sets of the cross-validation process were reduced (Tables 5 and 6). As with the balanced test sets, there were no statistically significant differences in TPRs of the classes of interest. All reduction methods improved clearly (20-30%) identification of these classes in comparison with the results of original data. Although the other classification measures of the NCL method were lower than the original ones, the decrease was slight and slightly smaller in comparison with OSS. These results suggest that our method may also be useful in building better classifiers for new unseen cases for the real-world data with imbalanced class distribution. Accuracies, TPRs and TNRs dropped only slightly, while TPRs of the classes of interest improved clearly.

NCL might allow us, for example, to generate classifiers that are able to identify examples of the sensory urge class better than the classifiers built from the original female urinary incontinence 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 [2]. In the preliminary tests with the two-fold cross-validation, we have found that a 3-NN classifier with the NCL reduced training sets has improved on average 20% the TPRs of the difficult sensory urge class, but the TPRs and TNRs of the larger classes have dropped slightly. As with the data analysis methods, the results of data reduction methods depend to some degree on the data and its partitions.

Unexpectedly, SWC, which was included in the study as a baseline method, outperformed OSS. A closer analysis revealed that the original data sets contained on average 17% noise identified by the ENN rule. Data sets reduced with SWC had approximately the same noise level as the original data, while the OSS reduced original data and cross-validation learning sets had on average 29% and 28% noise, respectively. It seems that the small sample sizes were clearly less problematic for SWC than the noise sensitiveness of the CNN rule for OSS.

There are some limitations to this study. Firstly, we did not study data with missing values, which are quite common in medical data, because the aim of our work was to study the applicability of the NCL rule. Secondly, it might be possible to reduce more the large classes by removing redundant data. Thirdly, since our method uses a proximity matrix, it takes $O(N^2)$ time, where N is the number of examples. A data pre-processing method should be efficient, because nowadays knowledge is mined quite often from very large data sets. We plan to address these limitations in future research, as well as other approaches, for example, bagging to improve the identification of the small difficult classes.

6 Conclusion

The neighborhood cleaning rule outperformed the other data reduction methods of this study as measured with the classification ability of the 3-NN and C4.5 classifiers. Our method seems to allow improved identification of difficult small classes both in descriptive and predictive analyzes, while keeping the classification ability of the other classes in an acceptable level.

Acknowledgments

This study was supported by grants from Oskar Öflund Foundation and Tampere Graduate School in Information Science and Engineering (TISE), Finland. Thanks go to Martti Juhola for commenting the manuscript and to M. Zwitter and M. Soklic for providing the Ljubljana breast cancer data.

References

1. Blake, C.L., Merz, C.J.: UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA, University of California, Department of Information and Computer Science (1998)
2. 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)
3. Kentala, E.: Characteristics of Six Otologic Diseases Involving Vertigo. Am. J. Otol. **17** (1996) 883-892
4. Laurikkala, J., Juhola, M.: A Genetics-Based Machine Learning System to Discover the Diagnostic Rules for Female Urinary Incontinence. Comput. Methods Programs Biomed. **55** (1998) 217-228
5. Kentala, E., Laurikkala, J., Pyykkö, I., Juhola, M.: Discovering Diagnostic Rules from a Neurotologic Database with Genetic Algorithms. Ann. Otol. Rhinol. Laryngol. **108** (1999) 948-954
6. Kubat, M., Holte, R.C., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Mach. Learn. **30** (1998) 195-215
7. 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, CA (1997) 179-186
8. Swingler, K.: Applying Neural Networks: A Practical Guide. Academic Press, London, (1996)
9. Cochran, W.G.: Sampling Techniques. 3rd edn. Wiley, New York (1977)
10. Aha, D.W., Kibler, D., Albert, M.K.: Instance-Based Learning Algorithms. Mach. Learn. **6** (1991) 37-66
11. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
12. Wilson, D.R., Martinez, T.R.: Reduction Techniques for Instance-Based Learning Algorithms. Mach. Learn. **38** (2000) 257-286
13. Weisberg, S.: Applied Linear Regression. 2nd edn. Wiley, New York (1985)

14. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in Databases, AI Magazine **17** (1996) 37-54
15. Dasarathy, B.V., Sanchez, J.S., Townsend, S.: Nearest Neighbour Editing and Condensing Tools-Synergy Exploitation. Pattern Analysis & Applications **3** (2000) 19-30
16. Wilson, D.R., Martinez, T.R.: Improved Heterogeneous Distance Functions. J. Artif. Intell. Res. **6** (1997) 1-34.
17. Laurikkala, J., Kentala, E., Juhola, M., Pyykkö, I., Lammi S.: Usefulness of Imputation for the Analysis of Incomplete Otoneurologic Data. Int. J. Med. Inf. **58-59** (2000) 235-242
18. Quinlan, J.R.: C4.5 Programs for Machine Learning. Morgan Kaufman, San Mateo, CA (1993)
19. Pett, M.A.: Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions. SAGE Publications, Thousand Oaks, CA (1997)