Class Noise and Supervised Learning in Medical Domains: The Effect of Feature Extraction

Mykola Pechenizkiy
Dept. of Math. IT
Univ. of Jyväskylä
Finland
mpechen@cs.jyu.fi

Alexey Tsymbal
Dept. CS, Trinity College Dublin
Ireland
tsymbalo@cs.tcd.ie

Seppo Puuronen
Dept. CS & ISs
Univ. of Jyväskylä
Finland
sepi@cs.jyu.fi

Oleksandr Pechenizkiy
Dept. CS & ISs
Univ. of Jyväskylä
Finland
olkpeche@cc.jyu.fi

Abstract

Inductive learning systems have been successfully applied in a number of medical domains. It is generally accepted that the highest accuracy results that an inductive learning system can achieve depend on the quality of data and on the appropriate selection of a learning algorithm for the data.

In this paper we analyze the effect of class noise on supervised learning in medical domains. We review the related work on learning from noisy data and propose to use feature extraction as a pre-processing step to diminish the effect of class noise on the learning process. Our experiments with 8 medical datasets show that feature extraction indeed helps to deal with class noise. It clearly results in higher classification accuracy of learnt models without the separate explicit elimination of noisy instances.

1. Introduction

Current electronic data repositories, especially in medical domains, contain huge amount of data including also currently unknown and potentially interesting patterns and relations that can be found using knowledge discovery and data mining (DM) methods [2]. Inductive learning algorithms can be used to form a generalization from a set of labeled (previously classified) instances so that the predictions can be performed for the previously unobserved instances. Inductive learning systems have been successfully applied in a number of medical domains, for example in localization of a primary tumor, prognostics of recurrence of breast cancer, diagnosis of thyroid diseases, and rheumatology [9].

However, it is generally accepted that the highest accuracy results that an inductive learning system can achieve depend on the quality of data and on the appropriate selection of a learning algorithm for the data. The quality of a dataset usually depends on a number of factors. In this paper we will emphasize the problem of class noise present in training data. Some DM researchers believe that more attention should be paid to the study of attribute noise in datasets as they think that at the time of insertion of new data (for example patient's data) a medical worker pays more attention to the diagnosis and to the correctness of its insertion. However, this is true only from the organizational perspective of data-entry process, and may be very different from other perspectives that we will address in this paper.

The goal of our study is three-fold: (1) to show the impact of class noise on supervised learning (SL) in medical domains, (2) to review the recent work on handling class noise, (3) and to investigate whether feature extraction (FE) can help to diminish the effect of class noise on SL.

FE is the process that discovers a new feature space having fewer dimensions through a functional mapping, keeping as much useful information about the data as possible [10]. Our hypothesis in this paper is that FE will help to produce more compact models during the SL process thus helping to avoid overfitting the noise included in the class attribute.

The rest of the paper is organized as follows. In Section 2 we consider class noise and its sources in medical domains. In Section 3 we review the related work on handling class noise for SL. In Section 4 FE techniques used in our experimental study are presented. In Section 5 we present the results of our experiments with 8 medical datasets. Finally, in Section 6 we briefly conclude with a summary and present some directions for further research.

2. Noise in data

2.1. Basic concepts

Data may contain various types of errors, either random or systematic. Random errors are often referred to as noise. However some authors are regarding as noise both mislabeled examples and outliers which are correctly classified but are relatively rare instances (also called exceptions).

According to [23], the quality of a dataset in SL is usually characterized by two information parts of instances, namely attributes and class labels. The quality of the attributes indicates how well they characterize instances for classification purposes, and the quality of class labels indicates the correctness of class labels' assignments. Noise is often similarly divided into two major categories that are class noise (misclassifications or mislabeling) and attribute noise (errors introduced to attribute values). Zhu and Wu [23] propose to distinguish the following examples of attribute errors: erroneous attribute values, missing or so-called 'don't know' values, and incomplete or so-called 'don't care' values.

The two major types of class noise are contradictory instances (instances with the same values of the attributes but different class labels, forming so-called irreducible or Bayes error) and wrongly classified (labeled) instances that are misclassifications (mislabelings). These errors may naturally occur in medical diagnostics when different classes have very similar or even overlapping symptoms.

In this paper our focus is on the study of the second type of class noise, namely the study of the performance of SL under the presence of certain amount of mislabeled instances in the data. We prefer the term 'mislabeled' instead of 'misclassified' because it is more general and can refer to several sources of class noise which we discuss in the following section.

2.2. Sources of class noise

It can be easily demonstrated that there is more than one reason why mislabeling is possible and that class noise is often present in real-world data. In particular, beside data-entry errors, subjectivity and the inadequacy of information used to label each instance constitute the major factors that may have impact on the amount of mislabeled instances in a dataset [1].

Domains in which medical experts may disagree are natural ones for subjective labeling errors [18]. In particular, if in some practical classification problem the absolute ground truth is unknown then experts must subjectively provide labels and mislabeled instances naturally appear [13]. Other types of instance mislabeling refer to the situations when an observation needs to be ranked according to a disease severity or when the information used to label an instance is different from the information to which the learning algorithm will have access (for example, when an expert relies on visual input rather than the numeric values of the attributes). Also, in medical domains the results of some tests (attribute values) may often be unknown (impossible to obtain or difficult to obtain, for example because of cost or time considerations), and such incompleteness of information may lead to class noise as well. Another source of misclassifications in data are the errors of devices used for automatic classification.

Two types of studies that aim either (1) at improving the quality of training data by identifying and eliminating mislabeled instances prior to applying a certain SL technique (and thereby decreasing the classification error) or (2) at developing such SL techniques that would be tolerant to class noise have been undertaken in the DM community [23].

3. Handling class noise

Noise handling techniques can be roughly divided into two categories: (1) so-called noise-tolerant techniques that try to avoid overfitting the possibly noisy training set during SL, and (2) filtering techniques that detect and eliminate noisy instances before SL [1, 4, 5, 6].

The noise-tolerant techniques handle noise implicitly, and the noise-handling mechanism is often embedded into either (1) search heuristics and stopping criteria used in model construction [11], or (2) post-processing such as decision tree post-pruning [15], or (3) model selection mechanism based for example on minimum description length (MDL) principle [17] or some compression measure which integrates measure of model complexity with an accuracy estimate [12].

The filtering techniques handle noise explicitly, and the noise-handling mechanism is often implemented as a filter that is applied before SL and it usually results in a reduced training set (when the noisy instances are not corrected but deleted).

Filtering algorithms can be broadly divided into single-algorithm filters and ensemble filters [1]. With single-algorithm filters one approach is to use the same learning algorithm to construct both the filter and the final classifier. This idea adapts the approach to remove outliers in regression analysis, where the same model is used to test for outliers and for fitting the final model to the cleaned data [21]. John [7] experimented with removing the training instances that are pruned by

C4.5 [16]. The tree was iteratively rebuilt from the reduced (filtered) set of training instances until no further pruning could be done. Brodley [1] used cross-validation over the training data instead of multiple iterations to find mislabeled instances. Another way to implement filtering is to construct a filter using one algorithm and to construct the final classifier using a different algorithm. This approach is based on the assumption that an algorithm may act as a better filter for another algorithm [1].

Ensemble filters detect mislabeled instances by constructing a set of base-level detectors (classifiers) and then use their classification errors to identify mislabeled instances [1]. Brodley [1] analysed the use of a majority vote filter (tags an instance as mislabeled if more than half of all base classifiers misclassify it), and a consensus filter (requires that all base classifiers must fail to classify an instance before tagging it as a misclassified one) for different benchmark datasets.

Gamberger et al. [4, 5] presented a series of noise detection experiments with coronary artery disease diagnosis (in [5]) and with 8 medical domains from the UCI repository (in [4]). They used a combined classification-saturation filter, and a consensus saturation filter. A saturation filter is based on the saturation property of the training set which assumes that the training set contains enough instances to learn a correct model. Earlier, it was shown how a simple compression measure can help to eliminate noise in a medical problem of rheumatic diseases diagnosis [6].

An interested reader can find further information about the behavior of various filtering approaches in [1,4]. We will just summarize that several researchers have acknowledged filtering approaches to be useful. They can help in noise elimination that results in the higher classification accuracy of a classifier built on a filtered dataset. Therefore, many researchers argue that the explicit elimination of noise during the preprocessing step is favorable since noisy examples in this case do not impact the SL process. Such argumentation seems to be rather obvious at least in theory. However, the same researchers have recognized some practical difficulties with filtering approaches.

One concern is that it is often hard to distinguish noise from exceptions (outliers) without the help of an expert, especially if the noise is systematic [19]. Another concern is that a filtering technique can use an expected level of noise as an input parameter, and this value is rarely known for a particular dataset. Only in some cases domain knowledge may help to estimate the level of noise in data. Besides, since the scarcity of training data is not an unusual problem in medical diagnosis, it is desirable to minimize the probability of discarding an instance that is an exception rather than an error.

In this paper we propose to use FE techniques to eliminate the effect of class noise on SL. This approach fits better to the second category of noise-tolerant techniques as it helps to avoid overfitting implicitly within learning techniques. However, this approach has also some similarity with the filtering approach as it clearly has a separate phase of dimensionality reduction which is undertaken before the SL process.

4. Feature Extraction Techniques Used

Principal Component Analysis (PCA) is one of the most commonly used FE techniques. It is based on extracting the axes on which data shows the highest variability [8]. Although PCA "spreads out" the data in the new basis (new extracted axes), and can be of great help in unsupervised learning, there is no guarantee that the new axes are consistent with the discriminatory features in a classification problem.

Another approach is to account for class information during the FE process. One technique is to use some class separability criterion (for example, from Fisher's linear discriminant analysis), based on a family of functions of scatter matrices: the within-class covariance, the between-class covariance, and the total covariance matrices [3]. The parametric and nonparametric eigenvector-based approaches use the simultaneous diagonalization algorithm to optimize the relation between the within- and between-class covariance matrices (thus taking into account class information) [3]. The difference between the approaches is in calculation of the between-class covariance matrix. The parametric approach accounts for one mean per class and one total mean, and therefore may extract at most *number_of_classes-1* features. The nonparametric method tries to increase the number of degrees of freedom in the between-class covariance matrix, measuring the between-class covariances on a local basis. Previous experiments with parametric and nonparametric FE approaches showed that nonparametric FE was often more robust to different dataset characteristics [20] and often resulted in higher classification accuracy of such basic SL techniques as Naïve Bayes, C4.5 and *k*NN comparing to parametric FE [14].

5. Datasets and experiment design

To evaluate the impact of class noise, we conducted experiments on 8 medical datasets (Table 1), with different levels of random class noise imputed (from 0

to 20% with a 2% step). Further information on these datasets and the datasets themselves are available at http://www.informatics.bangor.ac.uk/~kuncheva/activities/real_data.htm.

Table 1. Medical datasets used in the study

dataset	instances	features	classes
contractions	98	27	2
laryngeal1	213	16	2
laryngeal2	692	16	2
laryngeal3	353	16	3
rds	85	17	2
weaning	302	17	2
voice3	238	10	3
voice9	428	10	9

We applied *k*NN, Naïve Bayes (NB) and C4.5 decision tree learning algorithms (with and without FE techniques discussed in the previous section) to learn from these noisy datasets and to evaluate the impact of class noise on accuracy.

For each data set with each level of imputed class noise 30 test runs of Monte-Carlo cross validation were made for each classification technique with and without FE approaches. In each run, the data set is first split into the training set and the test set by stratified random sampling to keep class distributions approximately the same. Each time 30 percent instances of the dataset are first randomly held out to the test set. The remaining 70 percent instances form the training set, which is then corrupted with imputed random class noise. The test environment was implemented within the WEKA machine learning software in Java [22].

5.1. Impact of class noise on SL

The corresponding accuracy results are presented in Figure 1. The horizontal axis indicates the class noise level and the vertical axis represents the

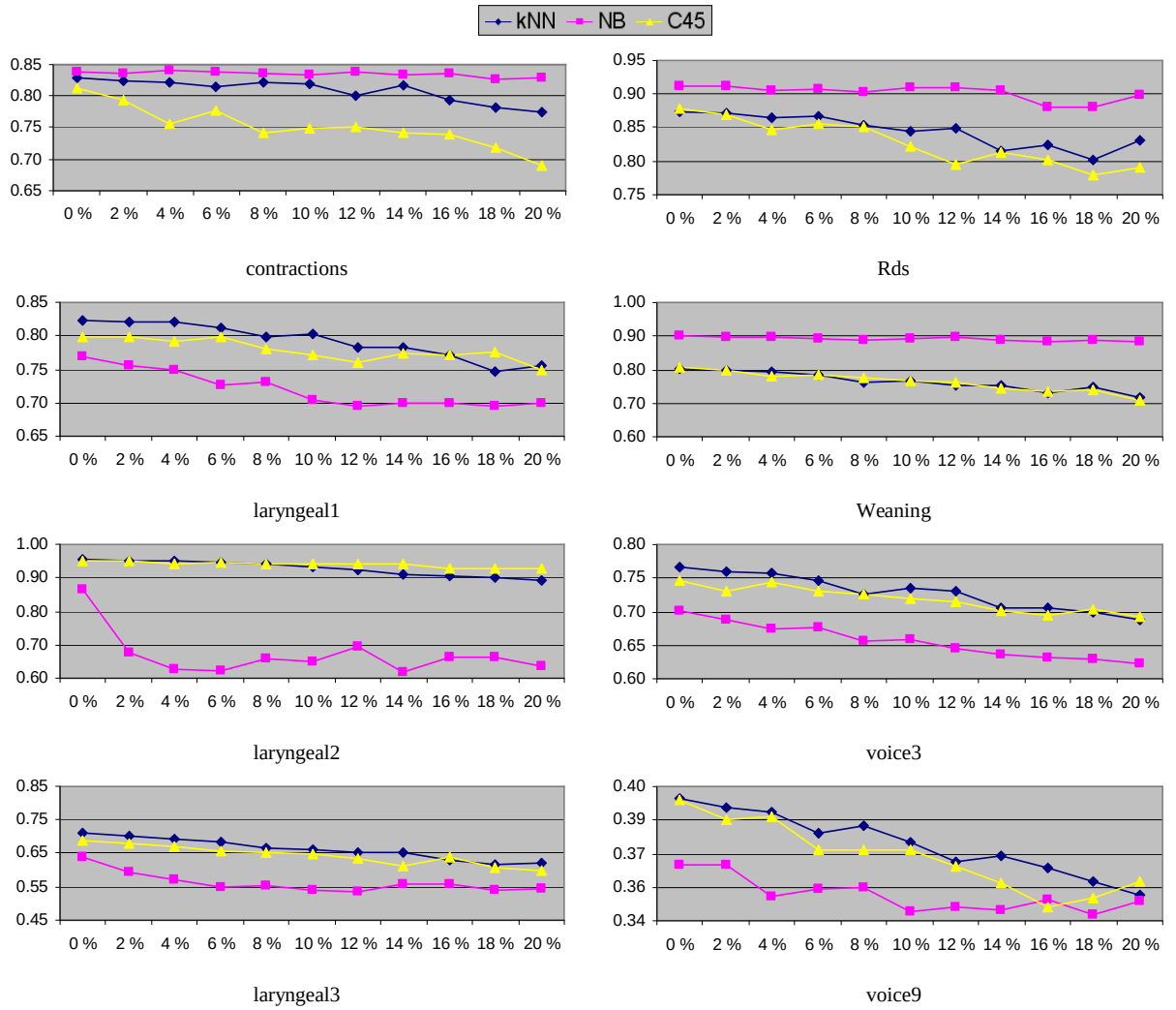


Figure 1. Accuracies for 8 datasets for *k*NN, NB and C4.5 classifiers with imputed class noise in training data

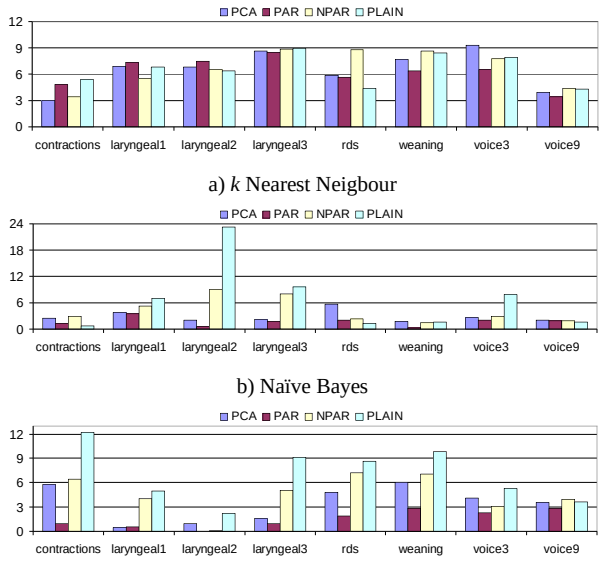
classification accuracy of different classifiers trained on a noisy training set and evaluated on a original test set (that is without imputing any class noise).

As we can see from Figure 1, when the level of noise increases, all classifiers trained on a noisy training set suffer from decreasing classification accuracy. In most cases the accuracy decrease is linear with respect to the increase of noise level. The behavior of different classifiers varies from one dataset to another. In particular, for *contractions*, *rds* and *weaning* datasets NB has higher accuracies for uncorrupted data and its accuracy deteriorates less in comparison with *k*NN and C4.5. On the contrary, on *laryngeal* and *voice* datasets *k*NN and C4.5 show better results, being more tolerant to class noise than NB.

5.2. Feature extraction and class noise

Further we compare how classification error increases for each classifier for situations when no FE was applied before SL (*PLAIN*) and when *PCA*, parametric FE (*PAR*) and nonparametric FE (*NPAR*) were applied before learning a classifier.

In Figure 2 we present the results of this comparison for each of the 8 datasets. The vertical axis shows the classification error increase due to the inclusion of 20% class noise in a dataset.



a) *k* Nearest Neigbour

b) Naïve Bayes

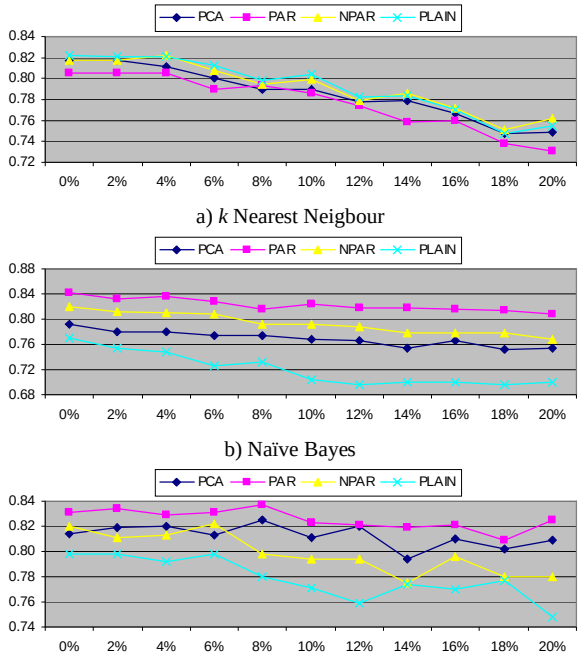
c) C4.5 decision tree

Figure 2. Error increase (%) due to class noise

In can be seen from the figure that with *k*NN the increase or error is 1-2% smaller (in comparison with *PLAIN*) when *PCA* is applied for *contractions*, when NPAR is applied for *contractions* and *laryngeal1*, and when PAR is applied for *wearing*, *voice3* and *voice9*. On *rds*, the increase in error is larger with any FE technique in comparison with *PLAIN*. In general, we can see that FE helps *k*NN to tolerate the presence of class noise rather modestly in comparison with NB and C4.5, where maximum difference in error increase is beyond 20% and 10% percent respectively. In particular, it can be seen that with NB the increase of error is smaller (in comparison with *PLAIN*) by about 3% for *laryngeal1*, 20% for *laryngeal2*, 8% for *laryngeal3*, 6% for *voice3* when PAR is applied. However with the other datasets (*contractions*, *rds*, *wearing* and *voice9*) FE is not helpful with respect to noise handling. And with C4.5 and NB the increase of error is smaller by about 10% for *contractions*, 4% for *laryngeal1*, 2% for *laryngeal2*, 8% for *laryngeal3*, 6% for *rds*, 6% for *wearing*, and 3% for *voice3* when PAR is applied. Only with *voice9* FE is not helpful with respect to noise handling.

Then, for each classifier, we compare the absolute values of classification accuracy and its decrease due to class noise.

In Figure 3 we demonstrate one set of representative results of this comparison (for *laryngeal1*).



a) *k* Nearest Neigbour

b) Naïve Bayes

c) C4.5 decision tree

Figure 3. Typical behavior of SL with and without FE when class noise is imputed (*laryngeal1)*

It can be seen from the figure that with *k*NN FE approaches have almost no effect on diminishing the

impact of class noise on classification; with NB the decrease of accuracy is less rapid (in comparison with *PLAIN*) when FE is applied (*PAR* shows the best result, then *NPAR* and then *PCA*); and with C4.5 the behavior is similar to NB, but *PCA* appears to be a better choice than *NPAR*.

6. Conclusions and future directions

The focus of this paper is on evaluating the impact of class noise on SL in medical domains and showing whether FE can diminish this impact. We demonstrated that, expectedly, class noise affects SL in many cases. The results of our experiments show that applying FE techniques before undertaking SL indeed enables decreasing the negative effect of the presence of mislabeled instances in the data. This is especially notable with NB and C4.5, which are less tolerant to the presence of class noise in data in comparison with *k*NN.

The directions of our future work include the comparison of FE techniques with other dimensionality reduction and instance selection techniques, and comparison of FE with filter approaches for class noise elimination.

Acknowledgments: This research is partially supported by the Academy of Finland and the Centre for International Mobility (CIMO), Finland and the Science Foundation Ireland under Grant No. S.F.I.-02IN.1I111. We are thankful to anonymous reviewers for useful suggestions and to Dr. Ludmila I. Kuncheva for the collection of real medical datasets, which is available at: http://www.informatics.bangor.ac.uk/~kuncheva/activities/real_data.htm.

7. References

[1] C.E. Brodley and M.A. Friedl, "Identifying Mislabeled Training Data", *Journal of Artificial Intelligence Research* 11, 1999, pp. 131-167.

[2] Fayyad, U., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, *Advances in Knowledge Discovery and Data Mining*, AAAI/ MIT Press, 1997.

[3] Fukunaga, K. *Introduction to statistical pattern recognition*. Academic Press, London, 1999.

[4] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise Detection and Elimination in Data Preprocessing: experiments in medical domains", *Applied Artificial Intelligence* 14, 2000, pp. 205–223.

[5] D. Gamberger, N. Lavrac, and C. Groselj, "Experiments with Noise Filtering in a Medical Domain", *Proc. of 16th ICML Conference*, San Francisco, CA, 1999, pp. 143–151.

[6] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise elimination in inductive concept learning: A case study in medical diagnosis", In *Proceedings of the 7th International Workshop on Algorithmic Learning Theory*, Berlin, Springer, 1996, pp. 199-212.

[7] G.H. John, "Robust decision trees: Removing outliers from data", In *Proc. of 1st Int. Conf. on Knowledge Discovery and Data Mining*, Menlo Park: AAAI Press 1995, pp. 174-179.

[8] Jollife, I.T. *Principal Component Analysis*, *Springer-Verlag*, New York. 1986.

[9] I. Kononenko, "Inductive and Bayesian learning in medical diagnosis", *Applied Artificial Intelligence* 7(4), 1993, pp. 317-337.

[10] Liu, H. *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Kluwer, 1998.

[11] J. Mingers, "An empirical comparison of selection measures for decision-tree induction", *Machine Learning* 3(4), 1989, pp. 319- 342.

[12] S. H. Muggleton, A. Srinivasan, and M. Bain, "Compression, significance and accuracy", In *Proc. of the 9th Int. Conf. on Machine Learning*, San Mateo, CA: Morgan Kaufmann, 1992, pp. 338 - 347.

[13] F.M. Nouri and N.B. Lincoln, "Predicting driving performance after stroke", *British Medical Journal* 307, 1993, pp. 482-483.

[14] M. Pechenizkiy, "Impact of the Feature Extraction on the Performance of a Classifier: kNN, Naïve Bayes and C4.5", In *Proc. of 18th CSCSI Conference on Artificial Intelligence AI'05*, LNAI 3501, Springer Verlag, 2005, pp. 268-279.

[15] J.R. Quinlan, "The Effect of Noise on Concept Learning", In Michalski, R.S., Carboneel, J.G. and Mitchell, T.M. (eds.), *Machine Learning*, Morgan Kaufmann, 1986.

[16] Quinlan J.R., *C4.5: Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann, 1993.

[17] J. Rissanen, "Modeling by shortest data description", *Automatica* 14, 1978, pp.465-471.

[18] P. Smyth, "Bounds on the mean classification error rate of multiple experts", *Pattern Recognition Letters* 17, 1996, pp. 1253-1327.

[19] A. Srinivasan, , S. Muggleton, and M. Bain, "Distinguishing exceptions from noise in non-monotonic Learning", In *Proc 2nd Int. Workshop on Inductive Logic Programming*, Tokyo, Japan, 1992.

[20] A. Tsymbal, S. Puuronen, M. Pechenizkiy, M. Baumgarten, and D. Patterson, "Eigenvector-based feature extraction for classification", In *Proc. 15th Int. FLAIRS Conf. on Artificial Intelligence*, AAAI Press, 2002, pp. 354-358.

[21] Weisberg, S. *Applied Linear Regression* (2nd ed.), New York: Wiley, 1985.

[22] Witten, I. and E. Frank, *Data Mining: Practical Machine Learning Tools with Java Implementations*, Morgan Kaufmann, San Francisco, 2000.

[23] X. Zhu, and X. Wu, "Class noise vs. attribute noise: a quantitative study of their impacts", *Artificial Intelligence Review* 22 (3-4), 2004, pp. 177-210.