Improving Classification by Removing or Relabeling Mislabeled Instances

Stéphane Lallich, Fabrice Muhlenbach, and Djamel A. Zighed

ERIC Laboratory – University of Lyon 2
5, av. Pierre Mendès-France
F-69676 BRON Cedex – FRANCE
{lallich, fabrice.muhlenbach, zighed}@univ-lyon2.fr

Abstract. It is common that a database contains noisy data. An important source of noise consists in mislabeled training instances. We present a new approach that deals with improving classification accuracies in such a case by using a preliminary filtering procedure. An example is suspect when in its neighborhood defined by a geometrical graph the proportion of examples of the same class is not significantly greater than in the whole database. Such suspect examples in the training data can be removed or relabeled. The filtered training set is then provided as input to learning algorithm. Our experiments on ten benchmarks of UCI Machine Learning Repository using 1-NN as the final algorithm show that removing give better results than relabeling. Removing allows maintaining the generalization error rate when we introduce from 0 to 20% of noise on the class, especially when classes are well separable.

1 Introduction – Outliers Issue

In this paper, we address the learning process of a categorical variable Y, on the basis of an example database described by p numerical attributes, denoted by X_1, X_2, ... X_p. Our focus is on mislabeled examples that constitute a specific category of outliers. We suggest a filtering strategy which identifies suspect examples in order to improve the generalization performance of the learning algorithm. We consider two options: removal or relabeling. This strategy is based on the *cut weighted edges statistic* [16] defined by geometrical neighborhood [17] and associated with 1-NN prediction.

Identifying outliers is an important step in any instance-based knowledge discovery process [2]. By outliers, we mean examples whose exceptional nature disturbs generalization. Barnett and Lewis [1] define an outlier as "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data".

Outliers can have different origins, and we will now handle the most common. We will first mention inclusion errors, that happen when an example is wrongly included in the learning set. Either the reference population was wrongly defined or the error happened at the time of sampling, as can be seen in medicine or marketing. Then, we consider appreciation, coding or data-input errors, be it on

M.-S. Hacid et al. (Eds.): ISMIS 2002, LNAI 2366, pp. 5–15, 2002.

the predictors or on the class. One should also talk about observations regarding rare examples, which are often associated with variable asymmetry and are an important cause for leverage points. Lastly, the model error can be strongly affected if a relevant attribute has been forgotten and this may wrongly show some examples as outliers.

Outliers disturb the learning process, mainly when the latter includes the variable's statistical moments; the mean and even more so the variance are usually very sensitive to exceptional values. Estimations and especially confidence intervals and the associated p-values may then be distorted. This can lead to faulty conclusions, particularly in a regression.

We distinguish between works which aim at identifying outliers for themselves (e.g., in the cases of fraud detection or records [7]) and those which look into limiting their noisy effects [6,4,5]). Our study belongs to this latter category. In order to remedy to the outliers issue in this second case, we can either remove the suspect examples from the learning set [6,5], or relabel them (cf. relaxation models [8]).

In the learning process, we often prefer to talk about noise, distinguishing the noise on the class from the noise on the attributes [10]. In this last case, Quinlan has shown that when the noise level increases, the fact of removing the noise from the attributes decreases the generalization performance of the classifier if the data to be classified presents the same attribute noise. As regards the class noise, the problem can be seen in different terms, because noise only concerns the learning set. Thus, Brodley and Friedl [4,5] have demonstrated that whatever the base or the filtering strategy experimented, identifying and removing those examples improves substantially the predictive accuracy in generalization as long as the level of noise does not exceed 20%, or even 30 to 40% in some cases.

2 Works Related to Mislabeled Examples

We will now get into the details of the works that deal with class noise. Because our aim is not to reduce the size of the learning set but to filter it, we will mainly discuss the work of Wilson [14], John [6] as well as Brodley and Friedl [4,5].

Wilson [14] has suggested the E-k-NN rule (Edited k Nearest Neighbor rule) which consists in using the k-NN classifier ($k = 3$) to filter the whole of the learning set before proceeding to the prediction by 1-NN. Only instances that the k-NN classifies properly are retained for the 1-NN. This rule edits out mislabeled instances, as well as the ones which are close border cases, leaving smoother decision boundaries while retaining all internal instances. Therefore this algorithm clearly is a filtering one rather than a reducing one. One may apply the E-k-NN rule repeatedly, until all remaining instances have a majority of same class neighbors. An alternative use was introduced by Tomek [12], who applies repeatedly the E-k-NN rule for growing values of k. This process was included by Wilson and Martinez [15] in different techniques of instances selection (DROP 3 for filtering) for instance-based learning algorithms.