Saturation filtering for noise and outlier detection

Nada Lavrač[1] and Dragan Gamberger[2]

[1] Institute Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia
nada.lavrac@ijs.si
[2] Rudjer Bošković Institute Bijenička 54, 10000 Zagreb, Croatia
Dragan.Gamberger@irb.hr

Instance selection is an important part of the KDD process. It is aimed at finding a representative data subset that can replace the original dataset, still solving a data mining task as if the whole dataset were used. For most data mining tasks, including classification tasks, the selected dataset should preferably exclude noisy instances, aimed at improving the predictive accuracy and applicability of induced knowledge significantly depend on appropriate noise handling procedures.

In order to present our approach to noise handling, we first introduce the underlying theory and intuitions. A *target concept* of a given problem domain is pre-defined as the source of all possible correct examples of the concept. An inductive learning task is to find a good representation of the target concept in a selected hypothesis language. This representation is called a *target theory*. The main property of a target theory is that it should be correct for all the correct domain examples. A domain may consist of instances of a single concept, a set of subconcepts, a main concept and some subconcepts, etc. These domain characteristics together with the restrictions of the used hypothesis language may cause that a target theory may either have the form of a single concept description, a set of subconcept descriptions, or a main concept description and its exceptions (descriptions of small subconcepts or outlier descriptions).

In an ideal inductive learning problem, the induced hypothesis is indeed a target theory that will 'agree' with the classifications of training examples and will thus perform as a perfect classifier of unseen instances. In practice, however, it frequently happens that available data contain various kinds of errors, either random or systematic. *Noise* usually means random errors in training examples (erroneous attribute values and/or erroneous classification). In this work, however, we use the term noise in a broader sense of outliers [20] denoting all examples that do not follow the same model as the rest of the data. This definition "includes not only erroneous data but also *surprising* veridical data" [6], including non-typical class representatives. Thus, in this work the term noise denotes errors (incorrect examples) as well as exceptions (correct examples representing some relatively rare subconcept of the target concept). Namely, from the induction point of view, exceptions have the same effect on the induction process as erroneous examples themselves.

Detection and filtering of noisy examples from the training set helps in the induction of the target hypothesis: a hypothesis induced from noiseless data will be less complex and more accurate when classifying unseen cases. For small datasets, noise filtering can be used as an important step in the preprocessing of training data; for large datasets, on the other hand, the sampling process that selects a subset of the training set should include or should be followed by noise filtering. Notice, however, that if noise filtering is combined with random sampling, the following phenomenon has to be taken into account. In domains with uneven concept (or class) distributions random sampling may result in only few instances of some subconcepts to be present in the sample [12]. Such instances, despite the fact that they are important representatives of a rare subconcept, are outliers in the sample. This observation is similar to the side-effect noticed in instance selection approaches based on critical and boundary points, like instance-based learning [1] and support vector machines [4] which tend to increase the relative number of noisy instances in the selected subset.

The above observations support the development of cautious procedures for explicit detection and filtering of noisy instances. The approach in this work [8] follows the Occam's razor principle, commonly attributed to William of Occam (early 14th century), that states: "Entities should not be multiplied beyond necessity." This principle is generally interpreted as: "Among the theories that are consistent with the observed phenomena, one should select the simplest theory" [14]. The Occam's razor principle suggests that among all the hypotheses that are correct for all (or for most of) the training examples one should select the simplest hypothesis; it can be expected that this hypothesis is most likely to capture the structure inherent in the problem and that its high prediction accuracy can be expected on objects outside the training set [17]. This principle is frequently used by noise handling algorithms (e.g., rule truncation and tree pruning) as noise handling aims at simplifying the generated rules or decision trees in order to avoid overfitting a noisy training set.

Despite the successful use of the Occam's razor principle as the basis for hypothesis construction, several problems arise in practice. First is the problem of the definition of the most appropriate complexity measure that will be used to identify the simplest hypothesis, since different measures can select different simplest hypotheses for the same training set. This holds for any Kolmogorov complexity based measure [14], including MDL (Minimal Description Length) [17], that use approximations of an ideal measure of complexity. Second, recent experimental work has undoubtly shown that applications of the Occam's razor may not always lead to best prediction accuracy. A systematic way of improving the results generated by the application of the Occam's razor principle has been presented by Webb [19]. Further empirical evidence against the use of Occam's razor principle is provided by boosting and bagging approaches [15] in which an ensemble of classifiers (a more complex hypothesis) typically achieves better accuracy than any single classifier. In addition to experimental evidence, much disorientation was caused by the so-called "conservation law of generalization performance" [18]. Although it is rather clear [16] that real-word learning tasks are different from the set of all theoretically possible learning tasks as defined in [18], there remains the so-called "selective superiority problem" that each algorithm performs best in some but not all domains [2]; the latter can be in a way explained by the bias-variance tradeoff [13]. Some of the above contradictory observations have been recently explained by Domingos [5].

Our work [8] contributes to the understanding of the abovementioned phenomena by studying for which domains (real or theoretically possible) an inductive learning algorithm based on the Occam's razor principle has a theoretical chance for successful induction. The paper [8] theoretically elaborates the conditions for Occam's razor applicability. For practical applications, the paper indicates the properties that need to be satisfied for the effective induction using the Occam's razor principle. This theoretical work is the basis for the *saturation filter* used for explicit noise and outlier detection presented in this work. The approach to explicit detection and filtering of noise and outliers is based on the observation that the elimination of noisy examples, in contrast to the elimination of examples for which the target theory is correct, reduces the CLCH value of the training set (CLCH stands for the Complexity of the Least Complex correct Hypothesis). This noise detection algorithm is called the saturation filter since it employs the CLCH measure to test whether the training set is saturated, i.e., whether, given a selected hypothesis language, the dataset contains a sufficient number of examples to induce a stable and reliable target theory.

An alternative approach to explicit noise detection and elimination has been suggested by Brodley and Friedl [3]. The basic idea of their noise elimination algorithm, called in this work the *classification filter*, is to use one or more learning algorithms (that may but do not have to include explicit noise handling) to create classifiers that serve as filters for the training data. The experimental results published in [3] provide evidence that classification filters can successfully deal with noise. It has also been shown that consensus filters are conservative at eliminating good data while majority filters are better at detecting bad data that makes them more appropriate for situations with an abundance of data.

This work discusses the performance of four filtering algorithms, all aimed at the detection and elimination of noisy instances: our noise filtering algorithm called the saturation filter [8, 10], the noise filtering algorithm by Brodley and Friedl called the classification filter [3], and two new algorithms that combine the two approaches to noise filtering. Specifically, the work presents a series of noise detection experiments performed in the problem of diagnosis of coronary artery disease [11], using the rule learning system ILLM [7] without noise handling, using either the saturation filter or the classification filter with ILLM, and by upgrading ILLM with two novel combined approaches: the combined classification-saturation filter and the consensus saturation filter. The best results were achieved by the latter, employing various levels of consensus achieved by different ways of n-fold saturation filtering, aimed at increasing the reliably of the detection of noisy instances.

References

1. Aha, D. W., Kibler, D., and Albert, M. K. (1991). Instance-based learning algorithms. *Machine Learning*, 6:37–66.

2. Brodley, M. (1995). Recursive automatic bias selection for classifier construction. *Machine Learning*, 20:63–94.
3. Brodley, C.E. and Friedl, M.A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167.
4. Burges, C. (1998). A tutorial on support vector machines. *Journal of data Mining and Knowledge Discovery*, 2(2).
5. Domingos, P. (1999). The role of Occam's razor in knowledge discovery. *Data Mining and Knowledge Discovery*, 3(4):409–425.
6. John, G.H. (1995). Robust decision trees: Removing outliers from data. In *Proc. of the First International Conference on Knowledge Discovery and Data Mining*, pp. 174–179. AAI Press.
7. Gamberger, D. (1995). A minimization approach to propositional inductive learning. In *Proc. of the Eighth European Conference on Machine Learning*, pp. 151–160. Springer-Verlag.
8. Gamberger, D. and Lavrač, N. (1997). Conditions for Occam's razor applicability and noise elimination. In *Proc. of the Ninth European Conference on Machine Learning*, pp. 108–123. Springer-Verlag.
9. Gamberger, D., Lavrač, N., and Grošelj, C. (1999). Experiments with noise filtering in a medical domain. In *Proc. of the Sixteenth International Conference of Machine Learning*, pp. 143–151. Morgan Kaufmann.
10. Gamberger, D., Lavrač, N., and Džeroski, S. (2000). Noise detection and elimination in data preprocessing: experiments in medical domains. *Applied Artificial Intelligence*, 14:205–223.
11. Grošelj, C., Kukar, M., Fetich, J., and Kononenko, I. (1997). Machine learning improves the accuracy of coronary artery disease diagnostic methods. *Computers in Cardiology*, 24:57–60.
12. John, G.H. and Langley, P.(1996). Static versus dynamic sampling for data mining. In *Proc. of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 367–370. AAI Press.
13. Kohavi, R. and Wolpert, D.H. (1996). Bias plus variance decomposition for zero-one loss functions. In *Proc. of the Thirteenth International Conference on Machine Learning*, pp. 275–283. Morgan Kaufmann.
14. Li, M. and Vitányi, P. (1993). An Introduction to Kolmogorov Complexity and its Applications. Springer-Verlag.
15. Quinlan, J.R. (1996). Bagging, boosting and C4.5. In *Proc. of the Thirteenth National Conference on Artificial Intelligence*, pp. 725–730. AAAI Press.
16. Rao, R., Gordon, D., and Spears, W. (1995). For every generalization action, is there really an equal or opposite reaction? Analysis of conservation law. In *Proc. of the Twelveth International Conference on Machine Learning*, pp. 471–479. Morgan Kaufmann.
17. Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*, 14:465–471.
18. Schaffer, C. (1994). A conservation law for generalization performance. In *Proc. of the Eleventh International Conference on Machine Learning*, pp. 259–265. Morgan Kaufmann.
19. Webb, G.I. (1996). Further experimental evidence against the utility of Occam's razor. *Journal of Artificial Intelligence Research*, 4:397–417.
20. Weisberg, S. (1985). Applied Linear Regression. John Wiley & Sons.