Data Sample Reduction for Classification of Interval Information Using Neural Network Sensitivity Analysis

Piotr A. Kowalski[1] and Piotr Kulczycki[2]

[1] Systems Research Institute, Polish Academy of Sciences,
ul. Newelska 6, PL-01-447 Warsaw, Poland
pakowal@ibspan.waw.pl,
[2] Cracow University of Technology,
Department of Automatic Control and Information Technology
ul. Warszawska 24, PL-31-155 Cracow, Poland
kulczycki@pk.edu.pl

Abstract. The aim of this paper is present a novel method of data sample reduction for classification of interval information. Its concept is based on the sensitivity analysis, inspired by artificial neural networks, while the goal is to increase the number of proper classifications and primarily, calculation speed. The presented procedure was tested for the data samples representing classes obtained by random generator, real data from repository, with clustering also being used.

Keywords: classification, interval information, data sample reduction, artificial neural networks, sensitivity method.

1 Introduction and Main Results

Recently, interest in interval analysis has grown notably in many practical applications [2]. Fundamental here is the assumption that the only available information about the investigated quantity $x \in \mathbb{R}$, is that it fulfills the condition $\underline{x} \leq x \leq \overline{x}$, and consequently can be treated as the interval $\left[\underline{x}, \overline{x}\right]$. The multidimensional case $x \in \mathbb{R}^n$ was also examined. The main subject of the research presented here is a reduction data sample for the classification task. The tested element is of interval form, but elements consisting of patterns of particular classes are defined uniformly. A classification procedure [3] was worked out for the removal from samples of those elements having negligible or even negative influence on the correctness of classification. Its concept is based on the sensitivity method [1], inspired by neural networks, while the goal is to increase the number of proper classifications as well as, primarily, calculation speed. The concept of classification is based on the Bayes approach, ensuring a minimum of potential losses arising from misclassification. For a such-formulated problem, the methodology of statistical kernel estimators [4] is used, which frees the investigated procedure from arbitrary assumptions concerning shapes of samples.

2 Numerical Experiments

As an illustrative example, consider the one-dimensional case with two classes, represented by 50-elements samples given by Gaussian generators N(0,1) and N(2,1).

D. Dicheva and D. Dochev (Eds.): AIMSA 2010, LNAI 6304, pp. 271–272, 2010.

The interval-type elements subjected to classification were calculated by generating intervals' centers and then their lengths: 0.1; 0.25; 0.5; 1.0; 2.0. Using the concept of the classification method without reduction, 16.21; 16.38; 16.42; 16.43; 16.45 percent were misclassifications, with respect to interval length. Next, by applying the data reduction procedure, the number of misclassifications decreased to 14.76; 14.87; 14.89; 14.96; 15.00 percent, respectively, when sample size was also significantly reduced. The common occurrence of the results in both the precision and calculation speed aspects, is worth underlining: about 10% improvement of classification accuracy with around 40% reduction of sample sizes were simultaneously obtained. Obviously, following a reduction in sample size, calculation speed was also significantly reduced. In the case when the samples had been obtained by the clustering k-means method, the classification algorithm led to the following results: 16.60; 16.34; 16.32; 16.33; 16.33 percent of misclassifications before the reduction procedure, and 14.67; 14.55; 14.51; 14.51; 14.50 after it. The results obtained here were better than in the basic case described above, thanks to a more effective reduction in atypical elements of patterns incorrectly treated during the clustering procedure.

3 Comments and Final Remarks

The investigated method is two-phased in its nature – the time-consuming procedures for defining the classifier and reduction data take place only once at the beginning, however, the interval classification procedure itself is performed in a relatively short period, mainly thanks to analytical forms of formulas obtained. The developed reduction algorithm was compared with simple and natural random reduction as well as with the k-NN method. For all these cases for reduction, the concept worked out here produced much better results. Furthermore, there is no need for any arbitrary assumption concerning an algorithm's parameters, which is another positive aspect of the procedure presented in this paper.

References

1. Engelbrecht, A.P.: Sensitivity Analysis for Selective Learning by Feedforward Neural Networks. Fundamenta Informaticae, vol. 46(3), pp. 219--252 (2001)
2. Jaulin, L., Kieffer, M., Didrit, O., Walter, E.: Applied Interval Analysis. Springer, Berlin (2001)
3. Kowalski, P.A.: Bayesian Classification of Imprecise Interval-Type Information (in Polish). Systems Research Institute, Polish Academy of Sciences, Ph.D. Thesis (2009)
4. Kulczycki, P.: Kernel Estimators in Industrial Applications. In: Soft Computing Applications in Industry, B. Prasad (ed.), pp. 69--91, Springer-Verlag, Berlin, (2008)