Instance selection for regression by discretization

Álvar Arnaiz-González, José F. Diez-Pastor, César García-Osorio, and Juan J. Rodríguez

Universidad de Burgos, España
{alvarag,jfdpastor,cgosorio,jjrodriguez}@ubu.es

Resumen This is a summary of our article published in Experts Systems with Applications [1], presented to the Multi-Conference CAEPIA'16 KeyWorks.

Keywords: instance selection, regression, mutual information, noise filtering, class noise

1 Summary

Automatic supervised learning begins with a data set of instances or examples, each of which is composed of input-output pairs. The learning problem consists in determining the relation between the input and the output values. When the output is a nominal or discrete value, the task is one of classification, as opposed to regression in which the value to predict is a continuous, numerical and non-discrete value.

This paper centres on the selection of instances, which has been widely studied, focusing above all on classification. The problem has not been studied as much in relation to regression data sets, among other reasons because of the complexity of this type of data set [3]. While in classification, the number of classes or values to be predicted is usually very low (the simplest example would be binary problems), the output variable in regression is continuous, such that the number of possible values to predict is unlimited. This paper seeks to apply all the algorithms conceived for classification purposes to regression problems, on the basis of a meta-model.

As we mentioned before, a few studies have been conducted on the application of instance-selection techniques to data sets with a numerical and non-discrete output variable and there are few algorithms that are specifically designed for that purpose [5].

The idea that we propose in this article is the possibility of applying to regression data sets, instance selection methods designed for classification tasks. To do so, the output variable is previously discretized in such a way that the data set is turned into a classification problem. Having completed the selection of instances, the numerical value of the output variable is recovered for the selected instances. Can be considered a meta-algorithm, as it allows to use in regression any previously existing classification method for instance selection, for example: MSS, DROP, LSBo. . .

The main advantages of the proposed method are its simplicity and its adaptability. The method is simple since the discretization stage is quite straightforward. On the other hand, it can be easily applied to any instance selection algorithm for classification, no changes are needed in the algorithms as they are used as black boxes and the discretization is completely independent of the chosen algorithm.

Besides simplicity, another major advantage of the method is that we now have at our disposal all the algorithms of instance selection for classification existing in the literature. The utility of the approach has been proved experimentally, it offers competitive results when compare to the few existing methods of instance selection for noise removal in regression [2,4], despite that the latter propose a more complex and sophisticated approaches to the problem. More specifically, its performance as a noise filter has been compared *i*) taking as reference the values of: the RMSE, the F_1 score, the *G mean* and the compression; *ii*) with different base regressors: *k*-nearest neighbours, RBF networks and REPTree, and *iii*) using 29 data sets to which several levels of noise were added.

Acknowledgments

This work was supported by the project TIN2011-24046 of the Spanish Ministry of Economy and Competitiveness.

Referencias

1. Álvar Arnaiz-González, José F. Díez-Pastor, Juan J. Rodríguez, and César Ignacio García-Osorio. Instance selection for regression by discretization. *Expert Systems with Applications*, 54:340 – 350, 2016.
2. A. Guillen, L.J. Herrera, G. Rubio, H. Pomares, A. Lendasse, and I. Rojas. New method for instance or prototype selection using mutual information in time series prediction. *Neurocomputing*, 73(10-12):2030 – 2038, 2010. Subspace Learning / Selected papers from the European Symposium on Time Series Prediction.
3. Mirosław Kordos, Szymon Białka, and Marcin Blachnik. Instance selection in logical rule extraction for regression problems. In Leszek Rutkowski, Marcin Korytkowski, Rafał Scherer, Ryszard Tadeusiewicz, LotfiA. Zadeh, and JacekM. Zurada, editors, *Artificial Intelligence and Soft Computing*, volume 7895 of *Lecture Notes in Computer Science*, pages 167–175. Springer Berlin Heidelberg, 2013.
4. Mirosław Kordos and Marcin Blachnik. Instance selection with neural networks for regression problems. In *Proceedings of the 22nd international conference on Artificial Neural Networks and Machine Learning - Volume Part II*, ICANN'12, pages 263–270. Springer-Verlag, Berlin, Heidelberg, 2012.
5. Miloš B. Stojanović, Miloš M. Božić, Milena M. Stanković, and Zoran P. Stajić. A methodology for training set instance selection using mutual information in time series prediction. *Neurocomputing*, 141(0):236 – 245, 2014.