Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Fusion of instance selection methods in regression tasks[☆]

Álvar Arnaiz-González[a], Marcin Blachnik[b], Mirosław Kordos[c], César García-Osorio[a,*]

[a] University of Burgos, Spain
[b] Silesian University of Technology, Poland
[c] University of Bielsko-Biala, Poland

ARTICLE INFO

Article history:
Received 3 February 2015
Revised 20 November 2015
Accepted 8 December 2015
Available online 14 December 2015

Keywords:
Instance selection
Regression
Ensemble models

ABSTRACT

Data pre-processing is a very important aspect of data mining. In this paper we discuss instance selection used for prediction algorithms, which is one of the pre-processing approaches. The purpose of instance selection is to improve the data quality by data size reduction and noise elimination. Until recently, instance selection has been applied mainly to classification problems. Very few recent papers address instance selection for regression tasks. This paper proposes fusion of instance selection algorithms for regression tasks to improve the selection performance. As the members of the ensemble two different families of instance selection methods are evaluated: one based on distance threshold and the other one on converting the regression task into a multiple class classification task. Extensive experimental evaluation performed on the two regression versions of the Edited Nearest Neighbor (ENN) and Condensed Nearest Neighbor (CNN) methods showed that the best performance measured by the error value and data size reduction are in most cases obtained for the ensemble methods.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

In real-world problems, we are often faced with regression tasks, aimed at predicting some real-value numbers. The quality of the prediction is often limited not by the learning algorithm, but by the quality of the data itself. There are various ways of improving the quality of the data on which the learning model is built. One of them is instance selection, in which we reject the training instances that are either not expected to improve or are expected to worsen prediction results.

In this paper, we use ‘edited’ or ‘filtered’ dataset interchangeably to refer to a dataset produced by an instance selection process.

Traditionally, instance selection methods are divided in two groups according to the approach used by algorithms to select an instance [29]:

- *Wrappers*: the algorithms of this family use the accuracy obtained by the classifier (usually the same that will be used after instance selection stage). Some of them are CNN [17], ENN [41], DROP1…5 [40], ICF [8], MSS [4]…
- *Filters*: as opposed to *wrappers*, the instance selection or rejection criterion is not based on a classifier. For example POP [32].

A large number of instance selection algorithms designed for classification tasks have been published in the scientific literature and new ones continue to appear. An up-to-date taxonomy can be found in [13].

However, two issues still require prompt attention: extending instance selection methodologies into regression tasks and applying the fusion of various sources to obtain better results. Regarding the first issue, we have already obtained some results in [21] and in this paper we propose a new approach based on the discretization of the output variable. The second issue has proven to be an effective approach in other domains [24], but to the best of our knowledge, no one has yet tried to apply ensembles to the combination of instance selection methods for regression.

The presence of noise is a common problem in Data Mining applications and noise filters are widely used to cope with this problem [35]. The usefulness and efficiency of instance selection methods for low quality and noisy data was confirmed in [23], which presented an experimental evaluation of several approaches to noise-resistant training of multilayer perceptron neural networks for classification and regression problems. Different noise reduction methods were evaluated with different noise levels in the data. For regression tasks, in most cases the solution that used the threshold-based ENN instance selection (in the following sections, we describe threshold-based ENN, short of Edited Nearest Neighbor, in detail) achieved the

[☆] This work was supported by the Project TIN2011-24046 of the Spanish Ministry of Economy and Competitiveness.

[*] Corresponding author. Tel.: +34947259358; fax: +34947258910.
E-mail addresses: alvarag@ubu.es (Á. Arnaiz-González), marcin.blachnik@polsl.pl (M. Blachnik), mkordos@ath.bielsko.pl (M. Kordos), cgosorio@ubu.es (C. García-Osorio).

http://dx.doi.org/10.1016/j.inffus.2015.12.002
1566-2535/© 2015 Elsevier B.V. All rights reserved.

lowest mean-square error (MSE) on the test set for various levels of noise added to the data. The only exception was when the noise level was extremely high and for that case the lowest MSE was obtained by a combination of the threshold-based ENN with modification of the error function.

This paper is structured as follows: Section 2 is a brief review of instance selection methods for classification; Section 3 explains the problems of applying them to regression tasks; Section 3.1 presents the threshold-based approach, while Section 3.2 suggests another idea based on numerical target discretization; Section 4 shows the possibility of applying the idea of ensembles to the instance selection paradigm. Then, in Section 5 an experimental study is performed to compare both approaches and the improvements achieved when they are used within an ensemble, and how, in this case, we can outperform the regressor trained with the whole dataset. Finally, Section 6 summarizes the most important conclusions, and Section 7 presents some future lines of research.

2. Instance selection for classification

In general, instance selection algorithms can be classified into three types: noise filters, condensation algorithms and prototype selection algorithms [19]. Noise filters remove instances that show extreme differences with their neighbors, as they are considered noise. In the Edited Nearest Neighbor (ENN) algorithm [41] used for classification tasks, an instance is removed if it belongs to a different class than that predicted by the k Nearest Neighbors algorithm (kNN), using the k nearest neighbors of the instance of interest. The condensation methods serve to reduce the number of instances in the dataset by removing instances that are overly similar to their neighbors. The Condensed Nearest Neighbor algorithm (CNN) [17] removes the instances in classification problems that are correctly predicted by kNN, as it is assumed that they do not bring any new information to build the classifier. Prototype selection methods do not select relevant instances, but rather transform the whole dataset into a few instances, each covering an area corresponding to the Voronoi cell[1] defined by each instance. An example of such methods is Learning Vectors Quantization (LVQ) [20].

The decisions about instance selection or rejection are very straightforward in instance selection for classification problems. They are based on the classification results obtained with an algorithm that is almost always kNN. The instance can either belong to the same class or to another that is different from the majority of its neighbors. The difference between the predicted and the real class of the instance determines the decision on its acceptance/rejection.

3. Instance selection for regression

There are very few papers that address the problem of instance selection for regression. The proposed methods fall into two categories: evolutionary-based and nearest neighbor-based (which includes our methods). The evolutionary based methods have very high computational cost, about 3 to 4 order of magnitude higher than the nearest neighbor-based methods for medium size datasets (according to the experiments and the data reported in [2]). For bigger datasets the difference is even larger, what makes the methods impractical in many applications.

However, we did not find in a literature any work on instance selection for regression tasks in a general setting, where the authors reported the improvements in terms of compression rate and classification accuracy even for a single arbitrary point, not to mention even the Pareto front, which we have examined in our work, so we were not able to use those methods for comparison.

Tolvi [39] presented the use of genetic algorithms for outlier detection. That approach does not rely on the dataset properties and purpose, so it can be used both for classification and regression problems. However they presented the results of only two very small datasets (35 instances with 2 features and 21 instances with 3 features).

In [2] Antoneli et al. also presented a genetic approach using quite complex multiobjective evolutionary algorithms with up to 300,000 evaluations of the fitness function.

In [15] Guillen et al. presented the use of mutual information for instance selection in time series prediction. Their idea was similar to k-NN based prediction, thus they determined the nearest neighbors of a given point and then instead of using k-NN to predict the output value, they calculated the mutual information (MI) between that point and each of their neighbors. Then they considered the loss of MI with respect to its neighbors in so that if the loss was similar to the vectors near the examined vector, this vector was included in the filtered dataset. They evaluated their methods on artificially generated data with one and two input features. Later on, Stojanović et. al [37] extended the previous idea to instance selection in time series prediction.

In [33] a Class Conditional Instance Selection for Regression (CCISR) is proposed. The method is based on CCIS for classification [27]. CCIS builds two graphs: one for the nearest neighbors of the same class as a given instance and another for other class instances. Then a scoring function based on the distances in graphs is used to evaluate the instances. In the regression version (CCISR), instead of using only the nearest instances to construct these graphs, the neighborhood is defined based on the probability density function of the examples. The authors provided results on several real-world datasets for fuzzy rule based systems in terms of MSE and the number of extracted rules.

In the following subsections, we present our approaches two instance selection for regression.

3.1. Threshold-based instance selection for regression

Instance selection methods developed for classification are based on the analysis of class labels of neighbor instances to determine the rejection/acceptance of a test instance. It usually takes the form of *if* statements where the condition of the *if* statement compares labels of the actual test instance with its neighbors. In the case of regression, label comparison is not possible since output values are not nominal but continuous. Thus, rather than labels, another quantity must be compared. An intuitive approach proposed in [21] is to use a similarity-based error measure such as:

$$Error = |Y_{real} - \bar{Y}_{predicted}| \tag{1}$$

and to compare the error to the predefined threshold, to determine whether the instance should be accepted or rejected:

$$\text{if } (Error > \theta) \text{ then} \ldots \tag{2}$$

where (…), as the consequence of the if, can denote either instance acceptance or rejection, depending on a given algorithm. To improve the performance of this method, the threshold can depend on the local properties of the dataset. When the standard deviation of the adjacent cases (to the test case) is high, the condition should be less sensitive to the value of error shown in Eq. (1). In other cases, when the local standard deviation is small, the condition should be more

[1] A Voronoi diagram (also known as Thiessen polygons or Dirichlet tessellation) is a geometrical construction to partition the Euclidean space. It divides the space into several polygons, Voronoi cells, each of them associated with a point, the seed. Each Voronoi cell defines a region of the space closer to its seed than to any other point, that is, the division is made according to the nearest neighbor rule [3].

Download English Version:

https://daneshyari.com/en/article/528025

Download Persian Version:

https://daneshyari.com/article/528025

Daneshyari.com