Expert Systems With Applications 54 (2016) 340–350

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Instance selection for regression by discretization

Álvar Arnaiz-González, José F. Díez-Pastor, Juan J. Rodríguez, César Ignacio García-Osorio*

University of Burgos, Civil Engineering, Escuela Politécnica Superior, Avda. Cantabria, s/n Burgos 09006, Province of Burgos, Spain

ARTICLE INFO

Keywords:
Instance selection
Regression
Mutual information
Noise filtering
Class noise

ABSTRACT

An important step in building expert and intelligent systems is to obtain the knowledge that they will use. This knowledge can be obtained from experts or, nowadays more often, from machine learning processes applied to large volumes of data. However, for some of these learning processes, if the volume of data is large, the knowledge extraction phase is very slow (or even impossible). Moreover, often the origin of the data sets used for learning are measure processes in which the collected data can contain errors, so the presence of noise in the data is inevitable. It is in such environments where an initial step of noise filtering and reduction of data set size plays a fundamental role. For both tasks, instance selection emerges as a possible solution that has proved to be useful in various fields. In this paper we focus mainly on instance selection for noise removal. In addition, in contrast to most of the existing methods, which applied instance selection to classification tasks (discrete prediction), the proposed approach is used to obtain instance selection methods for regression tasks (prediction of continuous values). The different nature of the value to predict poses an extra difficulty that explains the low number of articles on the subject of instance selection for regression.

More specifically the idea used in this article to adapt to regression problems "classic" instance-selection algorithms for classification is as simple as the discretization of the numerical output variable. In the experimentation, the proposed method is compared with much more sophisticated methods, specifically designed for regression, and shows to be very competitive.

The main contributions of the paper include: (*i*) a simple way to adapt to regression instance selection algorithms for classification, (*ii*) the use of this approach to adapt a popular noise filter called ENN (edited nearest neighbor), and (*iii*) the comparison of this noise filter against two other specifically designed for regression, showing to be very competitive despite its simplicity.

1. Introduction

Automatic supervised learning begins with a dataset 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.

The problem of using a finite set of examples to learn the relation between the values of the dependent and independent variables is not unique to Machine Learning, but is also a classic problem of statistics and pattern recognition. There are many real applications in which a solution to this problem would be of interest, among which figure image processing (Rui, Huang, & Chang, 1999; Wang, Huang, Luo, Wang, & Luo, 2011), speech recognition (Trentin & Gori, 2001), genome sequencing (García-Pedrajas, Pérez-Rodríguez, García-Pedrajas, Ortiz-Boyer, & Fyfe, 2012), industrial processes (Harding, Shahbaz, Srinivas, & Kusiak, 2005), fraud detection (Lei & Ghorbani, 2012), and software engineering (Serrano, Gómez-Sanz, Botía, & Pavón, 2009), finances (Sun & Li, 2011), to mention only a few.

Regardless of the dataset that is analyzed, the presence of noise in the real-world applications is common (Garcia, de Carvalho, & Lorena, 2015; Liu, Yamashita, & Ogawa, 1995; Sáez, Luengo, & Herrera, 2013; Wu & Zhu, 2008), besides reduce learning abilities of models (Zhu & Wu, 2004) and their elimination is by no means a clear cut process (García-Osorio, de Haro-García, & García-Pedrajas, 2010). Various methods have been proposed for their detection and elimination that follow various approaches. This paper centres on

* Corresponding author. Tel.: +34 947259358; fax: +34 947258910.
E-mail addresses: alvarag@ubu.es (Á. Arnaiz-González), jfdpastor@ubu.es (J.F. Díez-Pastor), jjrodriguez@ubu.es (J.J. Rodríguez), cgosorio@ubu.es (C.I. García-Osorio).

http://dx.doi.org/10.1016/j.eswa.2015.12.046
0957-4174/

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 datasets, among other reasons because of the complexity of this type of dataset (Kordos, Białka, & Blachnik, 2013). 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 (Kordos & Blachnik, 2012). This paper seeks to apply all the algorithms conceived for classification purposes to regression problems, on the basis of a meta-model.

The main contributions of the paper are:

- An approach for adapting to regression instance selection method initially designed for regression. It makes available a wide-range of instance selection methods for regression to researchers.
- The proposed approach was used to adapt ENN (Wilson, 1972) (edited nearest neighbor) to regression.
- The performance of the new model was compared against two state-of-the-art noise filters for regression (Guillen et al., 2010; Kordos & Blachnik, 2012).

This article has the following structure: first, the instance selection process for classification is explained and the problems involved in applying this technique to regression are analyzed. Then, the proposed method, consisting of the discretization of the numerical variable, is presented in the Section 3. In Section 4, the results of the experimentation are analyzed, and in the final section, the conclusions are extracted and future lines of work are suggested.

2. Instance-selection methods

Instance-based learners, also called lazy learners, are very effective, despite its simplicity (Brighton & Mellish, 2002) and nowadays are still frequently used in experimental studies in Machine Learning (Leyva, González, & Pérez, 2015). However, these methods suffer from various disadvantages (Kononenko & Kukar, 2007). They are very sensitive to the presence of noise in the training data (Nanni & Lumini, 2011). In addition, few algorithms are able to generate results within a reasonable time span when using large-sized datasets that need to be processed these days (Kordos & Blachnik, 2012).

According to Jankowski and Grochowski (2004), instance selection serves two purposes: to reduce noise and to eliminate outliers in the datasets (noise filters); and, to reduce the complexity of instance-based learning algorithms (Aha, Kibler, & Albert, 1991) (condensing algorithms), with the intention of reducing the number of examples in the training set.

Kordos and Blachnik (2012) explained that industries require expert and intelligent systems to optimize their processes. Datasets in industries are huge and require techniques able to reduce their complexity before building the prediction models. Moreover, instance selection has been used not only in industrial processes, below we described some other fields of application:

- Steel industry: We have found two works (Kordos & Blachnik, 2012; Koskimaäki, Juutilainen, Laurinen, & Roning, 2008) in which authors found necessary to reduce the size of datasets due to the great number of samples that are available in steel processes. Whereas Koskimaäki et al. (2008) tried to reduce the size as previous stage of clustering, Kordos and Blachnik (2012) used neural networks after instance selection process.
- Stock markets: In stock market analysis, Kim (2006) developed a new genetic instance selection method to reduce the complexity of induced solutions. They used the direction of change (increase or decrease of the stock index from one day to the next) in the daily Korea stock price index. Stock markets are complex and noisy, thus the instance selection method proposed gave the chance to improve the performance of artificial neural networks.
- Bankruptcy prediction: Financial institutions need accurate bankruptcy prediction models, since is essential for their risk management. In (Ahn & Kim, 2009) the combination of instance and feature selection was tested with the aim to improve, using genetic algorithms, the performance of case-base reasoning (CBR).
- Computer vision: A challenging problem in computer vision is the recognition of traffic signs (TSR) because in autonomous vehicles it has a crucial impact on driver safety. Chen, Lin, Ke, and Tsai (2015) used a genetic algorithm for feature and instance selection over a benchmark of TSR, as opposed to Ahn and Kim (2009), they performed instance and feature selection separately. They addressed the trade-off between accuracy, data set size reduction and time spent during the selection process, which is always present.
- Time series: The domains where time-series classification is used are wide, including finance, networking, medicine, astronomy, robotic, chemistry and industry (Keogh & Kasetty, 2002). Therefore the use of instance selection techniques for this problem is already under investigation (Buza, Nanopoulos, & Schmidt-Thieme, 2011; Guillen et al., 2010).

2.1. Instance selection for classification

Instance-selection algorithms are intended to reduce the complexity of the learning algorithms by reducing the number of examples, they extract the most significant and discard those that do not provide valuable information (García, Marqués, & Sánchez, 2012b), for example, the outliers and the examples introduced as a consequence of noise in the measurement process.

The literature contains a large number of instance-selection algorithms designed for classification purposes and new ones continue to appear. An up-to-date taxonomy may be found in García, Derrac, Cano, and Herrera (2012a).

The need for instance selection becomes obvious when the datasets used in real life are examined. Attempts to train a classifier, for example, on the basis of millions of instances can be a difficult, even an insurmountable task. The selection of instances therefore appears to be a good alternative, to reduce the complexity of the sample, enabling its subsequent treatment.

The term “instance selection” brings together a range of procedures and algorithms that are intended for the selection of a representative subset of the initial training dataset (Kim, 2006). A first classification of these techniques is usually done using as criteria the purpose of their application, dividing them into two large groups: editing (or noise filtering), and condensation algorithms.

2.1.1. Edition techniques

Noise filtering techniques attempt to eliminate the erroneously labelled instances from the training set and, at the same time, they attempt to clear possible overlaps between regions of different classes. In other words, their principal objective is to achieve compact and homogeneous groups; one of such techniques is the *Wilson's editing algorithm* (Wilson, 1972). If an instance is badly classified on the basis of the rule *k*-NN, it will basically eliminate that instance from the training set (Algorithm 1).

2.1.2. Condensation techniques

One of the problems that arises when real-world datasets are analyzed is the large number of examples that they contain, making the learning process computationally costly and preventing the