Information Sciences 266 (2014) 16–30

Contents lists available at ScienceDirect

Information Sciences

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

On the use of meta-learning for instance selection: An architecture and an experimental study ☆

Enrique Leyva *, Yoel Caises, Antonio González, Raúl Pérez

Dpto de Ciencias de la Computación e IA, ETSIIT, Universidad de Granada, Spain

ARTICLE INFO

Article history:
Received 14 October 2011
Received in revised form 28 November 2013
Accepted 1 January 2014
Available online 11 January 2014

Keywords:
Instance selection
Prototype selection
Meta-learning
Machine learning
Classification

ABSTRACT

Many authors agree that, when applying instance selection to a data set, it would be useful to characterize the data set in order to choose the most suitable selection criterion. Based on this hypothesis, we propose an architecture for knowledge-based instance selection (KBIS) systems. It uses meta-learning to select the best suited instance selection method for each specific database, among several methods available. We carried out a study in order to verify whether this architecture can outperform the individual methods. Two different versions of a KBIS system based on our architecture, each using a different learner, were instantiated. They were evaluated experimentally and the results were compared to those of the individual methods used.

© 2014 Elsevier Inc. All rights reserved.

1. Introduction

Instance selection (IS) is a strategy to solve the problem of applying machine learning techniques to large databases. It consists in searching for a set S in the vector space of the original set of instances T, with $|S| < |T|$, such that if a classifier is trained on S, its classification accuracy is as high as if it is trained on T. This reduction in the number of instances usually leads to a significant reduction in the training times.

Many different methods have been proposed since the 1960s[1] Most of them were developed to reduce the classification time and the storage requirements of the kNN classifier [14]. Traditionally, each IS proposal has a specific selection criterion that applies to any database. However, in recent years many authors have come to the conclusion that data characteristics are crucial to the success of the method used and that a single selection criterion is not sufficient to guarantee success over a wide range of environments.

An option for adapting the selection criteria to input data is the use of meta-heuristics like genetic algorithms. Genetic algorithms are able to explore wide regions of the solution space and, therefore, often achieve good results regardless of the environment. However, their high computational cost compared to other approaches is a major disadvantage for practical applications. A second alternative (which we followed for our work) is to use domain knowledge to choose the most appropriate selection strategy among several options. The arguments raised by most of the authors to which we referred earlier are in this line.

☆ This work has been partially funded by the Andalusian Regional Government Project P09-TIC-04813, the Spanish MEC Project TIN2012-38969, and cofinanced by FEDER funds (European Union).

* Corresponding author. Tel.: +34 958 244019; fax: +34 958 243317.
E-mail addresses: eleyvam@decsai.ugr.es (E. Leyva), ycaises@decsai.ugr.es (Y. Caises), A.Gonzalez@decsai.ugr.es (A. González), Raul_Perez@decsai.ugr.es (R. Pérez).

[1] For a survey of the main IS methods see [39].

0020-0255/$ - see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2014.01.007

These authors include Brighton and Mellish [8], who suggested that algorithms that retain instances near to class borders (border points) are successful in domains where the majority of instances are surrounded by instances from the same class (homogeneous class structures), whereas in non-homogeneous domains it is better to retain the instances that provide the best representation of their neighborhood (prototypes), which tend to be internal points.

In a survey on "creative prototype reduction schemes", Kim and Oommen [31], compared several methods using artificial and real life data, and concluded that no single scheme is superior to every other for all contexts, and that the data characteristics determine which method is preferred.

In another paper, Reinartz [42] analyzed some unresolved issues and research challenges for IS, and proposed among other research areas, to:

- "*develop more intelligent focusing solutions that provide data reduction techniques beyond pure statistical sampling and make use of the specific characteristics of concrete contexts in data mining*"
- "*extend attempts of analytical studies and experimental results to understand the relation between different instance selection techniques and to come up with reliable heuristics and guidelines for the selection of best suited focusing solutions given a specific data mining environment*"

Mollineda et al. [36] redefined, for multi-class problems, some data complexity measures previously presented for binary problems by Ho and Basu [29]. They also compared the results of the IS algorithms CNN [28] and ENN [55], as well as a combination of both, in several databases having different complexities. The same measures were later taken up by García et al. [21] to diagnose the effectiveness of an evolutionary IS method. Another set of measures is proposed by Caises et al. [9], as well as an algorithm that uses these measures in several empirical rules to determine the most suitable IS method (or combination of methods) for each database.

This paper presents a new approach based on meta-learning to obtain meta-models that will help to select the best IS method for each database. We propose an architecture for knowledge-based instance selection (KBIS) systems and instantiate two versions of a KBIS system using different learners. The article is organized as follows. Section 2 summarizes some background knowledge about the IS and meta-learning fields. Section 3 describes the proposed architecture and its instantiations. Finally, Section 4 compares experimentally both versions of the KBIS system with the individual methods they use.

2. Background

In this section, we will briefly review the main approaches in the IS field, some meta-learning foundations, and the relationship between both fields.

2.1. Instance selection

The IS problem has been addressed by many authors with different approaches since the 60s in the last century. Here we present a brief review on the most representative approaches in this field. Readers interested in a more comprehensive review can see [39], and two extensive experimental studies that have been published recently in [23,52].

The first proposal for IS was *condensed Nearest Neighbor* (CNN) [28]. Its strategy is to retain the class border instances and discard the internal ones. To accomplish this, CNN searches for a consistent subset, which means that every instance in T must be correctly classified by S. This strategy has been known as **condensation** and has been followed by several subsequent methods. Most of them try to improve the results of CNN by producing smaller subsets; some examples are *Reduced Nearest Neighbor* (RNN) [25], *Selective Nearest Neighbor* (SNN) [44], *Minimal Consistent Set* (MCS) [15], and *Fast Nearest Neighbor Condensation*(FCNN) [6]. Other proposals that retain border instances are *Editing by Ordered Projections* (EOP) [2] and *Patterns with Ordered Projections* (POP) [43], but unlike the previous ones, they do not try to select a consistent S. Both methods search for hyper-rectangles where all the contained instances have the same class. They retain an instance only if it is indispensable in order to delimit one of these hyper-rectangles in at least one dimension. All these methods are very sensitive to noise because noisy instances are interpreted as border points, so they are not removed from S and cause the noise proportion in S to be much higher than in T, thus affecting the classification accuracy of unseen instances.

Edition is the opposite strategy to condensation: It discards instances that disagree in classification with their neighborhoods. Methods of this kind are good noise filters but achieve little reductions in the number of instances. The first and most popular edition method was *Edited Nearest Neighbor* (ENN) [55], which removes all the instances that are misclassified by 3NN using T as knowledge base. Other edition techniques, such as *Repeated ENN* (RENN) and All-kNN [51], are modifications of ENN that apply this method iteratively. Nevertheless, several studies [11,26,56] show that such modifications produce slightly better levels of data reduction than ENN, and accuracies very similar or worse than those of this method. Despite its age, ENN still is the most used edition method, due to the quality of the results it provides and its low computational cost.

After more than two decades of proposals based on edition and condensation, a new trend emerged in the 90 s, focusing on getting benefits from the combination of both strategies. *Instance Based Learning 3* (IB3) [3] was the first member of this category of methods known as **hybrid**. Like CNN, it retains misclassified instances, but only uses for classification those

Download English Version:

https://daneshyari.com/en/article/392684

Download Persian Version:

https://daneshyari.com/article/392684

Daneshyari.com