Pattern Recognition 48 (2015) 1523–1537

Contents lists available at ScienceDirect

Pattern Recognition

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

Three new instance selection methods based on local sets: A comparative study with several approaches from a bi-objective perspective

Enrique Leyva*, Antonio González, Raúl Pérez

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

ARTICLE INFO

Article history:
Received 21 November 2013
Received in revised form
22 September 2014
Accepted 1 October 2014
Available online 12 October 2014

Keywords:
Local sets
Instance selection
Data reduction
Prototype-based classifiers
Instance-based learning

ABSTRACT

The local set is the largest hypersphere centered on an instance such that it does not contain instances from any other class. Due to its geometrical nature, this structure can be very helpful for distance-based classification, such as classification based on the nearest neighbor rule. This paper is focused on instance selection for nearest neighbor classification which, in short, aims to reduce the number of instances in the training set without affecting the classification accuracy. Three instance selection methods based on local sets, which follow different and complementary strategies, are proposed. In an experimental study involving 26 known databases, they are compared with 11 of the most successful state-of-the-art methods in standard and noisy environments. To evaluate their performances, two complementary approaches are applied, the Pareto dominance relation and the Technique for Order Preference by Similarity to Ideal Solution. The results achieved by the proposals reveal that they are among the most effective methods in this field.

1. Introduction

The k-Nearest Neighbors (kNN) classifier [1], despite its age, remains as one of the most used classifiers and it is frequently found as a benchmark for experimental studies in machine learning. It shows good performance on several domains due to its capacity to learn complex target functions. It is also popular for practical applications because it is easy to understand and to implement.

As kNN keeps all training instances, it has large storage requirements and needs a great deal of computational time for classification. There are several proposals to speed up kNN classification; some of them provide fast search methods to find the nearest neighbors, while others reduce the size of the case base and try to preserve the classification accuracy. The second approach can be focused on reducing the number of features [2,3], reducing the number of instances [4–6], or reducing both features and instances [7,8]. This paper is focused on the problem of reducing the number of instances, known as instance selection (IS).

Ideally, IS consists in searching for the minimal set S in the same vector space as the original set of instances T, subject to $accuracy(S) \geq accuracy(T)$, where the constraint means that the accuracy of any classifier trained with S must be at least as good as that of the same classifier trained with T. Nevertheless, in practice this is not a hard constraint, and IS is mostly treated as the bi-objective problem of maximizing accuracy and reduction. Hence, for most problems, there is no single best solution (method). Instead, there is a Pareto front with several no-dominated solutions.

Although there are plenty of IS methods in scientific literature, most of them fail to achieve a good balance between reduction and accuracy. On one hand, those that provide great reductions give accuracies below that achievable with all the instances. On the other hand, those that provide accuracy improvements barely reduce the number of instances. Moreover, even these methods tend to affect the classification accuracy in some databases. On top of that, few methods that seem to achieve an acceptable balance between reduction and accuracy exhibit a poor performance when noise is present.

Some previous works allowed us to realize that the local set (LS) concept [9] can be used to develop IS methods capable to overcome the limitations mentioned above. LS is defined as the largest hypersphere centered on an instance such that it does not contain instances from any other class. This concept emerged in the IS field and has remained relatively unused since its publication. However, it has great potential for several tasks related to machine learning, such as supervised clustering, data characterization and IS.

In [10], LSs were used as the basis of a supervised clustering algorithm. The results of this LS-clustering were also used in an IS

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

http://dx.doi.org/10.1016/j.patcog.2014.10.001
0031-3203/

method included in a selective combination of IS methods. More recently, in [11], five different IS strategies based on LSs were used in the context of a meta-learning system. In these two papers and in [12], several data-characterization measures based on LSs, and conceived for systems that apply meta-learning to IS, were used.

In addition to the empirical evidence of the usefulness of LSs provided by these works, there are some properties inherent to the concept that make it attractive in the context of IS for distance-based classification. First, it is a geometrical construct based on distances, intuitive, and easy to understand. Also, it can be computed with low computational cost. Furthermore, it is very sensitive to noise, a fact that can be used to detect and remove harmful instances. Finally, the LS provides a compact description of the instance neighborhood that can be used to determine whether it is relevant or superfluous for the classification of new instances.

Although the LS concept was successfully used in [10–12] for IS-related tasks, its potential as the basis to build IS methods was not addressed in these works. In this paper we propose three IS methods based on the experiences gathered from these works. They follow different and complementary strategies in order to determine which instances to remove or retain. One of them removes noisy and overlapped instances. It barely reduces the database, but increases the accuracy more than any other method. Moreover, it plays an important role in the other two methods. Another method selects the centroids from clusters, and its main strength lies in the reduction. The third method selects the border instances and achieves the best compromise between accuracy and reduction. Although they have different reduction–accuracy priorities, as will be seen in the experimental study, all of them offer non-dominated solutions in the Pareto frontier of the accuracy–reduction maximization problem.

The experimental study documented in this paper compares the performance of the different proposals and 11 of the most successful state-of-the-art methods on 26 known databases. This study is focused on selecting instances for the kNN classifier. The proposals and benchmark methods are strongly related to distance-based classifiers, so they are not expected to provide good selections for other families of classifiers.

We compared the performance of all the methods in terms of accuracy and reduction, and carried out statistical tests on them. Furthermore, since IS is a bi-objective problem, we assessed the performance using two well-known multi-criteria-decision-making tools: the Pareto dominance relation, and the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS). Both of them revealed that the proposals are top-performing methods in the field and overcome the previously mentioned limitations observed in IS methods.

The paper is organized as follows. In Section 2 we comment some of the most relevant approaches in the IS field. In Section 3 we review some concepts and three previous works related to the use of LSs for IS. Then, in Section 4 we describe the three proposed methods. Finally, in Section 5 we report the results of the experimental study.

2. Fundamental instance selection approaches

The first works in the field of IS date back to the 1960s. Since then, several methods have been proposed to tackle this problem. Most of them have been related to case based reasoning and aimed at reducing the storage requirements of instance based classifiers such as kNN. Nonetheless, some works have been focused on other machine learning paradigms, like fuzzy rule-based systems [13], neural networks [14], ensembles of classifiers [15], decision trees [16], and support vector machines [17].

Whether, and under which conditions, the same selection may be successfully used to train different classifiers is an open question, but this topic is out of the scope of this paper. For interested readers, it is worth to mention that some works [10,18–21] that address such reuses suggest that it depends at least on the selection method, the classifiers, and the database.

We mention here the IS methods that were considered in the study. All of them are related to kNN, and they are a representative sample of the fundamental approaches in this field.[1] Table 1 shows them grouped by selection strategy and sorted by publication year. There are three fundamental selection strategies in the literature (*condensation*, *edition*, and *hybrid*), each having several representatives. We included in the study the first representative of each family, as well as some recent proposals that exhibit good performance.

The *Condensed Nearest Neighbor* (CNN) method [22] was the first proposal for IS found in the literature. This method searches for a consistent subset, which means that every instance in T must be correctly classified by S; this strategy is known as condensation. CNN retains the class border instances and discards the internal ones. Most of the subsequent condensation methods try to improve the results of CNN by producing smaller subsets. Some examples are *Reduced Nearest Neighbor* (RNN) [23], *Selective Nearest Neighbor* (SNN) [24], *Minimal Consistent Set* (MCS) [25], and *Fast Nearest Neighbor Condensation* (FCNN) [26].

Along with CNN, we considered in this study one of the latest condensation methods, *Prototype Selection by Clustering* (PSC) [27]. This method uses clustering to divide the database in small subsets (clusters). Then, PSC selects the medoids of those clusters having only one class, and the border points of clusters containing instances from several classes. The main advantage of this method is its high efficiency.

In general, condensation methods are very noise sensitive because noisy instances are interpreted as border and remain in S. That causes a much higher noise proportion in S than in T, which affects the accuracy of classification for unseen instances.

An opposite strategy, known as *edition*, is followed in another pioneer IS method, *Edited Nearest Neighbor* (ENN) [28]. It discards the instances that disagree in classification with their neighborhoods. Methods like this one are characterized by being good noise filters, and achieve little reductions in the number of instances. Some other edition techniques have been proposed, such as *Repeated ENN* (RENN) and All-kNN [29], which are modifications of ENN that iteratively apply this method. Nevertheless, several studies [18,30,31] show that such modifications produce slightly better levels of data reduction than ENN, and their accuracy is very similar or worse. Despite its age, ENN remains as the most popular edition method because it provides good results with low computational costs.

Along ENN, we also considered in this study the three latest edition methods found in the literature: *Nearest Centroid Neighbor Edition* (NCNEdit) [32], *Edited Normalized Radial Basis Function* (ENRBF) [33], and *Edited Nearest Neighbor Estimating Class Probabilistic and Threshold* (ENNth) [34].

After more than two decades of edition and condensation proposals, a new trend emerged in the 90s with the intent to obtain benefits from the combination of both strategies. *Instance Based Learning 3* (IB3) [35] was the first member of this category of methods known as hybrid. It builds the selection incrementally, using previous selected instances to classify unprocessed ones. Like CNN, IB3 retains instances that are misclassified, but in order to classify new instances it uses only those instances whose

[1] Readers interested in a more comprehensive review can see [4]. Also, an extensive experimental study was recently published in [5].