Pattern Recognition 46 (2013) 355–364

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Predicting noise filtering efficacy with data complexity measures for nearest neighbor classification

José A. Sáez [a,*], Julián Luengo [b], Francisco Herrera [a]

[a] Department of Computer Science and Artificial Intelligence, University of Granada, CITIC-UGR, Granada 18071, Spain
[b] Department of Civil Engineering, LSI, University of Burgos, Burgos 09006, Spain

ARTICLE INFO

Article history:
Received 8 March 2012
Received in revised form
6 July 2012
Accepted 14 July 2012
Available online 23 July 2012

Keywords:
Classification
Noisy data
Noise filtering
Data complexity measures
Nearest neighbor

ABSTRACT

Classifier performance, particularly of instance-based learners such as *k*-nearest neighbors, is affected by the presence of noisy data. Noise filters are traditionally employed to remove these corrupted data and improve the classification performance. However, their efficacy depends on the properties of the data, which can be analyzed by what are known as data complexity measures. This paper studies the relation between the complexity metrics of a dataset and the efficacy of several noise filters to improve the performance of the nearest neighbor classifier. A methodology is proposed to extract a rule set based on data complexity measures that enables one to predict in advance whether the use of noise filters will be statistically profitable. The results obtained show that noise filtering efficacy is to a great extent dependent on the characteristics of the data analyzed by the measures. The validation process carried out shows that the final rule set provided is fairly accurate in predicting the efficacy of noise filters before their application and it produces an improvement with respect to the indiscriminate usage of noise filters.

© 2012 Elsevier Ltd. All rights reserved.

1. Introduction

Real-world data is commonly affected by noise [1,2]. The building time, complexity and, particularly, the performance of the model, are usually deteriorated by noise in classification problems [3–5]. Several learners, e.g., C4.5 [6], are designed taking these problems into account and incorporate mechanisms to reduce the negative effects of noise. However, many other methods ignore these issues. Among them, instance-based learners, such as *k*-nearest neighbors (*k*-NN) [7–9], are known to be very sensitive to noisy data [10,11].

In order to improve the classification performance of noise-sensitive methods when dealing with noisy data, noise filters [12–14] are commonly applied. Their aim is to remove potentially noisy examples before building the classifier. However, both correct examples and examples containing valuable information can also be removed. This fact implies that these techniques do not always provide an improvement in performance. As indicated by Wu and Zhu [1], the success of these methods depends on several circumstances, such as the kind and nature of the data errors, the quantity of noise removed or the capabilities of the classifier to deal with the loss of useful information related to the filtering. Therefore, the efficacy of noise filters, i.e., whether their usage causes an improvement in classifier performance, depends on the noise-robustness and the generalization capabilities of the classifier used, but it also strongly depends on the characteristics of the data.

Data complexity measures [15] are a recent proposal to represent characteristics of the data which are considered difficult in classification tasks, e.g., the overlapping among classes, their separability or the linearity of the decision boundaries.

This paper proposes the computation of these data complexity measures to predict in advance when the usage of a noise filter will statistically improve the results of a noise-sensitive learner: the nearest neighbor classifier (1-NN). This prediction can help, for example, to determine an appropriate noise filter for a concrete noisy dataset – that filter providing a significant advantage in terms of the results – or to design new noise filters which select more or less aggressive filtering strategies considering the characteristics of the data. Choosing a noise-sensitive learner facilitates the checking of when a filter removes the appropriate noisy examples in contrast to a robust learner—the performance of classifiers built by the former is more sensitive to noisy examples retained in the dataset after the filtering process. In addition, this paper has the following objectives:

1. To analyze the relation between the characteristics of the data and the efficacy of several noise filters.
2. To find a reduced set of the most appropriate data complexity measures for predicting the noise filtering efficacy.

* Corresponding author. Tel.: +34 958 240598; fax: +34 958 243317.
E-mail addresses: smja@decsai.ugr.es, tschigorine@gmail.com (J.A. Sáez), jluengo@ubu.es (J. Luengo), herrera@decsai.ugr.es (F. Herrera).

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2012.07.009

3. Even though each noise filter may depend on concrete characteristics of the data to work correctly, it would be interesting to identify common characteristics of the data under which most of the noise filters work properly.
4. To provide a set of interpretable rules which a practitioner can use to determine whether to use a noise filter with a classification dataset.

A web page with the complementary material of this paper is available at http://sci2s.ugr.es/filtering-efficacy. It includes the details of the experimentation, the datasets used, the performance results of the noise filters and the distribution of the data complexity metrics of the datasets.

The rest of this paper is organized as follows. Section 2 presents data complexity measures. Section 3 introduces the noise filters and enumerates those considered in this paper. Section 4 describes the method employed to extract the rules predicting the noise filtering efficacy. Section 5 shows the experimental study performed and the analysis of results. Finally, Section 6 enumerates some concluding remarks.

2. Data complexity measures

In this section, first a brief review of recent studies on data complexity metrics is presented (Section 2.1). Then, the measures of overlapping (Section 2.2), the measures of separability of classes (Section 2.3) and the measures of geometry (Section 2.4) used in this paper are described.

2.1. Recent studies on data complexity

There are some methods used in classification, either learner or preprocessing techniques, which work well with concrete datasets, while other techniques work better with different ones. This is due to the fact that each classification dataset has particular characteristics that define it. Issues such as the generality of the data, the inter-relationships among the variables and other factors are key for the results of such methods. An emergent field proposes the usage of a set of data complexity measures to quantify these particular sources of the problem on which the behavior of classification methods usually depends [15].

A seminal work on data complexity is [16], in which some complexity measures for binary classification problems are proposed, gathering metrics of three types: overlaps in feature values from different classes; separability of classes; and measures of geometry, topology and density of manifolds. Extensions can also be found in the literature, such as in the work of Singh [17], which offers a review of data complexity measures and proposes two new ones.

From these works, different authors attempt to address different data mining problems using these measures. For example, Baumgartner and Somorjai [18] define specialized measures for regularized linear classifiers. Other authors try to explain the behavior of learning algorithms using these measures, optimizing the decision tree creation in the binarization of datasets [19] or to analyze fuzzy-UCS and the model obtained when applied to data streams [20]. The data complexity measures have been referred to other related fields, such as gene expression analysis in Bioinformatics [21,22].

The research efforts in data complexity are currently focused on two fronts. The first aims to establish suitable problems for a given classification algorithm, using only the data characteristics, and thus determining their domains of competence. In this line of research recent publications, e.g., the works of Luengo and Herrera [23] and Bernadó-Mansilla and Ho [24], provide a first insight into the determination of an individual classifier's domains of competence. Parallel to this, Sánchez et al. [25] study the effect of data complexity on the nearest neighbor classifier. The relationships between the domains of competence of similar classifiers were analyzed by Luengo and Herrera [26], indicating that related classifiers benefit from common sources of complexity of the data.

Data complexity measures are increasingly used in order to characterize when a preprocessing stage will be beneficial to a subsequent classification algorithm in many challenging domains. García et al. [27] firstly analyzed the behavior of the evolutionary prototype selection strategy using one complexity measure based on overlapping. Further developments resulted in a characterization of when the preprocessing in imbalanced datasets is beneficial [28]. The data complexity measures can also be used online in the data preparation step. An example of this is the work of Dong [29], in which a feature selection algorithm based on complexity measures is proposed.

This paper follows the second research line. It aims to characterize when a filtering process is beneficial using the information provided by the data complexity measures. Noise will affect the geometry of the dataset, and thus the values of the data complexity metrics. It can be expected that such metrics will enable one to know in advance whether noise filters will be useful for the given dataset.

In this study, 11 of the metrics proposed by Ho and Basu [16] will be analyzed. In the following subsections, these measures, classified by their family, are briefly presented. For a deeper description of their characteristics, the reader may consult [16].

2.2. Measures of class overlapping

These measures focus on the effectiveness of a single feature dimension in separating the classes, or the composite effects of a number of dimensions. They examine the range and spread of values in the dataset within each class and check for overlapping among different classes.

- F1—*maximum Fisher's discriminant ratio*: This is the value of Fisher's discriminant ratio of the attribute that enables one to better discriminate between the two classes, computed as

$$\mathrm{F1} = \max_{i=1,\ldots,d} \frac{(\mu_{i,1}-\mu_{i,2})^2}{\sigma_{i,1}^2+\sigma_{i,2}^2} \tag{1}$$

  where d is the number of attributes, and $\mu_{i,j}$ and $\sigma_{i,j}^2$ are the mean and variance of the attribute i in the class j, respectively.
- F2—*volume of the overlapping region*: This measures the amount of overlapping of the bounding boxes of the two classes. Let $\max(f_i,C_j)$ and $\min(f_i,C_j)$ be the maximum and minimum values of the feature f_i in the set of examples of class C_j, let minmax_i be the minimum of $\max(f_i,C_j),(j=1,2)$ and maxmin_i be the maximum of $\min(f_i,C_j),(j=1,2)$ of the feature f_i. Then, the measure is defined as

$$\mathrm{F2} = \prod_{i=1\ldots d} \frac{\mathrm{minmax}_i-\mathrm{maxmin}_i}{\max(f_i,C_1\cup C_2)-\min(f_i,C_1\cup C_2)} \tag{2}$$

- F3—*maximum feature efficiency*: This is the maximum fraction of points distinguishable with only one feature after removing unambiguous points falling outside of the overlapping region in this feature [30].

2.3. Measures of separability of classes

These give indirect characterizations of class separability. They assume that a class is made up of single or multiple manifolds

Download English Version:

https://daneshyari.com/en/article/530614

Download Persian Version:

https://daneshyari.com/article/530614

Daneshyari.com