Pattern Recognition 70 (2017) 126–138

Contents lists available at ScienceDirect

Pattern Recognition

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

Feature selection for regression problems based on the Morisita estimator of intrinsic dimension

Jean Golay*, Michael Leuenberger, Mikhail Kanevski

Institute of Earth Surface Dynamics, Faculty of Geosciences and Environment, University of Lausanne, 1015 Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 22 March 2016
Revised 8 May 2017
Accepted 9 May 2017
Available online 10 May 2017

Keywords:
Feature selection
Intrinsic dimension
Morisita index
Measure of relevance
Data mining

ABSTRACT

Data acquisition, storage and management have been improved, while the key factors of many phenomena are not well known. Consequently, irrelevant and redundant features artificially increase the size of datasets, which complicates learning tasks, such as regression. To address this problem, feature selection methods have been proposed. This paper introduces a new supervised filter based on the Morisita estimator of intrinsic dimension. It can identify relevant features and distinguish between redundant and irrelevant information. Besides, it offers a clear graphical representation of the results, and it can be easily implemented in different programming languages. Comprehensive numerical experiments are conducted using simulated datasets characterized by different levels of complexity, sample size and noise. The suggested algorithm is also successfully tested on a selection of real world applications and compared with RReliefF using extreme learning machine. In addition, a new measure of feature relevance is presented and discussed.

1. Introduction

In data mining, it is often not known a priori what features (or input variables[1]) are truly necessary to capture the main characteristics of a studied phenomenon. This lack of knowledge implies that many of the considered features are irrelevant or redundant. They artificially increase the dimension E of the Euclidean space $\mathbb{R}^E$ in which the data points are embedded (E equals the number of input and output variables under consideration). This is a serious matter, since fast improvements in data acquisition, storage and management cause the number of redundant and irrelevant features to increase. As a consequence, the interpretation of the results becomes more complicated and, unless the sample size N grows exponentially with E, the curse of dimensionality [1] may reduce the overall accuracy yielded by any learning algorithm. Besides, large N and E are also difficult to deal with because of computer performance limitations.

In regression and classification, these issues are often addressed by implementing supervised feature selection methods [2–5]. Such methods can be broadly subdivided into filter (e.g. RReliefF [6], mRMR [7] and CFS [8]), wrapper [9,10] and embedded methods (e.g. the Lasso [11] and random forest [12]). Filters rank features, or subsets of features, according to a relevance measure independently of any predictive model, while wrappers use an evaluation criterion involving a learning machine. Both approaches can be used with search strategies, since an exhaustive exploration of the $2^{\#Feat.}-1$ models (all the possible combinations of features) is often computationally intractable. Greedy strategies [13,14], such as Sequential Forward Selection (SFS) [15], can be distinguished from stochastic ones (e.g. simulated annealing [16,17] and ant colony optimization [18,19]). Regarding the embedded methods, the feature selection is a by-product of a training procedure. It can be achieved by the addition of constraints in the cost function of a predictive model (e.g. the Lasso [11]), or it can be more specific to a given algorithm (e.g. random forest [12] and adaptive general regression neural networks [20]).

The present paper[2] deals with a new SFS filter algorithm. It relies on the idea that, although data points are embedded in E-dimensional spaces, they often reside on lower M-dimensional manifolds [22–24]. The value M ($\leq E$) is called Intrinsic Dimension (ID), and it can be estimated using the Morisita estimator of ID [25] which is closely related to the fractal theory. The proposed filter algorithm is supervised, designed for regression problems and based on this new ID estimator. It also keeps the simplicity of the Fractal Dimension Reduction (FDR) algorithm introduced in [26].

* Corresponding author.
E-mail address: jean.golay@unil.ch (J. Golay).

[1] In this paper, the term "feature" is used as a synonym for "input variable".

[2] The main idea of this paper was partly presented at the 23rd symposium on artificial neural networks, computational intelligence and machine learning (ESANN2015) [21].

http://dx.doi.org/10.1016/j.patcog.2017.05.008
0031-3203/

Finally, the results show the ability of the new filter to capture non-linear relationships and to effectively identify both redundant and irrelevant information.

The paper is organized as follows. Section 2 reviews previous work on ID-based feature selection approaches. The Morisita estimator of ID is shortly presented in Section 3 (for the completeness of the paper). Section 4 introduces the Morisita-based filter, and Section 5 is devoted to numerical experiments conducted on simulated data of varying complexity. In Section 6, real world applications from publicly accessible repositories are presented, and a comparison with a benchmark algorithm, RReliefF [6], is carried out using Extreme Learning Machine (ELM) [27]. Finally, conclusions are drawn in the last section with a special emphasis on future challenges and applications.

2. Related work

The concept of ID can be extended to the more general case where the data ID may be a non-integer dimension D [23,26,28]. The value D is estimated by using fractal-based methods which have been presented in [23,24,29] and successfully implemented in various fields, such as physics [30], cosmology [31], meteorology [32] and pattern recognition [33,34]. These methods rely on well-known fractal dimensions (e.g. the box-counting dimension [35,36], the correlation dimension [30] and Rényi's dimensions of qth order [37]), and they can be used in feature selection [26,38] and dimensionality reduction [23] to detect dependencies between variables (or features).

Traina et al. [26,39] have opened up new prospects for the effective use of ID estimation in data mining by introducing the Fractal Dimension Reduction (FDR) algorithm. FDR executes an unsupervised procedure of feature selection aiming to remove from a dataset all the redundant variables. The fundamental idea is that fully redundant variables do not contribute to the value of the data ID.

This idea can be illustrated by sampling two uniformly distributed variables V_1 and V_2. If they are independent, which means that they are not redundant, one has that:

$$ID(V_1, V_2) \approx ID(V_1) + ID(V_2) \approx 1 + 1 = 2 \tag{1}$$

where $ID(\cdot)$ denotes the ID of a dataset. It indicates that both V_1 and V_2 contribute to increasing the value of $ID(V_1, V_2)$ by about 1, which is, by construction, equal to the ID of each variable (i.e. $ID(V_1)$ and $ID(V_2)$). Conversely, the removal of either V_1 or V_2 would lead to a reduction in the data ID from about 2 (i.e. the dimension of the data space) to 1 (i.e. the ID of a single variable) and information would be irreparably lost. In contrast, if V_1 and V_2 are fully redundant with each other (e.g. $V_2 = V_1$), one has that:

$$ID(V_1, V_2) \approx ID(V_1) \approx ID(V_2) \approx 1 \tag{2}$$

where the ID of the full dataset is approximately equal to the topological dimension of a smooth line. This means that the contribution of only one variable is enough to reach the value of $ID(V_1, V_2)$ and the remaining one can be disregarded without losing any information.

Based on these considerations, the FDR algorithm removes the redundant variables from a dataset by implementing a Sequential Backward Elimination (SBE) strategy [13]. Besides, it uses Rényi's dimension of order $q = 2$, D_2, for the ID estimation. Following the same principles, De Sousa et al. [40] examined additional developments to FDR and presented a new algorithm for identifying subgroups of correlated variables.

FDR is designed to carry out unsupervised tasks, and it is not able to distinguish between variables that are relevant to a learning process and those that are irrelevant. The reason is that such variables can all contribute to the data ID. For instance, in Eq. (1), V_1 could be regarded as irrelevant to the learning of V_2, but it would be selected by FDR because it makes the data ID increase by about 1. Consequently, different studies were carried out to adapt FDR to supervised learning. Lee et al. [41] suggested decoupling the relevance and redundancy analysis. Following the same idea, Pham et al. [42] used mutual information to identify irrelevant features and combined the results with those of FDR. Finally, Mo and Huang [38] developed an advanced algorithm to detect both redundant and irrelevant information in a single step. Their algorithm follows a SBE search strategy and relies on the correlation dimension, df_{cor}, for the estimation of the data ID.

The filter algorithm suggested in the present paper is designed in such a way that it combines the advantages of both FDR and Mo's algorithm: it can deal with non-linear dependencies, it does not rely on any user-defined threshold, it can discriminate between redundant and irrelevant information, and the results can be easily summarized in informative plots. Moreover, it can cope with high-dimensional datasets thanks to its SFS search strategy, and it uses the Morisita estimator of ID which was shown to yield comparable or better results than D_2 and df_{cor} [25].

3. The Morisita estimator of intrinsic dimension

The Morisita estimator of ID, M_m, has been recently introduced [25]. It is a fractal-based ID estimator derived from the multipoint Morisita index $I_{m,\delta}$ [29,43] (named after Masaaki Morisita who proposed the first version of the index to study the spatial clustering of ecological data [44]). $I_{m,\delta}$ is computed by superimposing an E-dimensional grid of Q quadrats (i.e. cells) of diagonal size δ onto the data points. It measures how many times more likely it is that m ($m \geq 2$) randomly selected points will be from the same quadrat than it would be if all the N points of the studied dataset were distributed at random (i.e. according to a random distribution generated from a Poisson process). The formula is the following:

$$I_{m,\delta} = Q^{m-1} \frac{\sum_{i=1}^{Q} n_i(n_i-1)(n_i-2)\cdots(n_i-m+1)}{N(N-1)(N-2)\cdots(N-m+1)} \tag{3}$$

where n_i is the number of points in the ith quadrat. For a fixed value of m, $I_{m,\delta}$ is calculated for several values of δ on a chosen scale range. If a dataset approximates a fractal behaviour (i.e. is self-similar) within this range, the relationship of the plot relating $\log(I_{m,\delta})$ to $\log(1/\delta)$ is linear, and the slope of the regression line is defined as the Morisita slope S_m. Finally, M_m is expressed as:

$$M_m = E - \left(\frac{S_m}{m-1}\right). \tag{4}$$

In practice, each variable is rescaled to the [0, 1] interval (so is the grid), and δ can be replaced with the quadrat edge length ℓ, with ℓ^{-1} being simply the number of quadrats along each axis of the data space. Then a set of R values of ℓ (or ℓ^{-1}) is chosen so that it captures the linear part of the log-log plot. In the rest of this paper, only $M_{m=2}$ will be used, and it will be computed with an algorithm called Morisita INDex for Intrinsic Dimension estimation (MINDID) [25] whose complexity is $\mathcal{O}(N * E * R)$.

4. The Morisita-based filter for regression problems

The Morisita-Based Filter for Regression problems (MBFR) relies on three observations following from the work by Traina et al. [26], De Sousa et al. [40] and Mo and Huang [38]:

1. Given an output variable Y generated from k relevant and non-redundant input variables $X_1, \ldots, X_k$, one has that:

$$ID(X_1, \ldots, X_k, Y) - ID(X_1, \ldots, X_k) \approx 0 \tag{5}$$

where $ID(\cdot)$ denotes the (possibly non-integer) ID of a dataset.