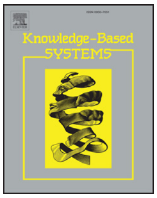
Contents lists available at ScienceDirect

Knowledge-Based Systems

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



Unsupervised feature selection based on the Morisita estimator of intrinsic dimension

Jean Golay*, Mikhail Kanevski

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

ARTICLE INFO

Article history:
Received 14 December 2016
Revised 12 July 2017
Accepted 8 August 2017
Available online xxx

Keywords:
Unsupervised feature selection
Morisita index
Intrinsic dimension
Redundancy minimization
Data mining

ABSTRACT

This paper deals with a new filter algorithm for selecting the smallest subset of features carrying all the information content of a dataset (i.e. for removing redundant features). It is an advanced version of the fractal dimension reduction technique, and it relies on the recently introduced Morisita estimator of Intrinsic Dimension (ID). Here, the ID is used to quantify dependencies between subsets of features, which allows the effective processing of highly non-linear data. The proposed algorithm is successfully tested on simulated and real world case studies. Different levels of sample size and noise are examined along with the variability of the results. In addition, a comprehensive procedure based on random forests shows that the data dimensionality is significantly reduced by the algorithm without loss of relevant information. And finally, comparisons with benchmark feature selection techniques demonstrate the promising performance of this new filter.

1. Introduction

Recent breakthroughs in technology have radically improved our ability to collect and store data. Consequently, more and more variables (or features[1]) are available to perform data mining tasks, but in general, a lot of them are redundant (i.e. they do not carry additional information beyond that subsumed by other features), or partially redundant, and contribute to the emergence of four major issues: (1) the reduction in the accuracy of learning algorithms because of the curse of dimensionality [1], (2) the computer performance limitations related to memory and processing speed, (3) the difficulty in visualizing large amounts of complex and high-dimensional data and (4) the interpretability of the results which becomes less tractable making it difficult to gain an insight into the mechanisms that generated the data.

Due mainly to these redundant and partially redundant features, data points do not occupy the full E-dimensional space $\mathbb{R}^E$ (E is the number of features in a dataset) in which they are embedded. Instead, they are often regarded as residing on a lower M-dimensional manifold where $M(\leq E)$ is the *Intrinsic Dimension* (ID) of data [2]. Dimensionality Reduction (DR) methods [3,4] can help remove redundant information by trying to map the original data space coordinates to an intrinsic coordinate system of dimensionality M. Depending on the assumptions made about the shape of the manifold, the mapping can be either linear (e.g. PCA [5]) or non-linear (e.g. kernel-PCA [6]), and a great advantage of the DR approach is its potential to capture complex dependencies. On the other hand, DR often leads to a deterioration in the physical interpretability of the data and to difficulties in the understanding of subsequent results. A possible solution to these drawbacks is the implementation of feature selection methods.

The goal of *feature selection* [7–11] is to select the smallest subset of original features which maintains some meaningful characteristics with respect to a chosen criterion. According to the possible use of output information (e.g. class labels), feature selection methods can be broadly classified as either supervised or unsupervised. Advanced supervised methods aim to select features which are both relevant to the prediction (i.e. classification or regression) of some output information and related as little as possible to one another (i.e. select relevant and non-redundant features). In contrast, unsupervised methods do not make use of any a priori knowledge regarding an output, and they can be further divided into two categories: Cluster Recognition (CR) and Redundancy Minimization (RM).

The CR methods aim to find the smallest subset of features that uncovers the most “interesting” and “natural” groupings (i.e. clusters) of data points [12–15]. They rely on criteria of relevance that do not involve any output information, and they can be categorized into filters and wrappers [16]. The former (e.g. the Laplacian score method [17], SPEC [18,19] and MCFS [20]) do not incorpo-

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

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

http://dx.doi.org/10.1016/j.knosys.2017.08.009
0950-7051/© 2017 Elsevier B.V. All rights reserved.

rate the clustering algorithm that will ultimately be applied, while the latter do (e.g. methods introduced in [12,21] or reviewed in [13]). In contrast, the RM methods are often not restricted to clustering problems, and they can be used as preprocessing tools in a wide variety of data mining approaches. Their goal is to select the smallest subset of features in such a way that all the information content of a dataset is preserved as much as possible. In other words, they aim to eliminate all the redundant information by selecting the most informative features (i.e. the non-redundant features). To achieve this goal, the RM methods often use criteria based on PCA loading values [22] or on measures of feature dependency, such as the maximal information compression index [23], mutual information [24,25] and fractal-based measures of ID [26–28]. More recently, Wang et al. [29] proposed a criterion that minimizes the reconstruction error of a linear projection of the original features, while ensuring low redundancy, and Zhu et al. [30] followed a similar objective by introducing the concept of self-representation for unsupervised feature selection. Further, the RM methods can be thought of as filters, and like many other methods of feature selection, they can rely on greedy (e.g. Sequential Forward Selection (SFS) [31] and Sequential Backward Elimination (SBE) [32,33]) or randomized (e.g. simulated annealing [34]) search strategies if they consider multivariate interactions and aim to find the best subset among the $2^E - 1$ combinations of features. Lastly, methods combining the CR and the RM approaches have also been developed. Many of them use a graph Laplacian matrix to preserve the data structure and involve a low redundancy constraint or a more advanced regularization term [35]. And the presence of noise and outliers motivated the work by Qian et al. [36] which proposed a framework to carry out both robust clustering and robust feature selection.

More specifically, the use of ID for unsupervised feature selection was introduced by Traina et al. [26,37]. They extended the concept of ID to fractal dimensions and proposed the Fractal Dimension Reduction (FDR) algorithm. FDR is a filter algorithm for non-linear RM that follows a SBE search strategy. It aims to eliminate the features which do not contribute to increasing the value of the data ID (i.e. the ID of the studied dataset), and it relies on Rényi's dimension of order 2 [38], D_2, for the ID estimation. An extension to FDR was proposed by De Sousa et al. [27] to identify subsets of correlated attributes in databases according to user-defined levels of correlation. Finally, Mo and Huang [28] modified FDR by replacing D_2 with the correlation dimension df_{cor} [39].

The present paper deals with a novel ID-based filter algorithm for RM. It relies on the recently introduced Morisita estimator of ID, M_m, which was shown to be more effective than D_2 and df_{cor} in situations where the data points were sparsely distributed [40]. Besides, the proposed algorithm follows a SFS search strategy; it can process large and highly non-linear data, and its implementation is straightforward in R and MATLAB. Another advantage is that the number of features to be selected can be determined directly from the results. And it is also worth mentioning that M_m was already used successfully to perform supervised feature selection in regression problems [41].

The remainder of this paper is organized as follows. Section 2 presents the Morisita estimator of ID, and Section 3 explains the relationship between ID and data redundancy. In Section 4, the proposed algorithm for RM is introduced, and Section 5 is devoted to numerical experiments conducted on simulated data and on real world case studies. The quality of the results is assessed using a comprehensive methodology based on random forests [42], and comparisons with benchmark feature selection techniques (including FDR) are also discussed. Finally, conclusions are drawn in the last section with a special emphasis on potentialities and future challenges.

2. The Morisita estimator of intrinsic dimension

2.1. Overview

The Morisita estimator of ID [40], M_m, is derived from the multipoint Morisita index $I_{m,\,\delta}$ [44–46]. $I_{m,\,\delta}$ is computed by means of an E-dimensional grid of Q cells (or quadrats) of diagonal size δ superimposed over the data points (see Fig. 1). It measures how many times more likely it is that m ($m \geq 2$) points selected at random will be from the same cell than it would be if the N points of the studied dataset were distributed according to a random distribution generated from a Poisson process (i.e. complete spatial randomness). $I_{m,\,\delta}$ is given by the following formula:

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

where n_i is the number of data points in the ith cell. In general, m is set to 2, and the computation of the index is iterated for R different values of δ. These values must be chosen by the user and determine the scales at which the phenomenon will be characterized. Within the range of these scale values, if the dataset follows a fractal behaviour (i.e. is self-similar), the functional relationship between $\log(I_{m,\,\delta})$ and $\log(1/\delta)$ is linear, its slope, S_m, is the Morisita slope, and M_m can be written as:

$$M_m = E - \left(\frac{S_m}{m-1}\right). \qquad (2)$$

In practice, each feature is rescaled to the [0, 1] interval (so is the grid), and δ is replaced with the edge length, ℓ, of the cells. In this context, ℓ^{-1} is simply the number of cells along each axis of the E-dimensional space where the data points are embedded.

2.2. Detailed procedure

In the remainder of this paper, the Morisita estimator of ID will be used only with $m = 2$ as advocated in [40]. The following steps summarize how to compute the ID of a dataset using $M_{m=2}$:

1. Rescale each of the E features to the [0, 1] interval.
2. Choose the values of the parameter ℓ^{-1} so that the functional relationship of Step 6 can be well approximated by a linear regression model (see Section 2.3).
3. Superimpose an E-dimensional grid over the data points. The size of the grid cells is controlled by the user through the parameter ℓ^{-1} which is simply the number of cells along each axis of the grid.
4. Count the number of data points falling into the cells of the grid. This step must be repeated for each value of the parameter ℓ^{-1} chosen by the user.
5. Compute the multipoint Morisita index $I_{m=2,\ell^{-1}}$ for each value of the parameter ℓ^{-1} using Eq. (1). Notice that the values of $I_{m=2,\ell^{-1}}$ are equal to those of $I_{m=2,\delta}$, since δ and ℓ^{-1} are two different ways of characterizing the size of the same cells.
6. Carry out the linear regression of $\log(I_{m=2,\ell^{-1}})$ on $\log(\ell^{-1})$. Then $S_{m=2}$ is simply the slope of the regression model.
7. Compute $M_{m=2}$ using Eq. (2).

The procedure is illustrated in Fig. 1 for $E = 2$. On the right, the two features F_1 and F_2 have been rescaled to the [0, 1] interval and a 2-dimensional grid is superimposed over the data points. The number of cells along each of the two axes of the grid is equal to 4. This is the value of the parameter ℓ^{-1} which allows the user to control the grid resolution. The calculation of $I_{m=2,\ell^{-1}}$ was iterated four times ($R = 4$) for $\ell^{-1} \in \{1, 2, 3, 4\}$, and the results were used to draw the log-log plot shown on the left of the figure. The dashed line represents the linear regression model of Step 6. Its slope is the Morisita slope S_2.