Pattern Recognition ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Pattern Recognition

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

A new estimator of intrinsic dimension based on the multipoint Morisita index

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 8 August 2014
Received in revised form
20 May 2015
Accepted 20 June 2015

Keywords:
Intrinsic dimension
Multipoint Morisita index
Fractal dimension
Multifractality
Dimensionality reduction

ABSTRACT

The size of datasets has been increasing rapidly both in terms of number of variables and number of events. As a result, the empty space phenomenon and the curse of dimensionality complicate the extraction of useful information. But, in general, data lie on non-linear manifolds of much lower dimension than that of the spaces in which they are embedded. In many pattern recognition tasks, learning these manifolds is a key issue and it requires the knowledge of their true *intrinsic dimension*. This paper introduces a new estimator of intrinsic dimension based on the multipoint Morisita index. It is applied to both synthetic and real datasets of varying complexities and comparisons with other existing estimators are carried out. The proposed estimator turns out to be fairly robust to sample size and noise, unaffected by edge effects, able to handle large datasets and computationally efficient.

1. Introduction

The 21st century is more and more data-dependent and, in general, when collecting data for a particular purpose, it is not known which variables matter the most. This lack of knowledge leads to the emergence of high-dimensional datasets characterized by redundant features which artificially increase the volume of data to be processed. As a result, the empty space phenomenon [1] and the curse of dimensionality [2] make it challenging to conduct pattern recognition tasks such as clustering and classification.

The goal of dimensionality reduction (DR) [3,4], sometimes called manifold learning, is to address this issue by mapping the N sampled data points into the lower dimensional space where they truly lie. Such a space is often considered as a manifold of intrinsic dimension M[1] embedded in a Euclidean space of dimension E with $E \geq M$. E equals the number of variables of a dataset and the intrinsic dimension (ID) of a manifold is equal to the theoretical ID of the data. If a manifold is space-filling, its dimension $M \approx E$. In contrast, if the Euclidean space is partially empty, $M < E$. The optimality of DR greatly depends on the accuracy of ID estimates. An underestimation of the theoretical ID will result in the implosion of the data manifold and information will be irreparably lost. On the contrary, an overestimation will lead to noise in the final mapping. From an application perspective, DR can be used to produce low dimensional syntheses of high dimensional datasets [5] and as a preprocessing tool for supervised learning [6,7] and data visualization [8].

DR methods perform variable transformations to capture the complex dependencies which generate redundancy within datasets. Nevertheless, it is often important not to recast data. The fractal dimension reduction (FDR) algorithm [9–12] was designed to this end. The fundamental idea is to remove from a dataset all the variables which do not contribute to increasing its ID. FDR can also be adapted to supervised feature selection methods [13]. The goal is then to reject irrelevant or redundant variables (or features) according to a prediction task (i.e. regression or classification). Although ID estimation lies at the core of FDR, more traditional unsupervised [14–16] and supervised [17–22] feature selection methods do not consider it. It has, however, a great potential in speeding up search strategies, such as those used in [23–26].

These different approaches highlight that ID estimation is a fundamental problem when dealing with high-dimensional datasets. Unfortunately, ID estimators [27,28] suffer from the curse of dimensionality as well. Their overall performance depends on many factors (to various degree), such as the number of data points, the theoretical ID of data and the shape of manifolds. The present research deals with a new ID estimator in order to provide a solution to the problems raised by these factors. It is based on the recently introduced multipoint Morisita index (m-Morisita) [29–31]. The m-Morisita index is a measure of global clustering closely related to the concept of multifractality and, so far, it has been successfully applied within the framework of (2-dimensional) spatial data analysis [29,30].

The paper is organized as follows: In Section 2, traditional fractal-based and maximum likelihood methods of ID estimation

* Corresponding author. Tel.: +41 21 692 35 41.
E-mail address: jean.golay@unil.ch (J. Golay).

[1] In Physics, mainly, M is often referred to as the degrees of freedom of data.

http://dx.doi.org/10.1016/j.patcog.2015.06.010
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

are presented. Section 3 derives a new ID estimator from the m-Morisita index and introduces a new algorithm for its application to high-dimensional datasets. Section 4 is devoted to comparisons between the proposed estimator and those of Section 3. Their behaviour regarding sample sizes, noise and the dimension of manifolds is analysed. A special attention is also paid to their bias and variance by using Monte-Carlo simulations and real world case studies from the UCI machine learning repository are examined. Finally, conclusions are drawn in Section 5.

2. Existing methods

Many ID estimation methods have been proposed [27,28,32–34] and they can be roughly divided into projection (e.g. PCA) and geometric methods (e.g. fractal, nearest-neighbour and maximum likelihood methods). This section focuses on fractal-based and maximum likelihood estimators. They are commonly used in a wide range of applications and they generally provide non-integer values as ID estimates.

2.1. Fractal-based estimation methods

The word *fractal* was first coined by Mandelbrot [35] to describe scale-invariant sets. At small scales δ, for a given point pattern, one has that

$$n_{box}(\delta) \propto \delta^{-D_0} \tag{1}$$

where $n_{box}(\delta)$ is the number of grid cells necessary to cover the whole pattern and D_0 is known as the box-counting dimension [35–37]. In practical applications, due to its simplicity, D_0 often replaces the Hausdorff dimension D (or fractal dimension) and it can be proved that D_0 is an upper bound of D [38].

In complex cases, the scaling behaviour of the moments of point distributions cannot be fully characterized by only one fractal dimension and a full spectrum of generalized dimensions, D_q, is required. Such distributions are referred to as being multifractal [39–42]. D_q is generally obtained by using a generalization of the box-counting method [39–41,43] based on Rényi's information, $RI_q(\delta)$, of qth order [44]. The central scaling law of this approach can be written as follows for $q \neq 1$:

$$\exp(RI_q(\delta)) \propto \delta^{-D_q} \tag{2}$$

where

$$RI_q(\delta) = \frac{1}{1-q}\log\left(\sum_{i=1}^{n_{box}(\delta)} p_i(\delta)^q\right) \tag{3}$$

In this last equation, $p_i(\delta) = n_i/N$ is the value of the probability mass function in the ith grid cell of size δ (n_i is the number of points falling into the ith cell) and $q \in \mathbb{R}\backslash\{-1\}$. Finally, one has that

$$D_q = \lim_{\delta \to 0}\frac{RI_q(\delta)}{\log\left(\frac{1}{\delta}\right)} \tag{4}$$

and

$$\lim_{q\to 1} D_q = df_i \tag{5}$$

$$D_2 = df_{cor} \tag{6}$$

where df_i and df_{cor} are, respectively, the information dimension [45,39] and the correlation dimension [46].

Usually, df_{cor} is computed with the Grassberger–Procaccia (GP) algorithm [46]. This algorithm is designed to better take advantage of the range of available pairwise distances between points. It can be introduced as follows: at small scales, for a point set, $X_N = \{x_1, \ldots, x_N\}$, one has that

$$C(\delta) \propto \delta^{df_{cor}} \tag{7}$$

where

$$C(\delta) = \frac{2}{N(N-1)}\sum_{i=1}^{N}\sum_{j=i+1}^{N} \mathbb{1}_{\{\|x_i - x_j\| \le \delta\}} \tag{8}$$

with $\mathbb{1}$ being an indicator function and df_{cor} can be expressed as

$$df_{cor} = D_2 = \lim_{\delta\to 0}\frac{\log(C(\delta))}{\log(\delta)} \tag{9}$$

The available values of $RI_q(\delta)$ and $\log(C(\delta))$ depend on the data resolution. A commonly used method for estimating D_q consists in plotting $RI_q(\delta)$ *vs* $\log(\delta^{-1})$ for a chosen scale interval. The final estimate is then the slope of the linear regression fitting the linear part of the resulting chart. The procedure is the same for the GP algorithm, except that df_{cor} and $\log(C(\delta))$ replace, respectively, D_q and $RI_q(\delta)$. Eventually, both D_q (in general $0 \le q \le 2$) and df_{cor} can be used as ID estimators.

Although these methods may entail some disadvantages due to the finiteness of datasets [33], they have been successfully applied in various fields, such as spatial [47,48] and time series [49] analysis, cosmology [50], climatology [36,51,52] and pattern recognition [53,54]. They have also been used in different procedures improving their overall performance [55].

2.2. Maximum likelihood estimation methods

The maximum likelihood estimation (MLE) of ID was introduced in [28]. The proposed method relies on the assumption that the k-nearest neighbours (k-NN) of any point x_i of a point set $X_N = \{x_1, \ldots, x_N\}$ are stemming from a uniform probability density function $f(x_i)$. As a consequence, for a fixed x_i, the observations are treated as a homogeneous Poisson process within a small sphere $S_{x_i}(R)$ of radius R centred at x_i. On this basis, the inhomogeneous binomial process $\{N(t, x_i), 0 \le t \le R\}$ with

$$N(t, x_i) = \sum_{j=1}^{N} \mathbb{1}_{\{x_j \in S_{x_i}(t)\}} \tag{10}$$

counts the number of observations within the distance t of x_i and can be approximated as a Poisson process. The rate of this process is

$$\lambda(t, x_i) = f(x_i)V(m(x_i))m(x_i)t^{m(x_i)-1} \tag{11}$$

where $m(x_i)$ is the dimension of the manifold on which x_i lies and $V(m(x_i))$ is the volume of the unit sphere in $\mathbb{R}^{m(x_i)}$ centred at x_i. The log-likelihood function of $N(t, x_i)$ can then be expressed as

$$L(m(x_i), \theta(x_i)) = \int_0^R \log(\lambda(t, x_i))\, dN(t, x_i) - \int_0^R \lambda(t, x_i)\, dt \tag{12}$$

where $\theta(x_i) = \log(f(x_i))$. Finally, the MLE for $m(x_i)$ provides a local estimator of ID [28,56,57]:

$$\hat{m}_k(x_i) = \left[\frac{1}{k-2}\sum_{j=1}^{k-1}\log\left(\frac{T_k(x_i)}{T_j(x_i)}\right)\right]^{-1} \tag{13}$$

where $k > 2$ is the number of NN taken into account and $T_k(x_i)$ is the distance between x_i and its kth NN.

If it is assumed that all the observations belong to the same manifold, one has that

$$\hat{m}_k = \frac{1}{N}\sum_{i=1}^{N}\hat{m}_k(x_i) \tag{14}$$

which is simply an average over the whole dataset and, for $k \in \{k_1, k_1+1, \ldots, k_2\}$ with $k_1 > 2$, the final estimate of ID is

Please cite this article as: J. Golay, M. Kanevski, A new estimator of intrinsic dimension based on the multipoint Morisita index, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.06.010