Information Sciences 328 (2016) 26–41

Contents lists available at ScienceDirect

Information Sciences

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

Intrinsic dimension estimation: Advances and open problems

Francesco Camastra*, Antonino Staiano

Department of Science and Technology, University of Naples Parthenope, Centro Direzionale Isola C4 - 80143 Napoli, Italy

ARTICLE INFO

Article history:
Received 2 January 2014
Revised 1 July 2015
Accepted 8 August 2015
Available online 24 August 2015

Keywords:
Intrinsic dimension
Curse of dimensionality
Maximum likelihood
Correlation dimension
Dimensionality reduction

ABSTRACT

Dimensionality reduction methods are preprocessing techniques used for coping with high dimensionality. They have the aim of projecting the original data set of dimensionality N, without information loss, onto a lower M-dimensional submanifold. Since the value of M is unknown, techniques that allow knowing in advance the value of M, called intrinsic dimension (ID), are quite useful. The aim of the paper is to review state-of-the-art of the methods of ID estimation, underlining the recent advances and the open problems.

1. Introduction

Dimensionality reduction methods are preprocessing techniques used for coping with high dimensionality. Dimensionality reduction methods aim to project the original data set $\Omega \subset \mathbb{R}^N$, without information loss, onto a lower M-dimensional submanifold of $\mathbb{R}^N$. Since the value of M is unknown, techniques that provide the value of M in advance, are quite useful. Following Fukunaga [34], the minimum number of parameters required to account for the observed properties of data is the intrinsic (or effective) dimension of the data set. The estimation of the intrinsic dimension (ID) of a data set is a classical problem of pattern recognition and machine learning. The first algorithm of data dimensionality estimation, by Bennett, dates back to 1969 [4]. ID estimation is relevant in machine learning not only for dimensionality reduction methods but also for several other reasons. First, using more dimensions than necessary leads to several problems, such as an increase in the space required to store data, and a decrease in the algorithm speed, since it generally depends on data dimensionality. Besides, building reliable classifiers becomes harder and harder when the dimensionality grows (*curse of dimensionality*) [3]. To this purpose, we recall that the capacity (*VC-dimension*) [96] of the linear classifiers, that determines their generalization capability, depends on ID. Nearest neighbor searching algorithms can profit from a good ID estimate, since the complexity of search data structures (e.g., kd-trees and R-trees) increases exponentially with ID [15]. Finally, ID is relevant for some prototype-based clustering algorithms. For instance, in a trained Neural Gas, data density P and density of the neural gas weight vectors ρ are related by $\rho \propto P^{\mu}$, where μ, called *magnification factor* [16,98], depends on ID according to $\mu = \frac{ID}{ID+2}$. Although in the literature there are surveys on ID estimation [12,45], they are dated so that they cannot take into account recent advances in the field.

The aim of the paper is to make state-of-the-art of the methods of ID estimation, underlining the advances and the open problems. By extending the taxonomy proposed by Jain and Dubes [45], we group the algorithms for estimating ID in three disjoint categories, i.e., *local*, *global*, and *pointwise*. In the local category, there are algorithms that provide an ID estimation by using information contained in sample neighborhoods. The algorithms, belonging to the global category, make use of the whole

* Corresponding author. Tel.: +39- 3384447991.
E-mail addresses: camastra@ieee.org, francesco.camastra@uniparthenope.it (F. Camastra), antonino.staiano@uniparthenope.it (A. Staiano).

http://dx.doi.org/10.1016/j.ins.2015.08.029
0020-0255/

data set providing a unique and global ID estimate for the data set. Finally, in the pointwise category, there are the algorithms that can produce both a global ID estimate of the whole data set and local ID estimate of particular subsets of the data set. In the paper the most relevant algorithms for each category, underlining their weak points, will be presented. The paper is organized as follows: In Section 2 the Intrinsic Dimension is defined; Section 3 introduces the concept of ideal ID estimator, discussing the properties that it should have; Sections 4, 5, 6 describe global, local and pointwise methods, respectively; In Section 7 the main ID estimation methods are discussed under the ideal ID estimator framework; finally, in Section 8 open problems are analyzed and some conclusion are drawn.

2. Intrinsic dimension

In the Introduction we have informally introduced the concept of the intrinsic dimension saying that it is given by the number of parameters (or degrees of freedom) required to describe all data. A more formal definition of the intrinsic dimension is the following, due to [33]:

Definition 1. A data set $\Omega \subseteq \mathbb{R}^N$ is said to have *intrinsic dimension* (ID) equal to M if its elements lie entirely, without information loss, within a M-dimensional manifold of $\mathbb{R}^N$, where $M < N$.

Since most methods of the ID estimation are based on mathematical concepts, we briefly review the definition of dimension in the mathematical domain. The definition of dimension in mathematics is not univocal. The first mathematical definition of dimension (*Hausdorff dimension*) is due to Hausdorff [38]. The *Hausdorff dimension* m_H of a set Ω is defined by introducing the quantity $\Gamma_H^m(r) = \inf_{s_i} \sum_i (r_i)^m$, where the set Ω is covered by cells s_i with variable diameter r_i, and all diameters satisfy $r_i < r$. In other words, we look for that collection of covering sets s_i with diameters less than or equal to r which minimizes the sum and denote the minimized sum $\Gamma_H^m(r)$. The *m-dimensional Hausdorff measure* is defined as $\Gamma_H^m = \lim_{r\to 0} \Gamma_H^m(r)$. The m-dimensional Hausdorff measure generalizes the usual notion of the total length, area and volume of simple sets. Hausdorff proved that Γ_H^m, for every set Ω, is $+\infty$ if m is less than some critical value m_H and is 0 otherwise. The critical value m_H is called the *Hausdorff dimension* of the set.

A further definition of dimension, strictly related to Hausdorff's one, is the so-called *Information Dimension* [44]:

Definition 2. The information dimension $m_H(P)$, of a probability measure, P, is defined to be the smallest Hausdorff dimension of sets that have measure 1, i.e.,

$$m_H(P) = \inf_B \{m_H(B) : P(B) = 1\}. \tag{1}$$

Besides, it is possible to define a notion of dimension strictly related to a single data point, namely the so-called *pointwise dimension* [102]:

Definition 3. Let $B_r(\vec{x})$ be a closed ball of radius r and centre the data point $\vec{x} \in \mathbb{R}^N$, if P is a probability measure such that the limit

$$q = \lim_{r\to 0} \frac{\ln P(B_r(\vec{x}))}{\ln r} \tag{2}$$

exists, then the limit q is called the *pointwise* (or *local Hausdorff*) dimension.

3. Ideal ID estimator properties

Before describing the different methods for estimating ID, it is necessary to define criteria that allow comparing them each with other. To this purpose, extending Pestov's axiomatic approach [75], we say that an *Ideal* ID estimator should:

1. be *computational feasible*;
2. be *robust to the multiscaling*;
3. be *robust to the high dimensionality*;
4. have a *work envelope* (or *operative range*);
5. be accurate, i.e., give an ID estimate *close to the underlying manifold dimensionality* (*accuracy*).

The first requirement is motivated by applicative reasons. An estimator that is hard to implement or requires huge computational resources that cannot be used in real world problems. The robustness of an ID estimator w.r.t. the multiscaling is motivated by ID dependence on the data scale [51,72,100]. In order to show this, we consider a two-dimensional data set (e.g., a K-Möbius strip [40]), and we add a three-dimensional gaussian noise to it. The data set, obtained this way, has an ID equal to two at a coarse scale, since the two-dimensional set is dominant. But if the same data set is observed at fine scale, the noise becomes dominant and the data set ID is three. ID estimator robustness w.r.t. high dimensionality is desirable since such an estimator should provide a reliable estimate even if the data set ID is very high, as it can happen in bioinformatics and text categorization applications. The fourth requirement is borrowed by Virtual Reality [10] where the work envelope (or *operative range*) indicates the range where a sensor gives reliable measures. In our specific case the work envelope of an ID estimator is the minimum cardinality that a data set should have so that the estimator gets a reliable estimate. In this way, we view implicitly the ID estimator as a