Expert Systems with Applications 40 (2013) 2841–2844

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Learning morphological maps of galaxies with unsupervised regression

Oliver Kramer [a,b,*], Fabian Gieseke [a], Kai Lars Polsterer [b]

[a] Department of Computing Science, University of Oldenburg, Germany
[b] Department of Physics and Astronomy, Ruhr-University of Bochum, Germany

ARTICLE INFO

Keywords:
Machine learning
Dimensionality reduction
Unsupervised nearest neighbors
Astronomy
Hubble sequence

ABSTRACT

Hubble's morphological classification of galaxies has found broad acceptance in astronomy since decades. Numerous extensions have been proposed in the past, mostly based on galaxy prototypes. In this work, we automatically learn morphological maps of galaxies with unsupervised machine learning methods that preserve neighborhood relations and data space distances. For this sake, we focus on a stochastic variant of unsupervised nearest neighbors (UNN) for arranging galaxy prototypes on a two-dimensional map. UNN regression is the unsupervised counterpart of nearest neighbor regression for dimensionally reduction. In the experimental part of this article, we visualize the embeddings and compare the learning results achieved by various UNN parameterizations and related dimensionality reduction methods.

1. Introduction

Galaxies are massive, gravitationally bounded systems of stars, gas, and dust. The number of stars typically varies in the range of 10^7–10^{14}. In 1926, Edwin Hubble introduced a classification scheme based on their shape. It became famous as Hubble sequence. The Hubble sequence has been extended in the past, but is still the most important classification scheme in modern astronomy. The advent of powerful computing strategies allows the application of advanced data mining to large astronomical databases. In our previous work, we applied classification methods to the quasar recognition problem (Gieseke et al., 2010).

The objective of this work is to automatically learn morphological maps of galaxies with similar characteristics like the Hubble sequence, i.e., similar types of galaxies are neighbored on the map. Images of galaxies allow the categorization of galaxies and the recognition of morphological relationships, similarities, and dissimilarities. From a machine learning perspective, images can be considered as high-dimensional patterns. The process of grouping and learning relationships between classes can be automatized with unsupervised methods. In this article, we employ UNN regression for learning two-dimensional latent topologies based on stochastic sampling (Kramer, 2012).

Let $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\} \subset \mathbb{R}^d$ be the set of high-dimensional patterns with corresponding pattern matrix $\mathbf{Y} = [\mathbf{y}_i]_{i=1}^N \in \mathbb{R}^{d \times N}$. Let $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathbb{R}^q$ be an arbitrary set of low-dimensional representations/latent points that define a manifold with $q < d$. Matrix $\mathbf{X} = [\mathbf{x}_i]_{i=1}^N \in \mathbb{R}^{q \times n}$ is the corresponding latent representation. The pairs $(\mathbf{x}_i, \mathbf{y}_i)$ with $1 \leqslant i \leqslant N$ are the patterns with their latent points (positions in latent space). The low-dimensional representation should represent typical characteristics of the high-dimensional data, and should loose as less information as possible, e.g., data space neighborhood relations and distances. The problem is a hard optimization problem, since the latent variables $\mathbf{X}$ are unknown.

In Section 2, we give a brief introduction to the Hubble sequence and the EFIGI galaxy data set we employ. Section 3 describes related work on galaxies and unsupervised learning. Section 4 revisits unsupervised nearest neighbor regression and introduces a variant for embedding patterns in two-dimensional (or higher dimensional) maps. An experimental study of this method on data from the galaxy data set will be presented in Section 5 followed by a summary in Section 6.

2. Hubble sequence and galaxy data set

2.1. Hubble sequence

The Hubble sequence is not only a morphological classification scheme for galaxies, the typical *fork* diagram shows a topological sorting of the galaxy classes. Neighbored classes in this diagram are *similar* to each other. Hubble's classification scheme differentiates between three main classes:

- Elliptical galaxies (type *E*) have smooth, featureless light distributions looking like ellipses. Galaxy *M*87 is a typical example for a supergiant elliptical galaxy, with a distance of 53.5 million light years from Earth.

* Corresponding author at: Department of Computing Science, University of Oldenburg, Germany.
E-mail address: oliver.kramer@uni-oldenburg.de (O. Kramer).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.12.002

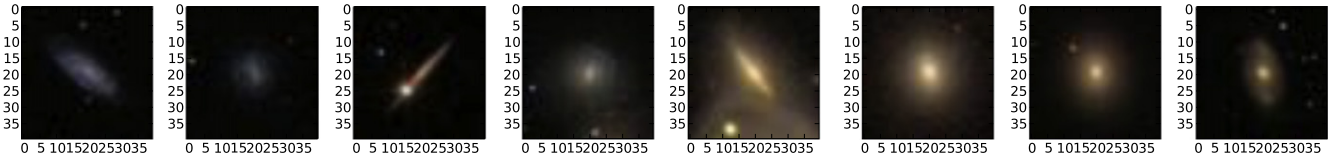

Fig. 1. Visualization of eight sample galaxies of the EFIGI data set Baillard et al., 2011. The EFIGI catalog contains 4458 nearby galaxies. In our experiments, we use a subset of 100 galaxies of types −6 to −4, and 7 to 11.

- Spiral galaxies (type *S*) have a very famous morphological structure. They consist of a flattened disk and stars that form the arms corresponding to the typical spiral structure and a concentration of stars in the center. In the Hubble sequence, regular spirals form the upper arm, while barred spirals can be found in the lower arm.
- Lenticular galaxies (type *S*0) have a bright central bulge, similar to an elliptical galaxy. They are often surrounded by an extended, disk-like structure.

For detailed introduction to galaxies and their morphologies, we refer the reader to the astronomical literature (e.g. Kitchin, 2007).

2.2. SDSS data

Our experimental study is based on data from the Sloan Digital Sky Survey (SDSS) (York et al., 2000) along with label information given the EFIGI catalog (Baillard et al., 2011) that contains 4,458 nearby galaxies. We concentrate on an EFIGI subset of 100 galaxies of types −6 to −4, and 7 to 11. Fig. 1 shows a selection of eight galaxies from the EFIGI data set. Each image is a vector of 40×40 aggregated RGB values $x_i \in [0,1]$, i.e., the data space dimensionality is $d = 1600$. As preprocessing, noisy parts of the image that are potentially not relevant to the galaxy shape are removed by setting all entries below a threshold $\epsilon = 0.1$–0.

3. Related work

3.1. Dimensionality reduction and galaxy images

Not many examples for the application of dimensionality reduction methods to astronomical data sets have been introduced in the past. Naim, Ratnatunga, and Griffiths (1997) provided a framework for learning two-dimensional self-organizing maps (SOMs) of morphological sequences. They employed images of approximately 3,000 galaxies from the Hubble Space Telescope Wide Field Planetary Camera 2 and discovered a population of bulge-dominated galaxies that are similar to peculiar galaxies. SOMs have also been used as classifiers, e.g., by Miller and Coe (1996). They trained SOMs with a small training set of galaxies and report recognition accuracies of up to 98% with a limiting magnitude of 20. Vanderplas and Connolly (2009) applied locally linear embedding (LLE) to spectral data of galaxies from the SDSS. They use the LLE model to improve the classification of emission-line spectra.

3.2. Unsupervised regression

In unsupervised regression, a regression model $\mathbf{f}_{\mathbf{X}}(\cdot) : \mathbb{R}^q \to \mathbb{R}^d$ maps every point $\mathbf{x}_i \in \mathbb{R}^q$ from latent to data space to reconstruct the corresponding pattern $\mathbf{y}_i$. We define the mapping $\mathbf{f}(\cdot) : \mathbb{R}^{q \times N} \to \mathbb{R}^{d \times N}$ from latent space $\mathbb{R}^q$ to data space $\mathbb{R}^d$ for matrices as follows

$$\mathbf{f}_{\mathbf{X}}(\mathbf{X}) = [\mathbf{f}_{\mathbf{X}}(\mathbf{x}_j)]_{j=1}^{N}. \tag{1}$$

For the optimal manifold $\mathbf{X}^*$, the data space reconstruction error (DSRE)

$$E_{\mathbf{X}}(\mathbf{X}) = \|\mathbf{f}_{\mathbf{X}}(\mathbf{X}) - \mathbf{Y}\|_F^2 \tag{2}$$

with Frobenius norm $\|\cdot\|_F^2$ is minimal, i.e., it holds

$$\mathbf{X}^* = \arg\min_{\mathbf{X} \in \mathbb{R}^{q \times N}} E_{\mathbf{X}}(\mathbf{X}). \tag{3}$$

Unsupervised regression has first been applied to kernel density regression (Meinicke, Klanke, Memisevic, & Ritter, 2005) and later to radial basis function networks (RBFs) (Smola, Mika, Schölkopf, & Williamson, 2001), Gaussian processes (Lawrence, 2005), and neural networks (Tan & Mavrovouniotis, 1995). Recently, we fitted nearest neighbor regression to the unsupervised regression framework (Kramer, 2011) and introduced extensions w.r.t. loss functions (Kramer). The stochastic extension we introduce in this work is based on the stochastic variant given in Kramer (2012).

4. Unsupervised nearest neighbor regression

In this section, we introduce UNN regression (Kramer, 2011) based on Gaussian sampling in latent space like introduced in Kramer (2012). Unsupervised nearest neighbors is an approach that iteratively constructs the manifold. For the first pattern $\mathbf{y}_1$, the latent position can be arbitrarily chosen, e.g., as vector of zeros $\mathbf{x}_1 = \mathbf{0}$. Latent matrix is $\overline{\mathbf{X}} = [\mathbf{x}_1]$, and the corresponding pattern matrix is $\overline{\mathbf{Y}} = [\mathbf{y}_1]$.

Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be the sequence of already considered patterns with associated embeddings $\mathbf{x}_1, \ldots, \mathbf{x}_n$. For the next pattern $\mathbf{y}_{n+1}$ with $n + 1 \leqslant N$, UNN searches for the closest already embedded pattern

$$\mathbf{y}^* = \arg\min_{\mathbf{y} = \mathbf{y}_1, \ldots, \mathbf{y}_n} \|\mathbf{y}_{n+1} - \mathbf{y}\|^2 \tag{4}$$

of pattern matrix $\overline{\mathbf{Y}} = [\mathbf{y}_j]_{j=1}^{n}$. Based on the latent position $\mathbf{x}^*$ belonging to pattern $\mathbf{y}^*$, κ candidate latent positions $\mathbf{x}_1^*, \ldots, \mathbf{x}_\kappa^*$ are generated with the Gaussian distribution

$$\mathbf{x}_l^* = \mathbf{x}^* + \mathbf{z}_l \tag{5}$$

for $l = 1, \ldots, \kappa$ with

$$\mathbf{z}_l \sim \sigma \mathcal{N}(0, 1) \tag{6}$$

and $\sigma = \|\mathbf{y}_{n+1} - \mathbf{y}^*\|$. The distance σ between pattern $\mathbf{y}_{n+1}$ that has to be embedded, and the closest embedded pattern $\mathbf{y}^*$ is employed as scaling factor of the Gaussian-based latent point sampling representing distance preservation of data space in latent space. The candidate latent point is chosen that minimizes the DSRE

$$\mathbf{x}_{n+1} = \arg\min_{\mathbf{x} = \mathbf{x}_1^*, \ldots, \mathbf{x}_\kappa^*} e_{\overline{\mathbf{X}}}(\mathbf{x}), \tag{7}$$

which is defined as

$$e_{\overline{\mathbf{X}}}(\mathbf{x}) = \|\mathbf{f}_{\overline{\mathbf{X}}}(\mathbf{x}) - \mathbf{y}_{n+1}\|^2. \tag{8}$$

Here, $\mathbf{f}_{\overline{\mathbf{X}}}(\cdot) : \mathbb{R}^q \to \mathbb{R}^d$ is the K-nearest neighbor (KNN) regression model defined as