Contents lists available at ScienceDirect

Neurocomputing

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

Incremental kernel spectral clustering for online learning of non-stationary data

Rocco Langone*, Oscar Mauricio Agudelo, Bart De Moor, Johan A.K. Suykens

Department of Electrical Engineering (ESAT)-STADIUS/iMinds Future Health Department, KU Leuven, B-3001 Leuven, Belgium

ARTICLE INFO

Article history:
Received 30 July 2013
Received in revised form
2 December 2013
Accepted 17 February 2014
Communicated by B. Hammer
Available online 13 April 2014

Keywords:
Incremental kernel spectral clustering
Out-of-sample eigenvectors
LS-SVMs
Online clustering
Non-stationary data
PM_{10} concentrations

ABSTRACT

In this work a new model for online clustering named Incremental kernel spectral clustering (IKSC) is presented. It is based on kernel spectral clustering (KSC), a model designed in the Least Squares Support Vector Machines (LS-SVMs) framework, with primal-dual setting. The IKSC model is developed to quickly adapt itself to a changing environment, in order to learn evolving clusters with high accuracy. In contrast with other existing incremental spectral clustering approaches, the eigen-updating is performed in a model-based manner, by exploiting one of the Karush–Kuhn–Tucker (KKT) optimality conditions of the KSC problem. We test the capacities of IKSC with some experiments conducted on computer-generated data and a real-world data-set of PM_{10} concentrations registered during a pollution episode occurred in Northern Europe in January 2010. We observe that our model is able to precisely recognize the dynamics of shifting patterns in a non-stationary context.

1. Introduction

In many real-life applications we face the ambitious challenge of online clustering of non-stationary data. Voice and face recognition, community detection of evolving networks such as the World Wide Web or the metabolic pathways in biological cell, object tracking in computer vision, represent just few examples. Therefore researchers perceived the need of developing clustering methods that can model the complex dynamics of evolving patterns in a real-time fashion. Indeed, in the recent past many adaptive clustering models with different inspiration have been proposed: evolutionary spectral clustering techniques [7,9,18,20], self-organizing time map [28], dynamic clustering via multiple kernel learning [27], incremental K-means [8] constitute some examples. Here we focus our attention on the family of the Spectral Clustering (SC) approaches [25,31,10], which has shown its practical success in many application domains. SC is an off-line algorithm, and the above-cited attempts to make it applicable to dynamic data-sets, although quite appealing, are at the moment not very computationally efficient. In [26] and more recently in [11], the authors propose some incremental eigenvalue solutions to continuously update the initial eigenvectors found by SC. In this paper, we follow this direction, but with an important difference. The incremental eigen-update we introduce is model-based and cast in a machine learning framework, since our core model is kernel spectral clustering (KSC, [3]). KSC is an LS-SVM formulation [29] of Spectral Clustering with two main advantages: an organized model-selection procedure based on several criteria (BLF, Modularity, AMS, [3,17,19]) and the extension of the clustering model to out-of-sample data. Moreover, it can scale to large data as it has been shown in [23,24] and very sparse models can be constructed [22,2]. In KSC a clustering model can be trained on a subset of the data and then applied to the rest of the data in a learning framework. The out-of-sample extension allows then to predict the memberships of a new point thanks to the previously learned model. The out-of-sample extension alone, without the need of ad-hoc eigen-approximation techniques like the ones proposed in [26] and[11], can be used to accurately cluster stationary data-streams. For instance, in [16], KSC has been applied for online fault detection of an industrial machine. In this work KSC was trained offline to recognize two main working regimes, namely good and faulty state. Then it was used in an online fashion via the out-of-sample extension to raise an early warning when necessary.

However, if the data are generated according to some distribution which change over time (i.e. non-stationary), the initial KSC model must be updated. In order to solve this issue we introduce

* Corresponding author.
E-mail addresses: rocco.langone@esat.kuleuven.be (R. Langone), mauricio.agudelo@esat.kuleuven.be (O. Mauricio Agudelo), bart.demoor@esat.kuleuven.be (B. De Moor), johan.suykens@esat.kuleuven.be (J.A.K. Suykens).

http://dx.doi.org/10.1016/j.neucom.2014.02.036
0925-2312/© 2014 Elsevier B.V. All rights reserved.

the Incremental Kernel Spectral Clustering Algorithm (IKSC). The IKSC method takes advantage of the work presented in [4] to continuously adjust the initial KSC model over-time, in order to learn the complex dynamics characterizing the non-stationary data.

The remainder of this paper is structured as follows: in Section 2 we briefly recall the KSC model. Section 3 introduces the new IKSC algorithm. Section 4 describes the data-sets used in the experiments. In Section 5 we discuss the simulation results and we compare our method with incremental K-means (IKM).To better understand our technique and the experimental findings we advice the readers to take a look at the demonstrative videos present in the supplementary material of this paper. Finally, Section 6 concludes the paper.

2. Kernel spectral clustering (KSC)

Spectral clustering methods use the eigenvectors of the graph Laplacian to unfold the data manifold and properly group the datapoints. In contrast with classical spectral clustering, KSC is considered in a learning framework. This allows the out-of-sample extension of the clustering model to test points in a straightforward way. With training data $\mathcal{D}=\{x_i\}_{i=1}^N, x_i \in \mathbb{R}^d$ and the number of clusters k, the kernel spectral clustering optimization problem can be stated in the following way [3]:

$$\min_{w^{(l)},e^{(l)},b_l} \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} - \frac{1}{2N}\sum_{l=1}^{k-1} \gamma_l e^{(l)^T} D^{-1} e^{(l)} \tag{1}$$

$$\text{such that } e^{(l)} = \Phi w^{(l)} + b_l 1_N. \tag{2}$$

This is a weighted kernel PCA formulation, being the weighting matrix equal to the degree matrix D associated to the training kernel matrix. The objective consists of minimizing the regularization terms and maximizing the weighted variance of the projections of the data points in the feature space. The score variables[1] are named $e^{(l)} = [e_1^{(l)}, \dots, e_N^{(l)}]^T$, $l = 1, \dots, k-1$ indicates the number of score variables needed to encode the k clusters to find, $D^{-1} \in \mathbb{R}^{N \times N}$ is the inverse of the degree matrix D, Φ is the $N \times d_h$ feature matrix $\Phi = [\varphi(x_1)^T, \dots, \varphi(x_N)^T]$ and $\gamma_l \in \mathbb{R}^+$ are regularization constants. The multiway clustering model in the primal space is expressed by a set of $k-1$ binary problems, which are combined in an Error Correcting Output Code (ECOC) encoding scheme:

$$e_i^{(l)} = w^{(l)^T} \varphi(x_i) + b_l, \quad i = 1, \dots, N, \quad l = 1, \dots, k-1. \tag{3}$$

where $w^{(l)} \in \mathbb{R}^{d_h}$ is the parameter vector in the primal space associated with the l-th binary clustering, b_l are bias terms, $\varphi: \mathbb{R}^d \to \mathbb{R}^{d_h}$ is the mapping of the input points x_i into a high-dimensional feature space of dimension d_h. The projections $e_i^{(l)}$ represent the latent variables of the group of $k-1$ binary clustering indicators given by $\text{sign}(e_i^{(l)})$. Thus every point x_i is associated with a latent variable $[e_i^{(1)}, \dots, e_i^{(k-1)}]$ which lives in the low-dimensional space spanned by $w^{(l)}$. The set of binary indicators $\text{sign}(e_i^{(l)}), i = 1, \dots, N, l = 1, \dots, k-1$ form a code-book $\mathcal{CB} = \{c_p\}_{p=1}^k$, where each code-word is a binary word of length $k-1$ representing a cluster.

As for all the kernel-based methods, since an explicit formula of the feature map $\varphi(\cdot)$ is in general unknown, the dual of problem (1) is derived. As a consequence, we go from the parametric representation of the clustering model expressed by Eq. (3) to a non-parametric representation in the dual space denoted by (5). Here only dot products between the mapped points in $\varphi(\cdot)$ appear, which can be easily computed using the kernel trick derived by the Mercer theorem: $\varphi(x_i)^T \varphi(x_j) = K(x_i, x_j)$. In Fig. 1 for the sake of clarity we illustrate, in the case of a synthetic dataset consisting of three intertwined spirals, the points mapped in the space of the eigenvectors $\alpha^{(l)}$ and the space of the latent variables $e^{(l)}$.

The Lagrangian associated with the primal problem, written in matrix form, is

$$\mathcal{L}(w^{(l)}, e^{(l)}, b_l, \alpha^{(l)}) = \frac{1}{2}\sum_{l=1}^{k-1} w^{(l)^T} w^{(l)} - \frac{1}{2N}\sum_{l=1}^{k-1} \gamma_l e^{(l)^T} D^{-1} e^{(l)} - \sum_{l=1}^{k-1} \alpha^{(l)^T} (e^{(l)} - \Phi w^{(l)} - b_l 1_N)$$

where $\alpha^{(l)}$ are the Lagrange multipliers. The KKT optimality conditions are the following:

$$\frac{\partial \mathcal{L}}{\partial w^{(l)}} = 0 \to w^{(l)} = \Phi^T \alpha^{(l)},$$
$$\frac{\partial \mathcal{L}}{\partial e^{(l)}} = 0 \to \alpha^{(l)} = \frac{\gamma_l}{N} D^{-1} e^{(l)},$$
$$\frac{\partial \mathcal{L}}{\partial b_l} = 0 \to 1_N^T \alpha^{(l)} = 0,$$
$$\frac{\partial \mathcal{L}}{\partial \alpha^{(l)}} = 0 \to e^{(l)} - \Phi w^{(l)} - b_l 1_N = 0.$$

Once we have solved the KKT conditions for optimality, we can derive the following dual problem:

$$D^{-1} M_D \Omega \alpha^{(l)} = \lambda_l \alpha^{(l)} \tag{4}$$

where Ω is the kernel matrix with ij-th entry $\Omega_{ij} = K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$, D is the related graph degree matrix which is diagonal with positive elements $D_{ii} = \sum_j \Omega_{ij}$, M_D is a centering matrix defined as

$$M_D = I_N - \frac{1}{1_N^T D^{-1} 1_N} 1_N 1_N^T D^{-1},$$

$\alpha^{(l)}$ are the dual variables, $\lambda_l = N/\gamma_l$ and $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is the kernel function and captures the similarity between the datapoints. The clustering model in the dual space evaluated on training data becomes

$$e^{(l)} = \Omega \alpha^{(l)} + b_l 1_N, \quad l = 1, \dots, k-1. \tag{5}$$

The eigenvectors $\alpha^{(l)}$ express an embedding of the input data that reveals the underlying clustering structure. They are linked to the $w^{(l)}$ through the first KKT condition.

In order to cope with truly non-stationary data arriving over time, the initial $\alpha^{(l)}$ must be modified in response to the new inputs. This issue is tackled by means of the incremental kernel spectral clustering algorithm, which will be explained in detail in the next section.

The out-of-sample extension is performed by the ECOC decoding scheme. In the decoding process the cluster indicators found in the validation/test stage are compared with the code-book and the nearest code-word indicated by the Hamming distance is selected. The cluster indicators are the results of binarizing the score variables for test points:

$$\text{sign}(e_{\text{test}}^{(l)}) = \text{sign}(\Omega_{\text{test}} \alpha^{(l)} + b_l 1_{\text{Ntest}}) \tag{6}$$

with $l = 1, \dots, k-1$. Ω_{test} is the $N_{\text{test}} \times N$ kernel matrix evaluated using the test points with entries $\Omega_{\text{test},ri} = K(x_r^{\text{test}}, x_i)$, $r = 1, \dots, N_{\text{test}}$, $i = 1, \dots, N$.

In the first two synthetic experiments that will be presented in Section 4.1.1 (Drifting Gaussians and Merging Gaussians) we use the RBF kernel function defined by $K(x_i, x_j) = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$. The symbol σ indicates the bandwidth parameter and x_i is the i-th data point. In the analysis of the third synthetic data

[1] We use interchangeably the terms projections, score variables, latent variables to name the $e^{(l)}$.