Neurocomputing 318 (2018) 213–226

Contents lists available at ScienceDirect

Neurocomputing

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

Supervised low rank indefinite kernel approximation using minimum enclosing balls

Frank-Michael Schleif[a,c,*], Andrej Gisbrecht[b], Peter Tino[a]

[a] School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK
[b] Helsinki Institute for Information Technology, Department of Computer Science, Aalto University, Finland
[c] School of Computer Science, University of Appl. Sc. Wuerzburg-Schweinfurt, Wuerzburg 97074, Germany

ARTICLE INFO

Article history:
Received 22 January 2018
Revised 28 June 2018
Accepted 20 August 2018
Available online 3 September 2018

Communicated by Ivor Tsang

Keywords:
Indefinite kernel
Kernel fisher discriminant
Minimum enclosing ball
Nyström approximation
Low rank approximation
Classification
Indefinite learning

ABSTRACT

Indefinite similarity measures can be frequently found in bio-informatics by means of alignment scores, but are also common in other fields like shape measures in image retrieval. Lacking an underlying vector space, the data are given as pairwise similarities only. The few algorithms available for such data do not scale to larger datasets. Focusing on probabilistic batch classifiers, the Indefinite Kernel Fisher Discriminant (iKFD) and the Probabilistic Classification Vector Machine (PCVM) are both effective algorithms for this type of data but, with cubic complexity. Here we propose an extension of iKFD and PCVM such that linear runtime and memory complexity is achieved for low rank indefinite kernels. Employing the Nyström approximation for indefinite kernels, we also propose a new almost parameter free approach to identify the landmarks, restricted to a *supervised* learning problem. Evaluations at several larger similarity data from various domains show that the proposed methods provides similar generalization capabilities while being easier to parametrize and substantially faster for large scale data.

1. Introduction

Domain specific proximity measures, like alignment scores in bioinformatics [1], the modified Hausdorff-distance for structural pattern recognition [2], shape retrieval measures like the inner distance [3] and many other ones generate non-metric or indefinite similarities or dissimilarities. Classical learning algorithms like kernel machines assume Euclidean metric properties in the underlying data space and may not be applicable for this type of data.

Only few machine learning methods have been proposed for non-metric proximity data, like the indefinite kernel Fisher discriminant (iKFD) [4,5], the probabilistic classification vector machine (PCVM) [6] or the indefinite Support Vector Machine (iSVM) in different formulations [7–9]. For the PCVM the provided kernel evaluations are considered only as basis functions and no Mercer conditions are implied. In contrast to the iKFD the PCVM is a *sparse* probabilistic kernel classifier pruning unused basis functions during training, applicable to arbitrary positive definite *and* indefinite kernel matrices. A recent review about learning with indefinite proximities can be found in [10].

While being very efficient these methods do not scale to larger datasets with in general cubic complexity. In [11,12] the authors proposed a few Nyström based (see e.g. [13]) approximation techniques to improve the scalability of the PCVM for low rank matrices. The suggested techniques use the Nyström approximation in a non-trivial way to provide *exact* eigenvalue estimations also for *indefinite* kernel matrices. This approach is very generic and can be applied in different algorithms. In this contribution we further extend our previous work and not only derive a low rank approximation of the indefinite kernel Fisher discriminant, but also address the landmark selection from a novel view point. The obtained Ny-iKFD approach is linear in runtime and memory consumption, for low rank matrices. The formulation is exact if the rank of the matrix equals the number of independent landmarks points. The selection of the landmarks of the Nyström approximation is a critical point addressed in previous work (see e.g. [14–16]). Most recently leverage scores [17] have been found very promising, but with quadratic costs. In general these strategies use the full positive semi-definite (psd) kernel matrix or expect that the kernel is of some standard class like an RBF kernel. In each case the approaches presented so far are costly in runtime and memory consumption as can be seen in the subsequent experiments.

* Corresponding author at: School of Computer Science, University of Birmingham, Birmingham B15 2TT, UK.
E-mail addresses: frank-michael.schleif@fhws.de, fschleif@techfak.uni-bielefeld.de (F.-M. Schleif), andrej.gisbrecht@aalto.fi (A. Gisbrecht), pxt@cs.bham.ac.uk (P. Tino).

https://doi.org/10.1016/j.neucom.2018.08.057
0925-2312/

Additionally, former approaches for landmark selection aim on generic matrix reconstructions of positive semi definite (psd) kernels. We propose a restricted reconstruction of the psd or non-psd kernel matrix with respect to a *supervised* learning scenario only. We no longer expect to obtain an accurate kernel reconstruction from the approximated matrix (e.g. by using the Frobenius norm) but are pleased if the approximated matrix preserves the class boundaries in the data space.

In [12] the authors derived methods to approximate large proximity matrices by means of the Nyström approximation and conversion rules between similarities and dissimilarities. These techniques have been applied in [11] and [18] in a proof of concept setting, to obtain approximate models for the Probabilistic Classification Vector Machine and the Indefinite Fisher Kernel Discriminant analysis using a random landmark selection scheme. This work is substantially extended and detailed in this article with a specific focus on *indefinite* kernels, only. A novel landmark selection scheme is proposed. Based on this new landmark selection scheme we provide detailed new experimental results and compare to alternative landmark selection approaches. The paper provides the following improvements over the current state of the art: (1) A linear costs approximation scheme for the Indefinite Kernel Fisher Discriminant (iKFD) and the probabilistic classification vector machine (PCVM) is provided. (2) A new supervised landmark selection scheme is proposed which can be also applied to indefinite input kernels to obtain a Nystroem approximation of the given indefinite kernel. (3) A variety of experimental results is provided showing the efficiency of the proposed approach and linked to related work.

Structure of the paper: First we give some basic notations necessary in the subsequent derivations. Then we review iKFD and PCVM as well as some approximation concepts proposed by the authors in [11] which are based on the well known Nyström approximation. Subsequently, we consider the landmark selection problem in more detail and show empirically results motivating a supervised selection strategy. Finally we detail the reformulation of iKFD and PCVM based on the introduced concepts and show the efficiency in comparison to Ny-PCVM and Ny-iKFD for various indefinite proximity benchmark data sets.

2. Methods

2.1. Notation and basic concepts

Consider a collection of N objects $\mathbf{x}_i$, $i = 1, 2, \ldots, N$, in some input space $\mathcal{X}$. Given a similarity function or inner product on $\mathcal{X}$, corresponding to a metric, one can construct a proper Mercer kernel acting on pairs of points from $\mathcal{X}$. For example, if $\mathcal{X}$ is a finite dimensional vector space, a classical similarity function is the Euclidean inner product (corresponding to the Euclidean distance) - a core component of various kernel functions such as the famous radial basis function (RBF) kernel. Now, let $\phi : \mathcal{X} \mapsto \mathcal{H}$ be a mapping of patterns from $\mathcal{X}$ to a Hilbert space $\mathcal{H}$ equipped with the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The transformation ϕ is in general a non-linear mapping to a high-dimensional space $\mathcal{H}$ and may in general not be given in an explicit form. Instead, a kernel function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is given which encodes the inner product in $\mathcal{H}$. The kernel k is a positive (semi) definite function such that $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle_{\mathcal{H}}$, for any $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. The matrix $K_{i,j} := k(\mathbf{x}_i, \mathbf{x}_j)$ is an $N \times N$ kernel (Gram) matrix derived from the training data. The motivation for such an embedding comes with the hope that the non-linear transformation of input data into higher dimensional $\mathcal{H}$ allows for using linear techniques in $\mathcal{H}$. Kernelized methods process the embedded data points in a feature space utilizing only the inner products $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ (kernel trick) [19], without the need to explicitly calculate ϕ. The kernel function can be very generic. Most prominent are the linear kernel with $k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ where $\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$ is the Euclidean inner product and ϕ identity mapping, or the RBF kernel $k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$, with $\sigma > 0$ as a free scale parameter. In any case, it is always assumed that the kernel function $k(\mathbf{x}, \mathbf{x}')$ is positive semi definite (psd). This assumption is however not always fulfilled, and the underlying similarity measure may not be metric and hence not lead to a Mercer kernel. Examples can be easily found in domain specific similarity measures as mentioned before and detailed later on. Such similarity measures imply *indefinite* kernels, preventing standard "kernel-trick" methods developed for Mercer kernels to be applied.

For a matrix A, A^{-1} denotes the inverse of A. We will still use this notation even when A is non-regular. In that case A^{-1} will represent an inverse obtained through an Singular Value Decomposition (SVD) - based regularization.

In what follows we will review some basic concepts and approaches related to such non-metric situations.

2.2. Krein and Pseudo-Euclidean spaces

A Krein space is an *indefinite* inner product space endowed with a Hilbertian topology.

Definition 1 (Inner products and inner product space)**.** Let $\mathcal{Q}$ be a real vector space. An inner product space with an indefinite inner product $\langle \cdot, \cdot \rangle_{\mathcal{Q}}$ on $\mathcal{Q}$ is a bi-linear form where all $f, g, h \in \mathcal{Q}$ and $\alpha \in \mathbb{R}$ obey the following conditions:

- Symmetry: $\langle f, g \rangle_{\mathcal{Q}} = \langle g, f \rangle_{\mathcal{Q}}$
- linearity: $\langle \alpha f + g, h \rangle_{\mathcal{Q}} = \alpha \langle f, h \rangle_{\mathcal{Q}} + \langle g, h \rangle_{\mathcal{Q}}$;
- $\langle f, g \rangle_{\mathcal{Q}} = 0 \; \forall g \in \mathcal{Q}$ implies $f = 0$

An inner product is positive definite if $\forall f \in \mathcal{Q}$, $\langle f, f \rangle_{\mathcal{Q}} \geq 0$, negative definite if $\forall f \in \mathcal{Q}$, $\langle f, f \rangle_{\mathcal{Q}} \leq 0$, otherwise it is indefinite. A vector space $\mathcal{Q}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{Q}}$ is called an inner product space.

Definition 2 (Krein space and pseudo-Euclidean space)**.** An inner product space $(\mathcal{Q}, \langle \cdot, \cdot \rangle_{\mathcal{Q}})$ is a Krein space if we have two Hilbert spaces $\mathcal{H}_+$ and $\mathcal{H}_-$ spanning $\mathcal{Q}$ such that $\forall f \in \mathcal{Q}$ we have $f = f_+ + f_-$ with $f_+ \in \mathcal{H}_+$ and $f_- \in \mathcal{H}_-$ and $\forall f, g \in \mathcal{Q}$, $\langle f, g \rangle_{\mathcal{Q}} = \langle f_+, g_+ \rangle_{\mathcal{H}_+} - \langle f_-, g_- \rangle_{\mathcal{H}_-}$. A finite-dimensional Krein-space is a so called pseudo-Euclidean space (pE).

Indefinite kernels are typically found through domain specific non-metric similarity functions (such as alignment functions used in biology [1]), specific kernel functions (e.g. the Manhattan kernel $k(\mathbf{x}, \mathbf{x}') = -||\mathbf{x} - \mathbf{x}'||_1$, tangent distance kernel [20]), or divergence measures plugged into standard kernel functions [21]. Another source of non-psd kernels are noise artifacts on standard kernel functions [7].

In such spaces vectors can have negative squared "norm", negative squared "distances" and the concept of orthogonality is different from the usual Euclidean case. In the subsequent experiments our input data are in general given by a symmetric indefinite kernel matrix K. We will use the symbol K to denote kernel matrices, whether psd or not. It will be clear from the context if the underlying space is a Hilbert or a Krein space. We use the symbol $\mathbf{S}$ for (symmetric) similarity matrices and $\mathbf{D}$ for a symmetric dissimilarity matrix.

In practical applications it may also happen that the given data are represented by non-metric dissimilarities. A prominent example is the dynamic timewarping score matrix which can be considered as a dissimilarity matrix of pairwise sequence alignments. Given a symmetric *dissimilarity* matrix $\mathbf{D}$ with zero