Information Sciences 436–437 (2018) 214–226

Contents lists available at ScienceDirect

Information Sciences

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

Data-independent Random Projections from the feature-map of the homogeneous polynomial kernel of degree two

Daniel López-Sánchez*, Juan Manuel Corchado, Angélica González Arrieta

Department of Computer Science, University of Salamanca, Salamanca, Spain

ARTICLE INFO

Article history:
Received 9 December 2016
Revised 28 August 2017
Accepted 13 January 2018
Available online 16 January 2018

MSC:
00-01
99-00

Keywords:
Random Projection
Non-linear dimensionality reduction
Polynomial kernel

ABSTRACT

This paper presents a novel non-linear extension of the Random Projection method based on the degree-2 homogeneous polynomial kernel. Our algorithm is able to implicitly map data points to the high-dimensional feature space of that kernel and from there perform a Random Projection to an Euclidean space of the desired dimensionality. Pairwise distances between data points in the kernel feature space are approximately preserved in the resulting representation. As opposed to previous kernelized Random Projection versions, our method is data-independent and preserves much of the computational simplicity of the original algorithm. This is achieved by focusing on a specific kernel function, what allowed us to analyze the effect of its associated feature mapping in the distribution of the Random Projection hyperplanes. Finally, we present empirical evidence that the proposed method outperforms alternative approaches in terms of pairwise distance preservation, while being significantly more efficient. Also, we show how our method can be used to approximate the accuracy of non-linear classifiers with efficient linear classifiers in some datasets.

1. Introduction

Random Projection (RP) is a simple yet computationally efficient dimensionality reduction method. While other dimensionality reduction methods compute the directions onto which the data will be projected by analyzing a training dataset, RP populates the projection matrix with a random normal distribution, thus requiring low computational effort. This technique is based on a theoretical result known as the Johnson–Lindenstrauss (JL) lemma: For any $0 < \epsilon < 1$ and $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ there is a map $f : \mathbb{R}^d \to \mathbb{R}^k$ for $k = \mathcal{O}(\epsilon^{-2}\log(n))$ such that:

$$\forall i, j \quad (1-\epsilon)||x_i - x_j||^2 \leqslant ||f(x_i) - f(x_j)||^2 \leqslant (1+\epsilon)||x_i - x_j||^2 \tag{1}$$

Furthermore, this map can be found in randomized polynomial time. For a simple proof of this lemma, refer to Dasgupta and Gupta [10]. Originally, the map f consisted of projecting the points from $\mathbb{R}^d$ to $\mathbb{R}^k$ by means of a $d \times k$ matrix whose elements were drawn from a Gaussian distribution. Later on, Achlioptas proved that the projection matrix could be populated according to a much simpler distribution [1]. Concretely, he showed that the entries of the projection matrix R can be computed as shown in Eq. (2); where Achlioptas used $s = 1$ and $s = 3$ (see [22]).

$$r_{ij} = \sqrt{s}\begin{cases} 1 & \text{with prob.} \quad 1/2s \\ 0 & \text{with prob.} \quad 1 - 1/s \\ -1 & \text{with prob.} \quad 1/2s \end{cases} \tag{2}$$

* Corresponding author.
E-mail address: lope@usal.es (D. López-Sánchez).

https://doi.org/10.1016/j.ins.2018.01.022

In fact, Achlioptas proved that the only requirement to satisfy the JL lemma is that the entries of the projection matrix must be independent random variables with zero mean and unit variance. Once the $d \times k$ matrix R has been populated, an arbitrary set of n points represented as an $n \times d$ matrix X can be projected from $\mathbb{R}^d$ to $\mathbb{R}^k$ according to Eq. (3) (see [1]).

$$X'_{n\times k} = \frac{1}{\sqrt{k}} X_{n\times d} R_{d\times k} \tag{3}$$

Using the distribution proposed by Achlioptas to populate the projection matrix reduces the computational cost of the projection. This is because the multiplication by $\sqrt{s}$ present in Eq. (2) can be delayed, so the computation of the projection itself reduces to aggregate evaluation (i.e. summation and subtraction but no multiplication).

Over the years, various authors [3,32] have explored the possibility of performing random projections from the feature space associated to different kernel functions. This is of interest because of two main reasons: (1) if the JL-lemma is satisfied, the pairwise distances between samples in the kernel feature space will be preserved in the resulting representation, so the low-dimensional projected points will preserve most of the structure from the kernel feature-space; and (2), since Random Projection preserves separability margins [3], classification problems may become more lineally solvable after the Random Projection form a kernel feature space. If this is the case, the classification accuracy of scalable linear classifiers will increase. Motivated by these advantages, Zhao et al. recently proposed a kernelized variant of the original Random Projection Algorithm, capable of working with an arbitrary kernel function [32]. As a drawback, their proposal lost some of the beneficial features of the original Random Projection algorithm. In particular, the method proposed in [32] is data-dependent and, unlike the original Random Projection algorithm, requires an expensive training phase. In addition, this algorithm is not compatible with the database-friendly distribution proposed by Achlioptas [1] and its distance-preserving capabilities have not been analyzed empirically. Moreover, it has been theoretically proven that formulating an algorithm capable of performing a Random Projection from the feature space of an arbitrary kernel function, by just having black-box access to the kernel function but no training samples (i.e., in a data-independent manner) is not possible [3,5]. However, the question of whether this could be achieved for particular kernel functions such as the polynomial kernel remains open.

In this paper, we present a novel method to efficiently perform random projections from the feature space of the homogeneous polynomial kernel of degree two. By focusing on a specific kernel function, our method overcomes the limitations of previous kernelization attempts of the Random Projection algorithm (i.e., data-dependence and training inefficiency). In addition, our method is compatible with the database-friendly distribution proposed by Achlioptas. Our experimental results evidence that, by using our method, one can efficiently generate a low-dimensional representation of data samples that condenses the structure of data in the feature space of the degree-2 homogeneous polynomial kernel, approximately preserving the pairwise distances between samples in that space. Because of the data-independent nature of the proposed method, it can be applied in online-learning scenarios [13] where no data samples are initially available. In addition, this representation can be used to train efficient linear classifiers that approximate the accuracy of their non-linear counterparts.

Formally, we propose replacing the inner product that takes place during the matrix multiplication of Eq. (3) (i.e. projection of data points of X onto directions in R) by the kernel function (see Eq. (4)). By doing so, the columns of the projection matrix and the data samples are mapped by $\phi(\cdot)$ and the projection takes place in the kernel feature space:

$$x'_{ij} = \frac{1}{\sqrt{k}}\, K\big([x_{i1}, x_{i2}, \ldots, x_{id}],\ [r_{1j}, r_{2j}, \ldots, r_{dj}]\big) = \frac{1}{\sqrt{k}}\, \langle \phi(row_i X), \phi(col_j R)\rangle \tag{4}$$

where $\langle\cdot,\cdot\rangle$ denotes the scalar product. However, this is not in general equivalent to explicitly mapping the data points to the extended feature space by means of $\phi : \mathbb{R}^d \to \mathcal{H}$ and then performing a Random Projection. This is because we cannot guarantee that the distribution of the elements in R will be preserved by the embedding $\phi(\cdot)$. Therefore, our goal will be to define the projection matrix R in such a way that when its columns are mapped by $\phi(\cdot)$ to the implicit, high-dimensional feature space $\mathcal{H}$ the result will be a set of valid random projection directions. This is the reason why we must focus on a specific kernel function, so we can analyze how its associated mapping $\phi(\cdot)$ affects the columns of the projection matrix.

The rest of this paper is structured as follows. Section 2 reviews the related work available in the literature. Section 3 describes the proposed approach and how it manages to efficiently perform a valid Random Projection form the feature space of the degree-2 homogeneous polynomial kernel. An improved method to construct the implicit projection hyperplanes is then described in Section 4. In Section 5, we provide empirical evidence that our method approximates a Random Projection form the feature space of the degree-2 homogeneous polynomial kernel, so the pairwise distances between points in the feature space are approximately preserved in the reduced representation. Additional experimental results regarding the suitability of our method for the task of classification are reported in Section 6. Finally, Section 7 summarizes the conclusions and future research lines.

2. Related work

This paper focuses on the design of a novel method to perform a Random Projection from the feature space associated to a specific kernel function, namely the degree-2 homogeneous polynomial kernel. By performing this random projection, it is possible to efficiently generate a compact representation of samples that captures the structure of data in the kernel feature space. Conveniently, this is achieved without any explicit computation of the expensive mapping function $\phi(\cdot)$. By focusing on a specific kernel function, we were able to preserve some of the most important beneficial features of the original