Neurocomputing 149 (2015) 767–776

Contents lists available at ScienceDirect

Neurocomputing

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

# A Projection Pursuit framework for supervised dimension reduction of high dimensional small sample datasets

Soledad Espezua [a,*], Edwin Villanueva [a], Carlos D. Maciel [b], André Carvalho [a]

[a] Department of Computer Science, ICMC - USP, University of São Paulo, Brazil
[b] Department of Electrical Engineering, São Carlos School of Engineering, University of São Paulo, Brazil

ARTICLE INFO

*Article history:*
Received 6 June 2013
Received in revised form
2 June 2014
Accepted 28 July 2014
Communicated by Shiguang Shan
Available online 12 August 2014

*Keywords:*
Projection Pursuit
Classification
Gene expression
Dimension reduction

ABSTRACT

The analysis and interpretation of datasets with large number of features and few examples has remained as a challenging problem in the scientific community, owing to the difficulties associated with the curse-of-the-dimensionality phenomenon. Projection Pursuit (PP) has shown promise in circumventing this phenomenon by searching low-dimensional projections of the data where meaningful structures are exposed. However, PP faces computational difficulties in dealing with datasets containing thousands of features (typical in genomics and proteomics) due to the vast quantity of parameters to optimize. In this paper we describe and evaluate a PP framework aimed at relieving such difficulties and thus ease the construction of classifier systems. The framework is a two-stage approach, where the first stage performs a rapid compaction of the data and the second stage implements the PP search using an improved version of the SPP method (Guo et al., 2000, [32]). In an experimental evaluation with eight public microarray datasets we showed that some configurations of the proposed framework can clearly overtake the performance of eight well-established dimension reduction methods in their ability to pack more discriminatory information into fewer dimensions.



## 1. Introduction

In the last few decades we have witnessed a rapid development and refinement of data acquisition technologies in several science and industrial areas [1]. This has led to the emergence of high-throughput technologies that are capable of generating datasets with the number of features ($p$) far greater than the number of examples ($n$), the so-called *large p small n* datasets. A representative example of these technological developments is the microarray technology [2], which has made possible the measurement of expression levels of thousands of genes in a relatively rapid and economic way, leading to significant advances in the understanding of severe diseases, like cancer, and raising hopes on possible cures [3,4].

Though the collection of *large p small n* datasets is nowadays a common practice in many fields, their analysis and interpretation is still a challenging task [5,6,1]. This difficulty is mainly originated by the so-called "curse of dimensionality" phenomenon, inherent in such a kind of data [7]. This phenomenon states that as the dimensionality increases, the corresponding space becomes emptier and the data points tend to be equidistant. This generates detrimental impacts in most machine-learning and pattern-recognition methods (including model-estimation instability, model over fitting and local convergence), compromising the generalization performance and reliability of such methods [5,6].

A common approach to circumvent the curse of dimensionality is by reducing it [6]. Two kinds of methods exist for this task: feature selection (FS) [8,9] and feature extraction (FE) [10,11]. The former methods try to find small subsets of original features that are relevant to the intended analysis. The latter methods reduce the dimensionality by building new features from combinations (linear or nonlinear) of the original features. FS has the benefit of keeping the original feature meaning, facilitating the interpretability by the domain expert [9]. However, it has been said [12] that FE is preferable over FS when the final goal is an accurate system for classifying new examples and interpretability is not as important. This is because FE is not tied to the original feature space, providing greater chances of finding more useful representations for the desired task [12].

Projection pursuit (PP) [13,14] is a FE method that has been successfully applied in several domains for both supervised and unsupervised analyses (e.g. [15–18]). PP seeks low-dimensional linear projections of the data that expose interesting aspects of them. To this end, a measure of "interestingness" is employed, which is known as *projection pursuit index* (PP index). A key advantage of PP is its flexibility to fit different pattern recognition tasks, depending on the PP index used. For example, PP can be

* Corresponding author. Tel.: +55 16 34132126.
*E-mail addresses:* sespezua@usp.br (S. Espezua), evillatal@usp.br (E. Villanueva), maciel@sc.usp.br (C.D. Maciel), andre@icmc.usp.br (A. Carvalho).

http://dx.doi.org/10.1016/j.neucom.2014.07.057
0925-2312/

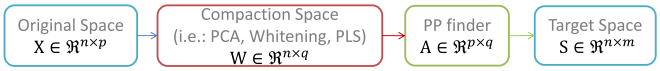


**Fig. 1.** Framework WSPP.

used to perform clustering analysis [19,20], classification [21–24], regression analysis [25] and density estimation [26] (some reviews of PP indexes can be found in [21,27,28]). Another advantage of PP is its out-of-sample mapping capability, that is, the possibility to map new examples in the projection space after the construction of it.

Despite the aforementioned advantages, the literature shows a limited use of PP in *large p small n* datasets, like those generated by microarray technology. This may be due to the high computational difficulty in finding optimal projection spaces for such cases. For instance, the projection of a dataset with $p=10$k features (a realistic number in microarray datasets) onto a target space of dimension $m=3$ will require the optimization of a projection matrix of $p \times m = 30$k elements. Evidently, the problem worsens as $p$ or $m$ increase. Traditional PP optimizers based on the gradients or Newton methods [29–31,19] are usually inadequate for such a kind of data due to the vastness of possible projections and, thus, the high susceptibility to find poor local optima [14]. More global PP optimizers were described recently, including genetic algorithms (GA) [32,33], simulating annealing (SA) [21], random scan sampling (RSSA) [34] and particle swarm optimization (PSO) [35]. However, none of these works have been directly applied in dimensionalities as high as those found in microarray data, which shows the difficulty of applying PP in such scenarios.

In this paper we present a framework to facilitate the applicability of PP on *large p small n* datasets with the aim of classification tasks. The framework is formed by two main stages (Fig. 1): a compaction stage and a PP optimization stage. The first stage is devised to rapidly transform the original data into a less sparse representation. The second stage is the PP part, which is responsible to find optimal projections taking the compacted representation as input.

For the compaction stage we use three well-known techniques: PCA, Whitening and Partial Least Squares. For the PP stage, we adopt the Sequential Projection Pursuit (SPP) approach [32] coupled with the GA optimizer (PPGA) we described recently [33], in which a specialized crossover operator showed excelling search capabilities. An experimental study is presented over eight public microarray datasets. The evaluation systematically tested several configurations of the framework, including variations of the compaction method, the PP index function and the target dimensionality. We used the predictive accuracy of two popular classification methods (LDA and 3NN) in order to assess the quality of the tested configurations. We also compare the framework against eight well-established dimension reduction methods, including FE and FS methods.

The paper is organized as follows. Section 2 introduces some important concepts of PP, SPP, PP optimization and PP indexes used in the paper. Section 3 describes the proposed framework. Section 4 presents the experimental evaluation, including the experimental setup, results and corresponding discussion. Finally, our conclusions are presented in Section 5.

## 2. Projection pursuit

The projection pursuit (PP) concept was formally introduced in the paper of Friedman and Tukey [13], although the seminal ideas were originally posed by Kruskal [36]. To describe the PP concept we assume that we have a data matrix $\mathbf{X}$ of $n \times p$ dimensions, where $n$ is the number of data examples or observations and $p$ is the number of attributes or variables. PP can be defined as the constrained optimization problem in (1), where the aim is to seek a $m$-dimensional projection space ($m < p$) (defined by the bases – columns – of $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m] \in \mathbb{R}^{p \times m}$) such that the projected data points in that space maximize a pre-defined objective function $\mathfrak{I}$, called the *projection pursuit index*. This function measures the degree of interestingness of the projected data. The constraint of orthonormality in $\mathbf{A}$ is necessary to ensure that each dimension in the target space shows different aspects of the data:

$$\mathbf{A}^* = \arg\max_{\mathbf{A}} \{\mathfrak{I}(\mathbf{XA})\}$$
$$\text{s.t } \mathbf{A}^T \cdot \mathbf{A} = \mathbf{I}. \tag{1}$$

### 2.1. Sequential projection pursuit

Sequential projection pursuit (SPP) [32] solves the PP problem in (1) by decomposing it into a sequence of $m$ optimization problems, each computing one base in $\mathbf{A}$.

The first base, $\mathbf{a}_1$, is obtained by searching a $p$-dimensional unit-length vector where the projected data $\mathbf{Xa}_1$ maximizes the one-dimension PP index $\mathfrak{I}$. Once $\mathbf{a}_1$ is found, SPP tries to remove all the information captured in that direction from the original data in order to avoid finding the same projection direction in subsequent iterations. For this task, the original SPP uses a "structure removal" procedure [14], which "Gaussianize" the data in the found direction, as follows: $\mathbf{X} = \mathbf{X} - \mathbf{Xa}_1\mathbf{a}_1^T$. The next base $\mathbf{a}_2$ is sought taking the updated $\mathbf{X}$ (also called *residual data*) as input data, subject to the constraint that $\mathbf{a}_2$ is orthogonal to $\mathbf{a}_1$. The process is iteratively repeated until all $m$ bases are obtained.

### 2.2. PP optimization

A key component in PP is the optimization process. Early approaches in this respect were based on the gradient techniques [30,29] and Newton–Raphson [31,37,14,13], where the projections are performed in at most three dimensions for visual exploratory tasks, the so-called exploratory projection pursuit (EPP). Further developments focused on developing more global methods for PP optimization, such as random search [38,39,29], genetic algorithm (GA) [32], random scan sampling algorithm (RSSA) [34], simulated annealing (SA) [21], particle swarm optimization (PSO) [35] and tribes [40]. In a previous work [33] we describe PPGA, a GA optimizer with a specialized crossover operator that often showed to find solutions better than those found by PSO, RSSA, and SA when used inside the SPP framework, reason why it is adopted for the present work.

Another important aspect in optimizing PP is how to ensure that each resulting dimension is associated with a different and complementary aspect of the data. Many PP methodologies, including SPP, address this task by using the "structure removal" procedure. However, it has been observed [41,38] that the successive application of this procedure (as done in the original SPP) may lead to data distortions, implying that an optimum found in residual data may not be longer related to any relevant aspect of the original data. Recently, Zhang and Chan [41,28] proposed an alternative approach to structure removal, which uses the *orthogonal complement space* concept.[1] In those works, the residual data is obtained projecting the current data onto the orthogonal complement of the found projection vector, which avoid data distortions and also ensures orthogonality of the projection bases.

[1] The orthogonal complement of one vector $x \in \mathbb{R}^n$ is the vector space $y$, all of which are orthogonal to $x$. Therefore, such space can be expanded by $n-1$ vector basis. That is, the orthogonal space of a vector $x$ $n$-dimensional is always dimensional size $n-1$.