Neural Networks 98 (2018) 154–177

Contents lists available at ScienceDirect

Neural Networks

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

Towards understanding sparse filtering: A theoretical perspective

Fabio Massimo Zennaro *, Ke Chen

School of Computer Science, The University of Manchester, Manchester, M13 9PL, UK

ARTICLE INFO

Article history:
Received 2 December 2016
Received in revised form 10 November 2017
Accepted 15 November 2017

Keywords:
Sparse filtering
Feature distribution learning
Soft clustering
Information preservation
Intrinsic structure
Cosine metric

ABSTRACT

In this paper we present a theoretical analysis to understand sparse filtering, a recent and effective algorithm for unsupervised learning. The aim of this research is not to show *whether* or *how well* sparse filtering works, but to understand *why* and *when* sparse filtering does work. We provide a thorough theoretical analysis of sparse filtering and its properties, and further offer an experimental validation of the main outcomes of our theoretical analysis. We show that sparse filtering works by explicitly maximizing the entropy of the learned representations through the maximization of the proxy of sparsity, and by implicitly preserving mutual information between original and learned representations through the constraint of preserving a structure of the data. Specifically, we show that the sparse filtering algorithm implemented using an absolute-value non-linearity determines the preservation of a data structure defined by relations of neighborhoodness under the *cosine distance*. Furthermore, we empirically validate our theoretical results with artificial and real data sets, and we apply our theoretical understanding to explain the success of sparse filtering on real-world problems. Our work provides a strong theoretical basis for understanding sparse filtering: it highlights assumptions and conditions for success behind this feature distribution learning algorithm, and provides insights for developing new feature distribution learning algorithms.

1. Introduction

Unsupervised learning deals with the *problem of modeling data*, stated as the problem of learning a transformation which maps data in a given representation onto a new representation. Contrasted with supervised learning, where we are provided labels and we learn a relationship between the data and the labels, unsupervised learning does not rely on any provided external semantics in the form of labels. In order to learn, unsupervised learning relies on the specification of assumptions and constraints that express our very understanding of the problem of modeling the data; for example, if we judge that a useful representation of the data would be provided by grouping together data instances according to a specific metric, then we may rely on distance-based clustering algorithms to generate one-hot representations of the data.

Often, the tacit aim of unsupervised learning is to generate representations of the data that may simplify the further problem of learning meaningful relationships through supervised learning. Coates, Ng, and Lee (2011) clearly showed that very simple unsupervised learning algorithms (such as *k*-means clustering), when properly tuned, can generate representations of the data that allow even basic classifiers, such as a linear support vector machine, to achieve state-of-the-art performances.

One common assumption hard-wired in several unsupervised learning algorithms is *sparsity* (for a review on the use of sparsity in representation learning see Bengio, Courville, & Vincent, 2013). Sparse representation learning aims at finding a mapping that produces new representations where few of the components are active while all of the others are reduced to zero. The adoption of sparsity relies both on biological analogies and on theoretical justifications (for discussion on the justification of sparsity see, for instance, Bengio et al., 2013; Földiák & Young, 1995; Ganguli & Sompolinsky, 2012; Olshausen & Field, 1997). Several state-of-the-art algorithms have been developed or have been adapted to learn sparse representations (for a recent survey of these algorithms, see Zhang, Xu, Yang, Li, & Zhang, 2015).

1.1. Sparse filtering and related work

In 2011, Ngiam, Chen, Bhaskar, Koh, and Ng (2011) proposed a novel unsupervised learning framework for generating sparse representations. Most of the successful unsupervised algorithms may be described as *data distribution learning* algorithms that try to learn new representations which better model the underlying probability distribution that generated the data. In contrast, they proposed the possibility of developing *feature distribution learning* algorithms that try to learn new representations having desirable

* Corresponding author.
E-mail addresses: zennarof@cs.manchester.ac.uk (F.M. Zennaro), chen@cs.manchester.ac.uk (K. Chen).

https://doi.org/10.1016/j.neunet.2017.11.010
0893-6080/

properties, without the need of taking into account the problem of modeling the distribution of the data.

Consistently with the feature distribution learning framework, they defined an algorithm named *sparse filtering*, which ignores the problem of learning the data distribution and instead focuses only on optimizing the sparsity of the learned representations. Sparse filtering proved to be an excellent algorithm for unsupervised learning: it is extremely simple to tune since it has only a single hyper-parameter to select; it scales very well with the dimension of the input; it is easy to implement; and, more importantly, it was shown to achieve state-of-the-art performance on image recognition and phone classification (Goodfellow, Erhan, Carrier, Courville, Mirza, Hamner, Cukierski, Tang, Thaler, Lee, Zhou, Ramaiah, Feng, Li, Wang, Athanasakis, Shawe-Taylor, Milakov, Park, Ionescu, Popescu, Grozea, Bergstra, Xie, Romaszko, Xu, Chuang, & Bengio, 2013; Ngiam et al., 2011; Romaszko, 2013). Thanks to its success and to the simplicity of implementing and integrating the algorithm in already existing machine learning systems, sparse filtering was adopted in many real-world applications (see, for instance, the works of Dong, Pei, He, Liu, Dong, & Jia, 2014; Lei, Jia, Lin, Xing, & Ding, 2015; Raja, Raghavendra, Vemuri, & Busch, 2015; Ryman, Bruce, & Freund, 2016).

Some studies have also provided sparse filtering with some biological support. Bruce, Rahman, and Carrier (2016) analyzed different biologically-grounded principles for representation learning of images, using sparse filtering as a starting point for the definition of new learning algorithms. Interestingly, Kozlov and Gentner (2016) used sparse filtering to model the receptive fields of high-level auditory neurons in the European starling, providing further support to the general hypothesis that sparsity and normalization are general principles of neural computation (Carandini & Heeger, 2012).

1.2. Problem statement

So far, sparse filtering has been successfully applied to many scenarios, and its usefulness repeatedly confirmed (see, for instance, its application in Dong et al., 2014; Han, Lee, Nam, & Lee, 2016; Liu, He, Xie, Gu, Liu, & Pei, 2016; Raja et al., 2015). In general, however, a clear theoretical explanation of the algorithm is still lacking. Ngiam et al. (2011) drew connections between sparse filtering, divisive normalization, independent component analysis, and sparse coding, while Lederer and Guadarrama (2014) provided a deeper analysis of the normalization steps inside the sparse filtering algorithm. However, the reasons why and on what conditions sparse filtering works are left unexplored. In this paper, we aim at understanding from a theoretical perspective *why* and *when* sparse filtering works. It is worth clarifying that our work does not concern itself with showing *whether* or *how well* well sparse filtering works, as there have been abundant evidence in literature on its successes in different real applications.

We begin by arguing that any unsupervised learning algorithm, in order to work properly, has to deal with the problem of preserving information conveyed by the probability distribution of the data. Given that feature distribution learning ignores the problem of learning the data distribution itself, a natural question arises: *how is the information conveyed by the data distribution preserved in feature distribution learning and, specifically, in sparse filtering?*

The actual success of sparse filtering suggests that the algorithm is indeed able to preserve relevant information conveyed in the distribution of the data. However, no explanation for this behavior has been given. We suggest that information may be preserved through the preservation of the structure of the data. To understand how this may be, we study the properties of the transformations within the algorithm and pose the following question: *is there any sort of data structure that is preserved by the processing in sparse filtering?*

Through a theoretical analysis we show that sparse filtering implemented using an absolute-value non-linearity does indeed retain information through the preservation of the data structure defined by the relations of neighborhoodness under the cosine distance. Relying on this, we investigate whether our theoretical results can be used to explain the success or the failure of sparse filtering in real applications. In particular we consider the following questions: *can the success of sparse filtering be explained in terms of the type of structure preserved? Can the failure of alternative forms of sparse filtering using different non-linearities be explained counterfactually on the grounds of information preservation? Is it possible to identify scenarios in which sparse filtering is likely to be helpful and other scenarios in which it is likely not to be useful?*

1.3. Contributions

We summarize the contributions made in this study as follows:

- We provide a theoretical analysis to understand why and when sparse filtering works. We show that the standard sparse filtering algorithm implemented with an absolute-value non-linearity implicitly works under the assumption of an intrinsic radial structure of the data. This assumption naturally makes the algorithm more suitable for certain data sets.
- We empirically validate our main theoretical findings, both on artificial data and real-world data sets.
- We provide useful insights for developing new feature distribution learning algorithms based on our theoretical understanding.

1.4. Organization

The rest of this paper is organized as follows. We first review the concepts and ideas forming the foundations of our work (Section 2). Next, we provide a formal theoretical analysis of the sparse filtering algorithm based on a rigorous conceptualization of feature distribution learning (Section 3). The theoretical results inform the following experimental simulations (Section 4). We then discuss the results we collected, in relation to sparse filtering, in particular, and to feature distribution learning, in general (Section 5). Finally, we draw conclusions by summarizing our contributions and highlighting future developments (Section 6).

To facilitate our presentation, Table 1 summarizes the notation system used in this manuscript.

2. Foundations

In this section we review basic concepts underlying our study. We provide a rigorous description of unsupervised learning, we present its formalization in information-theoretic terms, we formalize the property of sparsity, and, finally, we bring all these concepts together in the definition of the sparse filtering algorithm.

2.1. Unsupervised learning

Let $\mathsf{X} = \{\mathbf{X}^{(i)} \in \mathbb{R}^O\}_{i=1}^N$ be a set of N *samples* or *data points* represented as vectors in an O-dimensional space. We will refer to the given representation of a sample $\mathbf{X}^{(i)}$ in the space $\mathbb{R}^O$ as the *original representation* of the sample $\mathbf{X}^{(i)}$ and to $\mathbb{R}^O$ as the *original space*. From an algebraic point of view, we can formalize the data set as a matrix $\mathbf{X}$ of dimensions $(O \times N)$; from a probabilistic point of view, we can model the data points $\mathbf{X}^{(i)}$ as i.i.d. samples from a multivariate random variable $X = (X_1, X_2, \ldots, X_O)$ with pdf $p(X)$.

Unsupervised learning discovers a transformation $f : \mathbb{R}^O \to \mathbb{R}^L$ mapping the set X from an O-dimensional space to the set