Pattern Recognition Letters 53 (2015) 100–107

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Feature selection for unsupervised learning through local learning ☆

Jin Yao [a], Qi Mao [b], Steve Goodison [c], Volker Mai [d], Yijun Sun [b,e,*]

[a] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32610, USA
[b] Department of Microbiology and Immunology, State University of New York at Buffalo, Buffalo, NY 14201, USA
[c] Mayo Clinic, Jacksonville, FL 32216, USA
[d] Department of Epidemiology, Emerging Pathogens Institute, University of Florida, Gainesville, FL 32610, USA
[e] Department of Computer Science and Engineering, Department of Biostatistics, State University of New York at Buffalo, Buffalo, NY 14201, USA

ARTICLE INFO

Article history:
Received 2 March 2014
Available online 3 December 2014

Keywords:
Feature selection
Unsupervised learning
Clustering
Manifold learning

ABSTRACT

We consider the problem of feature selection for unsupervised learning and develop a new algorithm capable of identifying informative features supporting complex structures embedded in a high-dimensional space. The development of the algorithm is inspired by human learning in detecting complex data structures. We formulate it as an optimization problem with a well-defined objective function, and solve the problem by using an iterative approach. The algorithm can be easily implemented and is computationally very efficient. We use gap statistics to estimate the parameters so that the proposed method is completely parameter-free. We also develop a scheme based on permutation tests to estimate the statistical significance of the presence of a data structure. We demonstrate the effectiveness and versatility of the algorithm by comparing it with seven existing methods on a set of synthetic datasets with a wide variety of structures and cancer microarray gene expression datasets.

1. Background

The problem of unsupervised learning is that of trying to identify hidden structures in data. Due to the lack of label information, it is generally considered much more difficult than supervised learning. In applications involving high-dimensional data, the task becomes even more challenging since meaningful structures can be completely obscured by a large number of irrelevant features. A commonly used practice to alleviate the problem is to perform feature selection to remove irrelevant features to facilitate downstream analyses. In addition to defying the curse of dimensionality, eliminating irrelevant features can significantly reduce computational complexity and the cost of collecting irrelevant features. In some cases, it can also provide significant insights into the nature of the problems under investigation.

In processing high-dimensional data (e.g., biological data), unsupervised learning is commonly used for exploratory purposes. Before learning, we may only have limited knowledge on how data is distributed. It can be grouped into multiple but unknown numbers of clusters with arbitrary shapes, reside on multiple low-dimensional manifolds, encompass mixed data structures (e.g., clusters and manifolds), or may contain no structure at all (see Fig. 1). Our goal is to develop a *generic* feature selection algorithm capable of (1) identifying features supporting intrinsic geometry of high-dimensional data without explicitly assuming the form of data structures, and (2) providing us with statistics to indicate the absence of data structures if data does not contain any meaningful structure.

To the best of our knowledge, there is currently no effective method that can achieve the two goals. This paper presents a simple method that generally meets the above two requirements and addresses some of the limitations of existing methods.

1.1. Literature review

Feature selection for unsupervised learning is generally considered a much more difficult problem than that for supervised learning, due to the lack of label information that one can use to guide the selection of relevant features. Existing algorithms can be categorized as filter, wrapper or embedded methods. Filter methods are independent of learning algorithms and select useful features based on some statistical properties of data. Laplacian score [10] and SPEC [26] are two representative filter methods. Laplacian score weighs each feature according to its consistency with a Gaussian Laplacian matrix, and SPEC is a unified feature selection method for both supervised and unsupervised learning based on spectral graph theory. The two methods work well for low-dimensional data or data with a high signal-to-noise ratio (i.e., local structures are well preserved in the original space). However, when the number of irrelevant

☆ This paper has been recommended for acceptance by J. Laaksonen.

* Corresponding author. Tel.: +1 716 881 1374, +1 612 520 1208.
E-mail address: yijunsun@buffalo.edu (Y. Sun).

http://dx.doi.org/10.1016/j.patrec.2014.11.006
0167-8655/

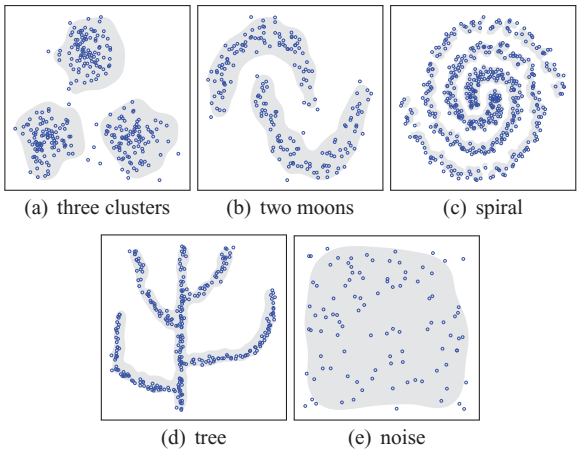

Fig. 1. Simulated datasets with a wide variety of data structures: (a) three clusters with regular shapes, (b) two moons with irregular shapes, (c) two twisted spirals, (d) tree, and (e) Gaussian noise. The shaded areas are the areas covered by the data.

features becomes excessively large, the assumption adopted by the two methods becomes invalid and both can perform poorly. In contrast to filter methods, a wrapper or embedded method uses the accuracy of a learning algorithm as a criterion to select useful features (see pioneering work by [7]). Since it is difficult to mathematically describe a complex data structure (e.g., manifolds), a clustering method is commonly used as the learning algorithm (e.g., MCFS [3], NDFS [13], RUFS [15], CGSSL [12] and sparse K-means [23]). However, as noted above, unsupervised learning is not *merely* limited to cluster analysis, and even if the clustering assumption holds, many methods require users to specify the number of clusters to be detected, which is generally unknown in advance in an exploratory data analysis. Moreover, the accuracy of a wrapper or embedded method to a large extent depends on the performance of a learning algorithm. For example, sparse K-means uses K-means as a base learner. It is well-known that K-means is prone to local minima, and consequently sparse K-means can perform poorly with the presence of copious irrelevant features.

2. Method

2.1. Motivation

Our goal is to develop a generic algorithm to identify relevant features supporting complex data structures hidden in a high-dimensional space. The essence is to define a criterion to *quantify* the presence of a data structure that can be easily optimized by using optimization techniques. Fig. 1 presents four toy examples of interest and Gaussian noise in a two-dimensional space. From the perspective of human learning, we say that the features in the first four examples provide more information than those in the last one. Without using a mathematical formulation to give a precise description of data, which is sometimes difficult or even impossible, a simple criterion we may use to quickly reach the above conclusion is that the areas of the blank space, not covered by data, in the first four subfigures are much larger than that in the noise case (since the data in the first four cases is tightly grouped). Then, a plausible approach to selecting informative features is to find a subspace where the areas of the blank space, or equivalently the difference between the total volume of the space spanned by data and the sum of the data volumes of clusters or those covered by manifolds (i.e., the shaded areas in the figures), is maximized. The problem now becomes how to define data volumes. For ease of presentation, we focus on clustering problems at the moment, but we will shortly see that our arguments hold for data with complex structures. Before learning, we do not know which data points belong to the same cluster. However, if data points are tightly grouped into clusters, the sum of the distances between data points and their nearest neighbors is small. Note that we herein do not require prior knowledge of the number of clusters and their shapes, and this idea works for other data structures (e.g., manifold) due to the *chaining effect*. However, finding a subspace where data groups tightly may lead to a trivial solution (all points would be collapsed into only one point). We also want the subspace spanned by data to be as large as possible. A natural idea is to maximize the average distance between data points and their mean vector. In a high-dimensional space, however, due to a phenomenon called distance concentration [17], samples tend to be closer to their center than to their nearest neighbors. This is an interesting but counter-intuitive phenomenon, manifesting the difficulty of high-dimensional data analysis due to the curse of dimensionality. To overcome the distance–concentration effect, we consider finding a subspace where the sum of the average distance of each sample to other samples is maximized. This can be understood as follows: if the territory of a country is large, the average travel distance from any city to all the other cities is large.

In the subsequent sections, we demonstrate how to formulate the problem of feature selection for unsupervised learning as an optimization problem and solve it using an iterative method. The proposed method is very simple but performs exceptionally well on a wide variety of complex data. Moreover, it does not require users to specify any parameter and the form of the structure one tries to detect.

2.2. Algorithm

Let $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N \subset \mathcal{R}^J$ be a training dataset. We seek to find a weighted subspace parameterized by $\mathbf{w}$ where the volume of the space spanned by the data is maximized and meanwhile the data is tightly grouped:

$$\max_{\mathbf{w}} \frac{1}{N}\sum_{n,i=1}^{N} d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w}) - \sum_{n=1}^{N} d(\mathbf{x}_n, \mathrm{NN}(\mathbf{x}_n)|\mathbf{w}),$$
$$\text{s.t. } \|\mathbf{w}\|_2^2 \le 1, \quad \mathbf{w} \ge \mathbf{0}, \tag{1}$$

where $\mathbf{w}$ is a non-negative feature weight vector, the magnitude of each component of which indicates the importance of the corresponding feature, and $\mathrm{NN}(\mathbf{x}_n)$ is the nearest neighbor of $\mathbf{x}_n$. The constraint $\|\mathbf{w}\|_2^2 \le 1$ prevents the objective function from increasing without an upper bound. $d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w})$ measures the distance between $\mathbf{x}_n$ and $\mathbf{x}_i$ with respect to $\mathbf{w}$. For numerical convenience, we use the block distance, which is also used in RELIEF [11] and LOGO algorithms [20] for feature selection in a supervised-learning setting. The above objective function has a close connection with that of K-means clustering, which is discussed in Section 2.5.

Since the nearest neighbor operator does not have an explicit form, it is difficult to directly solve (1) . To address the issue, we introduce a binary vector $\mathbf{p}_n \in \{0,1\}^{N-1}$ for each point $\mathbf{x}_n$. Computation of the distance between $\mathbf{x}_n$ to its nearest neighbor can be formulated as an optimization problem

$$d(\mathbf{x}_n, \mathrm{NN}(\mathbf{x}_n)|\mathbf{w}) = \min_{\mathbf{p}_n} \sum_{i\in\mathcal{M}_n} p_{ni} d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w}),$$
$$\text{s.t. } \sum_{i\in\mathcal{M}_n} p_{ni} = 1, \mathbf{p}_n \in \{0,1\}^{N-1},$$

where $\mathcal{M}_n = \{i : 1 \le i \le N, i \ne n\}$. Let $\mathcal{P} = \{\mathbf{p}_n\}_{n=1}^N$. Eq. (1) can then be transformed into the following optimization problem

$$\max_{\mathbf{w},\mathcal{P}} \frac{1}{N}\sum_{n,i=1}^{N} d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w}) - \sum_{n=1}^{N}\sum_{i\in\mathcal{M}_n} p_{ni} d(\mathbf{x}_n, \mathbf{x}_i|\mathbf{w})$$
$$\text{s.t. } \|\mathbf{w}\|_2^2 \le 1, \quad \mathbf{w} \ge \mathbf{0}, \sum_{i\in\mathcal{M}_n} p_{ni} = 1, \mathbf{p}_n \in \{0,1\}^{N-1}, \quad n = 1,\ldots,N, \tag{2}$$