Neurocomputing 150 (2015) 570–582

Contents lists available at ScienceDirect

Neurocomputing

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

Doubly supervised embedding based on class labels and intrinsic clusters for high-dimensional data visualization

Hannah Kim [a], Jaegul Choo [a,*], Chandan K. Reddy [b], Haesun Park [a]

[a] Georgia Institute of Technology, Atlanta, GA 30332, USA
[b] Wayne State University, Detroit, MI 48202, USA

ARTICLE INFO

Article history:
Received 2 December 2013
Received in revised form
20 September 2014
Accepted 25 September 2014
Available online 27 October 2014

Keywords:
Supervised dimension reduction
Multidimensional projection
T-distributed stochastic neighbor embedding
Visualization
Scatter plot
Clustering

ABSTRACT

Visualization of data can assist decision-making processes by presenting the underlying information in a perceptible manner. Many dimension reduction techniques have been proposed to generate faithful visualization snapshots given high-dimensional data. When class labels associated with the data are already provided, supervised dimension reduction methods, which utilize such pre-given label information as well as the data, have been effective in revealing the overall structure of data with respect to their pre-given class labels. However, the main principle of most of these supervised methods has been to enhance class separability, which generally leads to significant distortion of original relationships. To compensate for such distortion, we propose *a novel doubly supervised dimension reduction approach that highlights both natural groupings conforming to original relationships and classes determined by pre-given labels*. Our method imposes minimal supervision on the pre-given class information depending on their original distributions while imposing additional supervision on natural groupings to better preserve them in reduced feature space. Specifically, we apply the notion of doubly supervised dimension reduction to a state-of-the-art method called t-distributed stochastic neighbor embedding and present a new formulation and an algorithm. By performing both quantitative and qualitative analyses, we demonstrate the effectiveness of our method using various visualization examples on real-world data. Our results show that, compared to other existing methods, the proposed method better preserves the original high-dimensional relationships while simultaneously maintaining class separability and preserving cluster structures. In addition, due to the characteristics of preserving natural groupings, the visualization results generated by our method reveal interesting sub-groups that cohesively preserve the original relationships in the data.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

In the big data era, data are being collected easily in many settings in industry, science, and engineering. However, analysis on them is becoming more challenging than ever because of their complex nature and scale, often obfuscating the tasks to be solved. In these problematic situations, visualization can be helpful in facilitating decision-making processes by providing users with an overview of data.

Most real-world data are typically encoded as high-dimensional vectors as they can be effectively represented using a large number of features. One of the key methods for visualizing high-dimensional data in the form of a 2D/3D scatter plot is dimension reduction, and several well-known dimension reduction methods such as principal component analysis (PCA) [1] and multidimensional scaling (MDS) [2] have been widely applied in visualization applications. In general, the main idea behind most dimension reduction methods is to preserve the original high-dimensional relationships as much as possible in a lower-dimensional space. For example, PCA achieves this goal by maximizing the variance of the data in a low-dimensional space, and MDS tries to approximate all given pairwise similarity/distance values.

In many cases, however, additional information is available about the high-dimensional data. One of such additional information is class labels of the individual data points, indicating pre-given groupings of data. Unlike the previous unsupervised methods, which use only high-dimensional data as input, another type of methods called supervised dimension reduction utilizes such pre-given class labels when reducing the dimensions. Supervised dimension reduction, such as linear discriminant analysis (LDA) [3–5], has been successfully applied in numerous classification applications in machine learning and data mining (e.g., facial recognition [6]), and visual analytics [7,8]. Given these labels, supervised dimension reduction generally enhances class separation in lower-dimensional

* Corresponding author.
E-mail addresses: hannahkim@gatech.edu (H. Kim), jaegul.choo@cc.gatech.edu (J. Choo).

http://dx.doi.org/10.1016/j.neucom.2014.09.064
0925-2312/© 2014 Elsevier B.V. All rights reserved.

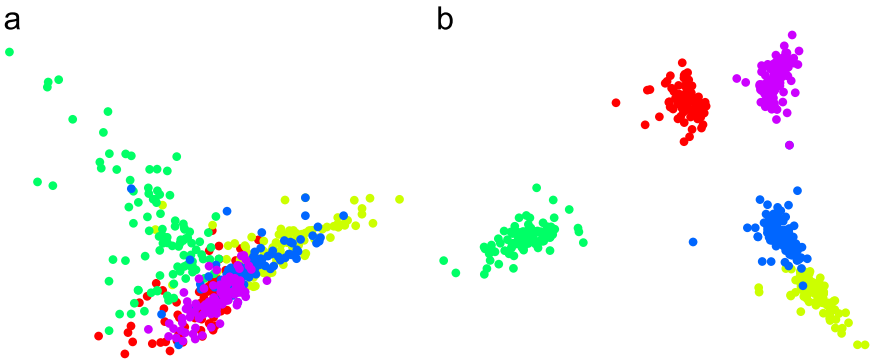

Fig. 1. Comparison of 2D visualization of Medline data described in Section 5.1. (a) Principal component analysis (PCA); (b) Linear discriminant analysis (LDA).

representations of data. Unlike unsupervised methods such as PCA, as shown in Fig. 1(a), supervised dimension reduction plays an important role in visualization by highlighting the class structure of data, as shown in Fig. 1(b). In this manner, supervised dimension reduction effectively performs the first step in a well-known visual information seeking mantra: Overview first, zoom and filter, then details-on-demand [9] by producing a visual overview with the class information highlighted.

Nonetheless, the two different criteria of separating classes and preserving original high-dimensional relationships could conflict with each other. For example, the widely used LDA aims at maximizing inter-class distances while minimizing intra-class ones, which could significantly distort the original relationships of the high-dimensional data. Although this issue might not be important in other applications such as classification, it could cause a problem in visualization tasks. That is, together with an overview at a class level, users would want the original relationships in data to be faithfully represented in visualization.

To effectively handle this trade-off problem in visualization applications, we propose a novel supervised dimension reduction approach. Basically, our proposed method incorporates a concept of *intrinsic clusters*, which takes into account natural groupings inherent in the data, to dimension reduction. As opposed to the classes that are externally formed by pre-given label information, intrinsic clusters are computed by a clustering algorithm purely based on the original high-dimensional relationships. Such intrinsic clusters provide a means to better preserve the original relationships in addition to the class separation capabilities available in the existing supervised dimension reduction methods.

The proposed method contains two important characteristics. First, we adaptively impose supervision on different classes depending on how clearly they are separated from the rest. In other words, we impose strong supervision on poorly separated classes so that they are visually distinct while imposing weak supervision on already well separated ones. In this manner, unnecessary distortion will be avoided. Second, we try to actively enhance the structure of intrinsic clusters by highlighting the separation between them. As a result, our method can properly capture original relationships while maintaining class separation in visualization. To realize our approach, we have chosen a state-of-the-art dimension reduction method, t-distributed stochastic neighbor embedding (t-SNE) [10], which has been applied successfully in various visualization applications, and we have extended it to what we call doubly supervised t-SNE.

The contribution of our work is summarized as follows:

- We introduce a novel concept of *double supervision* on dimension reduction based on pre-given class information as well as inherent clusters reflecting the natural groupings of data.
- We develop the formulation and algorithm of our novel dimension reduction method, doubly supervised t-SNE, which can separate pre-given classes as well as preserve the high-dimensional structure of the data.
- We evaluate the proposed method on various real-world data sets and demonstrate both quantitative and qualitative results.

The rest of the paper is organized as follows. Section 2 describes prior work related to dimension reduction. Section 3 introduces a widely used dimension reduction technique, t-SNE, as well as its basic extensions to supervised t-SNE. Next, Section 4 discusses the proposed methodology, and Section 5 presents our experiments. Finally, Section 6 concludes the paper.

2. Related work

Many dimension reduction techniques have been proposed in the past. The main goal of dimension reduction is to model high-dimensional data in a low-dimensional space such that the original information conveyed in a high-dimensional space is preserved as much as possible. Dimension reduction techniques attempt to achieve this goal by optimizing various objective and cost functions. For example, MDS [2] minimizes the sum of squared errors in terms of the pairwise distances of data items between high- and low-dimensional spaces. Isomap [11] works similar to MDS except that it uses geodesic pairwise distances approximated by the shortest path distances on k-nearest neighbor graphs instead of Euclidean pairwise distances of MDS. Another family of methods employs probabilistic formulation and objective functions. For example, stochastic neighbor embedding (SNE) [12] and t-distributed SNE (t-SNE) minimize the Kullback–Leibler divergence, a commonly used difference measure in probability, between the probability distributions derived from pairwise distance relationships in high- and low-dimensional spaces. However, these methods do not directly consider the original data grouping information, and thus they are called unsupervised methods.

Unlike the above-mentioned unsupervised methods, supervised methods (e.g., LDA [3]) assume that label information indicating the class structure in the data is already given and try to directly incorporate it into the dimension reduction process. In visualization, this grouping information has been actively used in various methods such as self-organizing maps (SOM) [13], where data clusters are naturally revealed during the dimension reduction process. To highlight the pre-given class structure, most supervised methods minimize distances within the same classes while maximizing those between different classes. In other supervised methods, a simple supervised extension of unsupervised methods via pre-given label information is to append the given label information as an additional dimension to the original high-dimensional representation of the data, which has been applied previously in SOM [13]. However, since the label information is generally represented as a numeric vector, if the scale of this label

Download English Version:

https://daneshyari.com/en/article/407548

Download Persian Version:

https://daneshyari.com/article/407548

Daneshyari.com