Semi-Supervised Dimension Reduction Techniques to Discover Term Relationships

Manuel Martín-Merino
University Pontificia of Salamanca, Spain

INTRODUCTION

The analysis of high dimensional datasets remains a challenging task for common machine learning techniques due to the well known 'curse of dimensionality' (Aggarwal, 2013; Cherkassky, 2007). It has been suggested in the literature (Kuman, 2006; Cevikalp, 2008) that the dimension reduction techniques can help to overcome this problem because they reduce the noise keeping the main structure of the dataset. Several algorithms have been proposed to this aim such as Principal Component Analysis (PCA), Correspondence Analysis or neural based techniques (see for instance (Kuman, 2006; Borg, 2005; Stuhlsatz, 2012)). In this article, we study two non-linear techniques, the Sammon mapping (Martín-Merino, 2004) and the Self Organizing Maps (SOM) (Kaski, 2006). Both have been widely applied to visualize term relationships.

Non-linear dimension reduction techniques have been applied to discover semantic relations among terms or documents in textual databases (Kaski, 2006). However, the algorithms proposed in the literature often have a low discriminant power, that is, different topics of the textual collection often overlap strongly in the projection. This is mainly due to the well known 'curse of dimensionality' (Aggarwal, 2013; Wang, 2013) and to the unsupervised nature of the algorithms proposed. Therefore, the projections are often useless to identify the different semantic groups in a given textual collection (Martín-Merino, 2005; Gönen, 2010).

Unfortunately, the words of a textual collection cannot be organized in a supervised manner, because no a priori classification of terms into topics is usually available (Martín-Merino, 2004). However, several search engines such as Yahoo provide a categorization for a small subset of documents (Martín-Merino, 2005; Manning, 2008) that may help to improve the discriminant power of the dimension reduction techniques. The semi-supervised dimension reduction techniques proposed in the literature (Gönen, 2010) cannot be applied to this problem because only the documents are categorized, not the terms.

Therefore new techniques should be developed that are able to reduce the term dimensionality considering the a priori categorization of a small subset of documents.

In this article we first study several unsupervised dimension reduction techniques that have been widely applied for textual data analysis. Next, we present a new semi-supervised versions of the Sammon mapping and SOM that profit from the document categorization carried out by human experts to reduce the term dimensionality. To this aim, rather than modifying the error function, a semi-supervised similarity is first defined that takes into account the document class labels. Next, the standard algorithms are applied to represent this dataset in a low dimensional space considering this semi-supervised similarity. Finally the algorithms presented have been tested using a real textual collection and have been exhaustively evaluated through several objective functions. The experimental results suggest that considering the categorization of a small subset of documents helps to improve the discriminant power of standard dimension reduction techniques.

BACKGROUND

Non-Linear Unsupervised Dimension Reduction Techniques

Non linear dimension reduction techniques help to discover term relationships missed by linear ones. They are able to represent the original dissimilarity matrix

DOI: 10.4018/978-1-4666-5888-2.ch721

in a space of smaller dimension. In this section we introduce briefly the Sammon mapping and the Self Organizing Maps (SOM).

The Sammon's mapping (Borg, 2005) was originally proposed as the non-linear map $\mathbb{R}^m \mapsto \mathbb{R}^d$ $d \ll m$ arising from the minimization of:

$$E = \frac{\sum_{i<j} \frac{(\delta_{ij} - d_{ij})^2}{\delta_{ij}}}{\sum_{i<j} \delta_{ij}}, \quad (1)$$

where δ_{ij} and d_{ij} are the inter-pattern distances in $\mathbb{R}^m$ and $\mathbb{R}^d$ respectively.

Notice that the numerator of E is weighed by δ_{ij}^{-1}. Therefore, the smaller dissimilarities will have more weight in the error function than the larger ones. This option is preferred over more drastic weights such as δ_{ij}^{-2} that do not help to achieve a balance between local and global structure preservation.

The error function is non linear and can be optimized by an iterative algorithm such as Newton's method.

The Self Organizing Maps (SOM) (Dhat, 2011) is a nonlinear visualization technique for high dimensional data. Input vectors are represented by neurons arranged according to a regular grid (usually 1D-2D) in such a way that similar vectors in input space become spatially close in the grid. Figure 1 shows a diagram of this neural network.

The results obtained by the SOM algorithm are equivalent to that obtained by optimizing the following energy function:

$$E(\mathcal{W}) = \sum_r \sum_{x_\mu \in V_r} \sum_s h_{rs} D(\mathbf{x}_\mu, \vec{w}_s) \approx \underbrace{\sum_r \sum_{x_\mu \in V_r} D(\mathbf{x}_\mu, \vec{w}_r)}_{\text{Quantization error}} + K \underbrace{\sum_r \sum_{s \neq r} h_{rs} D(\vec{w}_r, \vec{w}_s)}_{\text{C measure}} \quad (2)$$

where we have considered that the number of prototypes is large enough so that

$$D(\mathbf{x}_\mu, \vec{w}_s) \approx D(\mathbf{x}_\mu, \vec{w}_r) + D(\vec{w}_r, \vec{w}_s).$$

h_{rs} is a neighborhood function (for instance the Gaussian kernel) that transforms nonlinearly the neuron distances, D denotes the squared Euclidean distance and V_r is the Voronoi region corresponding to prototype $\vec{w}_r$ (Cherkassky, 2007).

Equation (6) shows that the SOM energy function may be decomposed as the sum of a quantization error and a $\mathcal{C}$ measure. The first one minimizes the information lost when the input patterns are represented by a set of prototypes. The second one maximizes the correlation between the prototype dissimilarities and the corresponding neuron distances in the grid of neurons. This is a multidimensional scaling algorithm that organize input neurons such that neighboring neurons in the grid correspond to prototypes that are close in input space.

The SOM energy function may be optimized by an iterative algorithm made up of two steps (Cherkassky, 2007):

1. A quantization algorithm is run that represents each pattern by the nearest neighbor prototype. This operation minimizes the first term in equation (6) and is equivalent to compute the Voronoi regions V_r. This step assigns each pattern x_μ to the Voronoi region V_r corresponding to the nearest neighbor prototype using the following equation: $r = \arg\min_s D(x_\mu, w_s)$.
2. The prototypes are organized along the grid of neurons by minimizing the second term in the function error. The optimization problem can be solved explicitly using the following adaptation rule for each prototype (Cherkassky, 2007):

$$\vec{w}_s = \frac{\sum_{r=1}^{M} \sum_{x_\mu \in V_r} h_{rs} \vec{x}_\mu}{\sum_{r=1}^{M} \sum_{x_\mu \in V_r} h_{rs}} \quad (3)$$

where M is the number of neurons and h_{rs} is for instance a Gaussian kernel of width $\sigma(t)$. The kernel width is adapted in each iteration using the rule proposed by (Cherkassky, 2007) $(\sigma(t) = \sigma_i(\sigma_f / \sigma_i)^{t/N_{iter}})$, where $\sigma_i \approx M/2$ is usually considered in the literature and σ_f is a parameter that determines the degree of smoothing of the principal curve generated by SOM.