Features Learning and Transformation Based on Deep Autoencoders

Eric Janvier[1], Thierry Couronne[1], and Nistor Grozavu[2](✉)

[1] Mindlytix, 33 Avenue Robert Andr Vivien, 94160 Saint-Mand, France
{e.janvier,t.couronne}@mindlytix.com
[2] LIPN CNRS UMR 7030, CNRS - Université Paris 13,
99, av. J-B Clement, 93430 Villetaneuse, France
Nistor.Grozavu@lipn.univ-paris13.fr

Abstract. Tag recommendation has become one of the most important ways of an organization to index online resources like articles, movies, and music in order to recommend it to potential users. Since recommendation information is usually very sparse, effective learning of the content representation for these resources is crucial to accurate the recommendation.

One of the issue of this problem is features transformation or features learning. In one hand, the projection methods allows to find new representations of the data, but it is not adapted for non-linear data or very sparse datasets. In another hand, unsupervised feature learning with deep networks has been widely studied in the recent years. Despite the progress, most existing models would be fragile to non-Gaussian noises, outliers or high dimensional sparse data. In this paper, we propose a study on the use of deep denoising autoencoders and other dimensional reduction techniques to learn relevant representations of the data in order to increase the quality of the clustering model.

In this paper, we propose an hybrid framework with a deep learning model called stacked denoising autoencoder (SDAE), the SVD and Diffusion Maps to learn more effective content representation. The proposed framework is tested on real tag recommendation dataset which was validated by using internal clustering indexes and by experts.

1 Introduction

Data mining, or knowledge discovery in databases (KDD), an evolving area in information technology, has received much interest in recent studies. The aim of data mining is to extract knowledge from data. The data size can be measured in two dimensions, the size of features and the size of observations. Both dimensions can take very high values, which can cause problems during the exploration and analysis of the dataset [12]. Models and tools are therefore required to process data for an improved understanding. Indeed, datasets with a large dimension (size of features) display small differences between the most similar and the least similar data. In such cases it is thus very difficult for a learning algorithm to detect the similarity of variables that define the clusters [9].

A. Hirose et al. (Eds.): ICONIP 2016, Part III, LNCS 9949, pp. 111–118, 2016.
DOI: 10.1007/978-3-319-46675-0_13

In hybrid methods, learning of item representations (also called item latent factors in some models) is crucial for the recommendation accuracy especially when the tag-item matrix is extremely sparse [15].

The main purpose of unsupervised learning methods is to extract generally useful features from unlabelled data, to detect and remove input redundancies, and to preserve only essential aspects of the data in robust and discriminative representations. Unsupervised methods have been routinely used in many scientific and industrial applications. In the context of neural network architectures, unsupervised layers can be stacked on top of each other to build deep hierarchies [7].

Unsupervised feature learning algorithms aim to find good representations for data, which can be used for different tasks i.e. classification, clustering, reconstruction, visualization,... Recently, deep networks such as stacked autoencoders (SAE) and diffusion maps (DM) have shown high feature learning performance [8].

Despite the progress, robust feature learning is still faced with challenges due to noise and outliers which are commonly appeared in the real-world data. In order to improve the antinoise ability of the deep networks, a new method was proposed by modifying the traditional stacked autoencoder to learn useful features from corrupted data and developed the stacked denoising autoencoder (SDAE) [8,14]. By corrupting the input data and using denoising criterion, the SDAE could learn robust representations and achieve good performance under different types of noises and to learn only the relevant features structure.

In this study, we focus on reducing the dimensions of the feature space as part of the unsupervised learning through different methods: Singular Value Decomposition (SVD), Diffusion Maps (DM) and Stacked Denoising Autoencoder (SDAE).

After transforming the features space, the new dataset will be clustered in order to detect relevant groups of tags which will be used furtherer for the recommendation. In this work a two-level topological clustering linked with the hierarchical clustering is used to visualize the results and to improve the computational time of the clustering model.

The rest of this paper is organized as follows: we present the proposed feature learning framework in Sect. 3 after introducing the feature transformation problem in Sect. 2. Section 3 introduces the use of the two-level topological clustering: the Self-Organizing Maps (SOM) with the Hierarchical Clustering which further is used for the tag recommendation. In the Sect. 4 we show the first experimental results on a real dataset. Finally we drew some conclusions and the possibilities of further research in this area.

2 Unsupervised Transformation of the Feature Space

Predictive models capable to classify new objects generally require learning by using labeled data. Unfortunately, only a small amount of labeled learning data may be available because of the cost of manual annotation of the data. Recent research has been focused on the use of large amounts of available unlabeled data, including: the transformation, the reduction of dimensionality, hierarchical representations of the variables ("deep learning"), kernel based learning, etc.

The unsupervised learning is often used for clustering data and rarely as a data preprocessing method. However, there are many methods that produce new data representations from unlabeled data. These unsupervised methods are sometimes used as a preprocessing tool for supervised or unsupervised learning models [4].

Given a data matrix represented as vectors of variables (p observations and n features), the goal of the unsupervised transformation of feature space is to produce another data matrix of dimension (p, n') (the transformed representation of n' new latent variables) or a similarity matrix between the data of size (p, p). Applying a model on the transformed matrix should provide better results compared to the original dataset.

The transformation of the feature space is done in two steps. First, we decompose the sparse data matrix using a normalization method and the SVD (Singular Value Decomposition). Then the matrix of latent variables obtained after this decomposition is used to learn the feature representation space using the Diffusion Maps and the SDAE method.

2.1 Matrix Decomposition and Normalization

The approximate factorization and tensor factorization (or decomposition) of a matrix have a main contribution in the improvement of data and the extraction of latent components. A common point for noisy detection, reduction of the model, the reconstruction of feasibility is to replace original data by an approximate representation of reduced dimensions obtained via a matrix factorization or decomposition. The concept of matrix factorization is used in a wide range of important applications and each matrix factorization is a different assumption about the components (factors) of matrices and their underlying structures, and this choice is an essential process in each application domain [4].

Very often, the datasets to be analyzed are nonnegative (or partially positive), and sometimes they also have a sparse representation. For these datasets, it is better to take into account these constraints in the analysis and to extract factors with physical meaning or a reasonable interpretation, and thus to avoid absurd or unpredictable results.

The singular value decomposition (SVD) treats the rows and columns in a symmetrical manner, and thus provides more information on the data matrix. This method also allows us to sort the information in the matrix so that, in general, the relevant part becomes visible. This property makes the SVD so useful in data mining and many other areas.

The bidiagonalisation GK (Golub-Kahan) method was originally formulated [3] for computing the SVD. This method can be also used to calculate a partial bidiagonalisation:

$$AQ_k = P_{k+1}B_{k+1}$$

where A is the data matrix, B_{k+1} are bidiagonal, and the clones Q_k and P_{k+1} are orthonormal.

With this decomposition, the approximations of singular values and singular vectors can be calculated similarly by tridiagonalisation. Indeed, it can be shown that the procedure of the GK bidiagonalisation is equivalent to applying the Lanczos tridiagonalisation on a symmetric matrix with a particular initial vector.

In our method we use this technique for the sparse data and Principal Component Analysis (PCA) is used for non-sparse datasets.

2.2 Diffusion Maps

The diffusion maps (DM) are based on defining a Markov random walk on the graph of the data. By performing the random walk for a number of timesteps, a measure for the proximity of the datapoints is obtained. The DM distance is defined using this measure. The key idea behind the diffusion distance is that it is based on integrating over all paths through the graph. This makes the diffusion distance more robust than, e.g., the geodesic distance employed in Isomap [7].

2.3 Deep Autoencoders

Denoising autoencoder (DAE) was proposed to overcome the limitations of autoencoders by reconstructing denoised inputs x from corrupted, noisy inputs $\tilde{x}$. DAEs avoids overfitting and learns better, non-trivial features by introducing stochastic noises to training samples. To generate corrupted inputs $\tilde{x}$ from their original value x it can be done with several different stochastic corruption criteria $q_D(\tilde{x}|x)$, including adding Gaussian random noise, randomly masking dimensions to zero, etc.

The objective function of DAEs remains the same as typical autoencoders. Note that the objective function minimizes the discrepancy between reconstructions and original, uncorrupted inputs x, not the corrupted inputs $\tilde{x}$ [6].

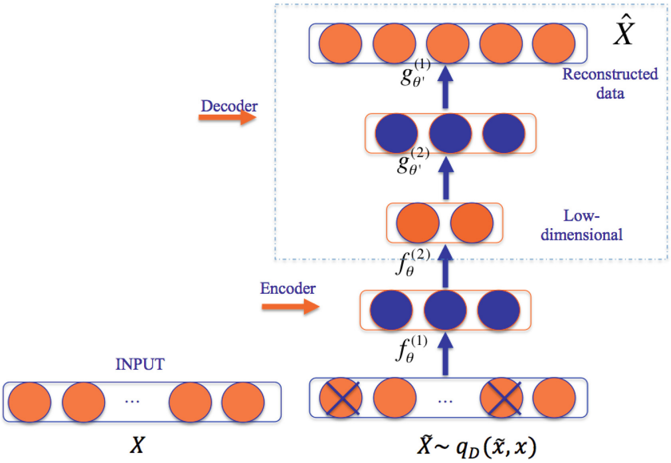

Fig. 1. Denoising autonecoder

Algorithme 1 . Transformation of the feature space and data coding

```
Inputs:
Learning (Training) data
Output:
New representation of the dataset)
Begin
1. Apply the diagonalization and factorization of the initial matrix (training data);
2. Apply the Diffusion Maps on the dataset;
3. Train the DSAE on the dataset;
4. Concatenate the obtained factors;
2-levels Clustering:
5. Construct the prototypes matrix using the SOM algorithm
6. Apply the hierarchical clustering on the prototypes map.
End
```

Stacking DAES (Fig. 1) on top of each other allows the model to learn more complex mapping from input to hidden representations. Just as other deep models including deep belief networks, training stacked DAEs is also done in twophase: layerwise, greedy pre-training and fine-tuning.

Unlike typical deep models that are extended by adding layers from bottom to top in pre-training, stacked DAEs are extended by adding layers in the middle of them. More specifically, the pre-training of stacked DAEs is done by the following steps. First, train bottom layer DAE with encoding function $y^{(1)} = f^{(1)}(x, \theta_f^{(1)})$ and decoding function $z^{(1)} = g^{(1)}(y^{(1)}, \theta_g^{(1)}$.

Train more DAEs in a similar way until the desired number of layers is achieved. After pre-training, the weights and biases of stacked DAE are fine-tuned by back-propagation as ordinary neural networks [6,7].

For a training dataset A, the first step of the proposed method is presented as following:

1. Normalization: $\widehat{A} = A * diag(std(A))^{\frac{1}{2}}$
2. Dimensionality reduction of the dataset $\widehat{A}$ by matrix factorisation: $svd(\widehat{A}) = [U_{\widehat{A}} S_{\widehat{A}} V_{\widehat{A}}]$
   For each column of $U_{\widehat{A}}$, $U_k = \frac{U_k}{\|U_k\|}$, where k is the number of retained eigenvectors

In the following (Algorithm 1) we present the proposed unsupervised learning algorithm for feature space transformation.

3 Topological Clustering

Topological learning is a recent direction in Machine Learning which aims to develop methods grounded on statistics to recover the topological invariants from the observed data points. Most of the existed topological learning approaches are based on graph theory or graph-based clustering methods. The topological learning is one of the most known technique which allow clustering and visualization simultaneously. At the end of the topographic learning, the “similar” data will be collect in clusters, which correspond to the sets of similar observations. These clusters can be represented by more concise information than the brutal listing of their patterns, such as their gravity center or different statistical moments. As expected, this information is easier to manipulate than the original data points. The neural networks based techniques are the most adapted to topological learning as these approaches represent already a network (graph) [5]. The models that interest us in this paper are those that could make at the same time the dimensionality reduction and clustering using Self-Organizing Maps (SOM) [10] in order to characterize clusters. SOM models are often used for visualization and unsupervised topological clustering. Its allow projection in small spaces that are generally two dimensional. Some extensions and reformulations of the SOM model have been described in the literature [1,5,11].

For map clustering we use traditional hierarchical clustering combined with Davides-bouldin index to choose optimal partition [13].

4 Experimental Results

This experiment is conducted with 3 datasets containing the description of 54000 web domains made with 800 topics. The 800 topics are extracted via a semantic analysis of words crawled on each domain, then the count of specific words is used to profile the domains by topic. The matrix is quite empty with a vast majority of domains qualified by very few topics. Topics are also classified into 50 groups according to level of similarity/dissimilarity.

Data set 1 contains all the domains/lines (54000) and all the topics/columns (800). Data set 2 (short) contains only the domains where largest number of topics/lines is informed (4910) and all the topics/columns (800). Data set 2 (short) contains only the domains where largest number of topics/lines is informed (4910) and topics/columns with highest weight (280).

Since clustering is an unsupervised process and most of theses algorithms are very sensitive to their initial assumptions, some evaluation is required to describe/analyze the clustering results [2]. Cluster validity represents the goodness measure of a clustering result relative to others created by other clustering algorithms, or by the same algorithm using different parameter values.

In general, there are three fundamental criteria to investigate the cluster validity: external criteria, internal criteria, and relative criteria. In the following we show main clustering validity indices.

Table 1 shows the quality of the clustering results in terms of the Davies-Bouldin index. It is easy to see that the proposed method outperforms the classical clustering for different numbers of clusters. The expert validated the results by indicated that the number of clusters should be 50, that means that the method proposed here outperforms a lot the clustering results for 50 clusters.

The same analysis can be made for the quality of the obtained topological map, where the quantization and topographic error decrease by using the deep learning features transformation (Table 2).

Table 1. Davies-Bouldin index obtained on the datasets

Method	Proposed model				Classical clustering			
nb cl.	5	15	30	50	5	15	30	50
dataset1	0.9259	0.7911	0.7301	0.7445	0.9736	0.8107	0.8005	0.7742
dataset2	0.5775	0.5870	0.7316	0.6867	0.5840	0.6776	0.7806	21.8992
dataset3	0.9177	0.8524	0.7511	6.5137	0.9682	0.9404	0.7911	27.8449

Table 2. Topological and quantization errors of the maps

Method	Proposed model		Classical clustering	
nb cl.	Quantization error	Topographic error	Quantization error	Topographic error
dataset1	0.56	0.41	3.245	0.73
dataset2	0.059	0.044	2.974	0.047
dataset3	0.124	0.142	3.436	0.071

5 Conclusion

In this work, we proposed k with a deep learning model called stacked denoising autoencoder (SDAE), the SVD and Diffusion Maps to learn more effective content representation. The transformed data was clustered using a two-level clustering model: SOM and hierarchical clustering to cluster the users web behaviour. The results on a real tag recommender dataset show that this approach aouperforms the classical clustering method and was also validated by the experts. As future works, we plan to test this method on different synthetic datasets and to compare it with other approaches. Also, some current work is made on the evaluation of the recommender system which use this approach.

References

1. Bishop, C.M., Svensén, M., Williams, C.K.I.: GTM: The generative topographic mapping. Neural Comput. **10**(1), 215–234 (1998)
2. Saporta, G.: Probabilits, analyse des donnes et statistiques. Editions Technip (2006)
3. Golub, G.H., Kahan, W.: Calculating the singular values and pseudo-inverse of a matrix. SIAM J. Numer. Anal. **2**, 205–224 (1965)
4. Grozavu, N., Bennani, Y., Labiod, L.: Feature space transformation for transfer learning. In: The 2012 International Joint Conference on Neural Networks (IJCNN), Brisbane, 10–15 June 2012, pp. 1–6 (2012)
5. Grozavu, N., Bennani, Y., Lebbah, M.: From variable weighting to cluster characterization in topographic unsupervised learning. In: Proceedings of International Joint Conference on Neural Network. IJCNN (2009)
6. Kang, L., Lee, K.T., Eun, J., Park, S.E., Choi, S.: Stacked denoising autoencoders for face pose normalization. In: Lee, M., Hirose, A., Hou, Z.-G., Kil, R.M. (eds.) Neural Information Processing. Theoretical Computer Science and General Issues, vol. 8227, pp. 241–248. Springer, Heidelberg (2013)

7. Van der Maaten, L., Postma, E., Van den Herik, H.: Dimensionality reduction: a comparative review. Technical report TiCC TR 2009–005 (2009)
8. Qi, Y., Wang, Y., Zheng, X., Wu, Z.: Robust feature learning by stacked autoencoder with maximum correntropy criterion. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, 4–9 May 2014, pp. 6716–6720 (2014). doi:10.1109/ICASSp.2014.6854900
9. Roth, V., Lange, T.: Feature selection in clustering problems. In: Thrun, S., Saul, L., Schölkopf, B. (eds.) Advances in Neural Information Processing Systems, vol. 16. MIT Press, Cambridge (2003)
10. Kohonen, T.: Self-organizing Maps. Springer, Heidelberg (2001)
11. Verbeek, J., Vlassis, N., Krose, B.: Self-organizing mixture models. Neurocomputing **63**, 99–123 (2005)
12. Verleysen, M., Francois, D., Simon, G., Wertz, V.: On the effects of dimensionality on data analysis with neural networks. In: Mira, J., Álvarez, J.R. (eds.) IWANN 2003. LNCS, vol. 2687, pp. 105–112. Springer, Heidelberg (2003). doi:10.1007/3-540-44869-1_14
13. Vesanto, J., Alhoniemi, E.: Clustering of the self-organizing map. IEEE Trans. Neural Netw. **11**(3), 586–600 (2000)
14. Vincent, P., Larochelle, H., Bengio, Y., Manzagol, P.A.: Extracting and composing robust features with denoising autoencoders. In: Proceedings of the 25th International Conference on Machine Learning. ICML 2008, pp. 1096–1103. ACM, New York (2008)
15. Wang, H., Shi, X., Yeung, D.Y.: Relational stacked denoising autoencoder for tag recommendation. In: Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence. AAAI 2015, pp. 3052–3058. AAAI Press (2015). http://dl.acm.org/citation.cfm?id=2888116.2888141