DATA CLEANING BY DEEP DICTIONARY LEARNING

Zhongqi Lu & Qiang Yang
Department of Computer Science and Engineering
the Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{zluab,qyang}@cse.ust.hk

ABSTRACT

The soundness of training data is important to the performance of a learning model. However in recommender systems, the training data are usually noisy, because of the randomness nature of users' behaviors and the sparseness of the users' feedback towards the recommendations. In this work, we would like to propose a noise elimination model to preprocess the training data in recommender systems. We define the noise as the abnormal patterns in the users' feedback. The proposed deep dictionary learning model tries to find the common patterns through dictionary learning. We define a dictionary through the output layer of a stacked autoencoder, so that the dictionary is represented by a deep structure and the noise in the dictionary is further filtered-out.

1 INTRODUCTION

Data noise has been a major cause of poor performances in recommender systems. Most recommender systems make recommendations based on the feedbacks of users, i.e., if a user has purchased an item, then the recommender system will compute for the user's interests based on this action. They then make their future recommendations based on predicted interests of the user.

However, there are various reasons that may cause the recorded user actions to be faulty. First, as shown in Campos et al. (2012), the user's account could be used to serve other users, such as the household members. This is especially true in the online shopping systems, where people purchase gifts for others. This leads to inaccurate feedbacks. Second, people could be misled occasionally to purchase things that are unrelated to their true interests, because of the bias of advertisements and the irrational behaviours on purchasing. Therefore, not all feedbacks can be used to reflect users' true intentions.

In this work, we organize our data in the form of transactions of items purchased. Our goal is to clean these transactions by removing some faulty ones, based on our judgement of whether the item is a "noise" to the user's purchasing history. The denoised purchasing history will be used as the training data for the existing learning based recommender systems.

As a preliminary study of the noise in transactions, we demonstrate several examples of typical noises from the TMall online shopping website `https://www.tmall.com/`, specifically we would like to focus on the women cloth category.

First, we consider one typical real world transaction to illustrate our work, as shown in Figure 1. There are four items purchased in one transaction. As we have seen from the tuple of items purchased by an user within certain time interval (and treated as a *transaction* in our setting), item 1, item 2 and item 3 are normally for the young ladies, while item 4 is clearly not. There would be multiple reasons for the exceptional purchasing of item 4 (e.g., the young lady purchased item 4 as a gift to her mother.), and it is desirable to remove the item 4 from the training data, when training a model to predict the user's interests.

The second step is to identify what should be defined as exceptional purchased items. Based on our previous definition of *transaction*, we make the basic assumption that the exceptional purchased items are not in the frequent purchasing combinations in users' transactions. And we verify this assumption by the following simple experiment: We calculate the co-occurrence of the items in the

Figure 1: A transaction in the online purchasing log. Item 4 is a "noise" in the transaction.

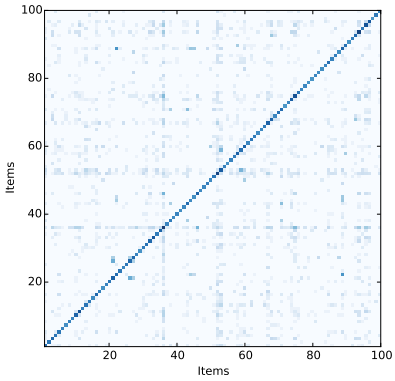

Figure 2: Visualization of the co-occurrence of purchasing some randomly selected items. Darker color indicates more co-occurrence.

transactions, and show the co-occurrence of purchasing a pool of selected items in Figure 2. In the figure, each row/column represents an item, and the darker pixel indicates the item represented by columns are more likely to be purchased together with the one represented by row. As shown in Figure 2, for each row (which represents one item), we can clearly see some pixels are darker than the others. This means if an item is purchased, then the items corresponding to the darker pixels are also likely to be purchased. The dark points clearly indicate that there exist some purchasing patterns statistically.

2 Deep Dictionary Learning

2.1 Problem Statement

To automatically filter out those items that cannot reflect users' true interests, we would like to introduce a data cleaning approach in online shopping recommendation scenario.

Traditionally the data are cleaned by some sophisticated rules designed by the domain experts and senior data engineers. However, other than the high costs of human efforts, the rules could be subjective to their designers' instinct. Besides, as more and more people using online shopping nowadays, the daily cumulated data in industrials usually measured by Tera-Bytes and the dimension of a data sample could be of thousands, which makes rule based solutions even impractical.

We look into the causes of the noise in online shopping data, and propose to clean up the data by a deep dictionary learning approach. Intuitively, the "noise" in online shopping systems could be

viewed as the minority purchased items in majority purchasing patterns. For example, if the lipstick is purchased along with shaver and shaving cream, we would consider the lipstick be a noise to this purchase record. Notice that we mark the lipstick to be the noise, instead of the shaver, based on the prior knowledge that shaving cream is usually purchased with shaver. However, there are millions of possibilities that no rules could cover all.

It is necessary to automatically discover the majority purchasing patterns as part of the data cleaning process. As the largest online shopping operator in China, we have already got the purchasing records for millions of active users. The purchasing records can be viewed as the most direct way to express the common sense of the public. In this work, the scale of items is millions, and we would like to discuss our method in a big data scenario.

We empirically define a transaction to be a collection of items purchased by a user within a certain time window $T \in \{[t_0, t_1), [t_1, t_2), \ldots\}$. In this work, because we are only focusing on a problem of item clustering, we would like to simplify the presentation of a transaction by omitting the time and user factors in the notation, i.e., a transaction is defined as $\mathbf{x} = \{x_i\}_{i=1}^N$, where $x_i = 1$ if any user purchased item x_i in any time window T; and $x_i = 0$ otherwise.

2.2 Dictionary Learning

The classical Dictionary Learning techniques Aharon et al. (2006); Lee et al. (2006) consider a finite training set of signals $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbf{R}^{m \times n}$ and optimize the empirical cost function

$$f_n(\mathbf{D}) = \frac{1}{n} \sum_{i=1}^{n} l(\mathbf{x}_i, \mathbf{D}), \tag{1}$$

where $\mathbf{D} \in \mathbf{R}^{m \times k}$ is the dictionary. Each column of the dictionary $\mathbf{D}$ represents a basis vector. And l is a loss function that would be small if $\mathbf{D}$ is "good" at representing the signal $\mathbf{x}$. As in Lee et al. (2006), we define $l(\mathbf{x}, \mathbf{D})$ as the optimal value of the l_1-sparse coding problem:

$$l(\mathbf{x}, \mathbf{D}) = \min_{\alpha \in \mathbf{R}^k} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\alpha\|_2^2 + \lambda \|\alpha\|_1, \tag{2}$$

where λ is a regularization parameter.

By learning a dictionary and using the sparse representations, we could partially eliminate some noise in the transactions. However, in the settings of recommender systems, the learned dictionary is still possible to be corrupted, because of the randomness of users' purchasing behaviors. To tackle the randomness issue, we assume a corruption while learning the dictionary. During the encoding step, we first corrupt the initial input $\mathbf{D}$ into $\widetilde{\mathbf{D}}$ by means of a stochastic mapping $\widetilde{\mathbf{D}} \sim p(\widetilde{\mathbf{D}} \mid \mathbf{D})$, and then encode $\widetilde{\mathbf{D}}$ by a deep neural network $f_\theta(.)$:

$$\mathbf{y} = f_\theta(\widetilde{\mathbf{D}})$$

where θ is the set of parameters in a deep neural net structure Le (2013); Lu et al. (2013). Then a decoding step is to reconstruct a clean "repaired" input from the corrupted version of it:

$$\mathbf{z} = g_{\theta'}(\mathbf{y})$$

where θ' is defined as the transpose of θ. Because the objective is denoising, we would like to see a set of parameters that is obtained by minimizing the average reconstruction error:

$$\theta = \arg\min \mathbf{L}_2(\mathbf{D}, \mathbf{z}) \tag{3}$$

where $\mathbf{L}_2$ is a square error loss.

Above all, in the deep dictionary learning model, we redefine $l(\mathbf{x}, \mathbf{D})$ in Equation (2):

$$l(\mathbf{x}, \mathbf{D}) = \min_{\alpha \in \mathbf{R}^k} \frac{1}{2} \|\mathbf{x} - \mathbf{D}\alpha\|_2^2 + \lambda \|\alpha\|_1 + \gamma \|\mathbf{D} - g_{\theta'}(f_\theta(\mathbf{D}))\|^2 \tag{4}$$

As discussed in Mairal et al. (2009), to prevent $\mathbf{D}$ from being arbitrarily large, which is resulted by arbitrarily small values of α, we adopt a constrain:

$$\mathcal{C} = \{\mathbf{D} \in \mathbf{R}^{m \times k} \quad \text{s.t. } \forall j = 1, \ldots, k,\ \mathbf{d}_j^T \mathbf{d}_j \leq 1\}.$$

Although the problem of minimizing the empirical cost $f_n(\mathbf{D})$ in Equation (1) is not convex with respect to $\mathbf{D}$, it can be optimized by alternatively minimizing over $\mathbf{D}$, α, or θ, while keeping the other one fixed, as proposed by Lee et al. (2006).

REFERENCES

Michal Aharon, Michael Elad, and Alfred Bruckstein. The k-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *Signal Processing, IEEE Transactions on*, 54 (11):4311–4322, 2006.

Pedro G Campos, Alejandro Bellogin, Fernando Díez, and Iván Cantador. Time feature selection for identifying active household members. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pp. 2311–2314. ACM, 2012.

Quoc V Le. Building high-level features using large scale unsupervised learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8595–8598. IEEE, 2013.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pp. 801–808, 2006.

Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *INTERSPEECH*, pp. 436–440, 2013.

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 689–696. ACM, 2009.