Non-negative Tri-factorization with a Dynamic Laplacian Constraint for Associating Features

Hongjie Zhai Makoto Haraguchi

We deal with the problem of embedding features from two different data sets into a common space for finding associated features. In the previous works, we have proposed a embedding framework based on Non-negative Tri-factorization and a laplacian constraint. In this framework, features are embedded into a common space where associated ones will have the same vector representation. However, the experiments we have conducted show that many unrelated features are also near to each other in the common space. The reason is that we can not use negative samples because of the limitation of Non-negative Tri-factorization. In this paper, to address this issue, we introduce the Dynamic Laplacian Constraint. That is, during the optimization process, we dynamically find new associated features and add them into the laplacian constraint. By this way, we can refine the common space step by step. Furthermore, when finding new associated features, we can introduce extra constraints to explicitly excluded the unrelated features. Experiments show the Dynamic Lapalcian Constraint can deliver quality improvements over our previous framework.

1 Introduction

In this paper, we study a problem of guessing feature association. That is, considering two object sets $O^1 = \{o_1^1, o_2^1, \ldots, o_n^1\}$ and $O^2 = \{o_1^2, o_2^2, \ldots, o_m^2\}$, where objects in O^1 are described by feature set $F^1 = \{f_1^1, f_2^1, \ldots, f_u^1\}$ and objects in O^2 are described by another feature set $F^2 = \{f_1^2, f_2^2, \ldots, f_v^2\}$. Some "hints" of associations, which are partial associations between two feature sets, are assumed already known. The target to finding the association between remaining features.

We proposed a novel method for finding the feature associations. To guess the missing associations, our method tries to find new associations by the knowledge from known associations (hints). Once new associations are detected, they can be added to known association set. Thus, we are able to use the new known association set to furtherly find new associations. By repeating this process, finally all the possible associated feature pairs can be detected. For the performance and scalability, we formulize this clustering-based methed by Non-negative Tri-factorization with a dynamic laplacian constraint.

2 Basic Idea and Algorithm

The basic idea of our method is illustrated in Figure 4. Here, we have two object-feature relation tables, where $f_1^1, f_2^1, f_3^1, \ldots$ are the features of objects $o_1^1, o_2^1, o_3^1, \ldots$. If a object contains the feature, the corresponding position will be 1 as shown in Figure 1. Additionally, we also have two associated feature pairs: $f_1^1 - f_1^2$ and $f_2^1 - f_2^2$. Firstly, we merge the associated features into one. After that, by only focus on the merged ones, we can construct the vector representation of objects with the same dimensions. For example, in Figure 2, f_1^1

ジェイ 泓杰, 原口 誠, 北海道大学情報科学研究科, Graduate School of IST, Hokkaido University.

and f_1^2 are merged as f_1 as well as that f_1^2 and f_2^2 are merged into f_2. With these constructed vectors, we perform clustering on objects. As the result, the objects in different sets may be clustered into one object cluster just like Figure 2. Here, cluster c_1 contains object o_1^1, o_3^1 and o_3^2 while cluster c_2 contains object o_2^1 and o_2^2. Moreover, by representing features with object clusters, we can perform clustering on features to find new associations. This is illustrated in Figure 3. It is easy to find that under the object cluster representation, f_4^1 and f_4^2 have the same vector. Thus, we found a new association $f_4^1 - f_4^2$. Once new associations are found, we merge into one and repeat the whole process until no new associations can be found.

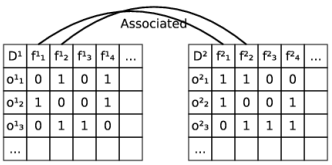

図 1 Merge associated features

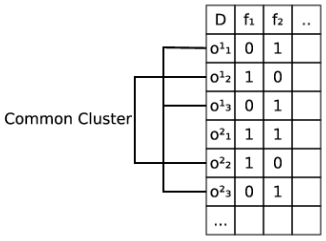

図 2 Clustering in common feature space

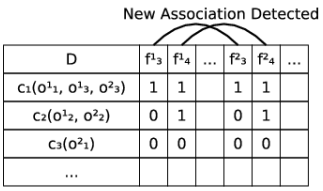

図 3 Mining new associations in common object cluster space

図 4 Illustration of Idea

Conclusively speaking, our method can be concluded as the following *two-phase* process:

- Use associated features to build the common space, clustering objects in this common feature space.
- Use object clusters to build the common space, clustering features in this common object space.

To improve the performance for large scale data, instead of the two-phases algorithm, we embed features into a common space by tri-factorization proposed by [3]. In this common space, the associated features are guarantee to have the same vector representation. Furthermore, we adapted a "dynamic" laplacian scheme to extend the hint set dynamically. We show the details in the algorithm 1.

Here, $D^1 \in \mathbb{R}+^{|O^1|\times|F^1|}$, $D^2 \in \mathbb{R}+^{|O^2|\times|F^2|}$ are the relation matrix, where $d_{ij} = 1$ if object o_i contains feature f_j. $W \in \mathbb{R}^{|F^1|+|F^2|\times|F^1|+|F^2|}$ is called feature relation matrix (laplacian). It is constructed from the following rules:

- if $i \leq |F^1| and j \leq |F^1|$, $w_{ij} = 0$
- if $i \geq |F^1| and j \geq |F^1|$, $w_{ij} = 0$
- if $i \leq |F^1| and j \geq |F^1|$, if f_i and f_j are associated, $w_{ij} = 1$, else $w_{ij} = 0$
- if $i \geq |F^1| and j \leq |F^1|$, $w_{ij} = w_{ji}$

By considering each feature as a vertex and connect the associated feature pairs, we can get a bigraph G. It is easy to know that the matrix K is the laplacian matrix of graph G. According to [2], the laplacian constraint can make sure the associated features always have similar vector representation in the common space. We should point out that the relation matrix (laplacian) will change according to the optimization. In each epoch, we select the nearest features pairs as the new association and add them to the laplacian. Thus, the laplacian is extended during optimization until all the feature are associated.

Data: Object-feature relation matrix: D^1, D^2, feature relation matrix: W, feature set F^1 and F^2, object set O^1, O^2, Dimension Parameter: N,M

Result: Associated feature pairs

Initialize random non-negative matrix L^1, C^1, R^1, L^2, C^2, R^2, where $C^{\{1,2\}} \in \mathbb{R}+^{N\times M}$;

while *features are remaining to be associated* **do**

  $K = D - W$, where $d_{ii} = \sum_j w_{ij}$;

  Solve the following tri-factorization problem: $argmin_{L^1,C^1,R^1,L^2,C^2,R^2}|D^1 - L^1C^1R^1| + |D^2 - L^2C^2R^2| + \lambda(\frac{R^1}{R^2})^T K(\frac{R^1}{R^2})$;

  Find $f_1 \in F^1$ and $f_2 \in F^2$ where $|f_1 - f_2|_F^2$ is minimum. Set $w_{i,|F^1|+j} = 1$.

end

Assign column vector in R^1 and R^2 to each feature in order;

for *each vector* v_j^1 *in* R^1 **do**

  Find the nearest vector v_j^2 in R^2;

  Print (F_i^1, F_j^2) as associated feature pair.

end

Algorithm 1: Tri-factorization for Association Learning

3 Experiment

To validate the ability of proposed method, we performed a preliminary experiment. We take the Spanish/Italian news articles between 1996-08-20 and 1996-08-25 from Reuters Corpora [1]. We randomly selected $4,000$ from both Spanish and Italian articles as the objects. After morphological analysis by tree-tagger [5], we only keep nouns as features. To generate the hints, we used google translate service. Each Spanish noun is translated into Italian and vice versa. We randomly selected several pairs as the hints. Because of the time limitation, the experiment result is still under preparation. We would like to report the details in oral representation.

4 Conclusion and Future Works

This paper proposed a general method for feature association guessing and give a tri-factorization formulation of the method. Our algorithm adapts a dynamic laplacian approach which dynamically extend the set of hints. This scheme allow us starting from a small set of known associations and extend them to all the dataset. However, there are still several problems remaining to be solved: (1) The convergence of Tri-factorization with dynamic laplacian still need to be investgated. (2) Experiments on large dataset (e.g. $20,000$ documents) still need to be done. (3) Dynamic laplacian needs selecting new associations during optimization. Thus, a guideline for selecting now associations should be designed.

参 考 文 献

[1] Lewis David D., et al. "Rcv1: A new benchmark collection for text categorization research." Journal of machine learning research 5.Apr (2004): 361-397.

[2] Cai Deng, et al. "Graph regularized nonnegative matrix factorization for data representation." IEEE Transactions on Pattern Analysis and Machine Intelligence 33.8 (2011): 1548-1560.

[3] Long Bo, Zhongfei Mark Zhang and Philip S. Yu. "Co-clustering by block value decomposition." Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. ACM, 2005.

[4] Lee Daniel D. and H. Sebastian Seung. "Algorithms for non-negative matrix factorization." Advances in neural information processing systems. 2001.

[5] Schmid, Helmut. "Probabilistic part-ospeech tagging using decision trees." New methods in language processing. Routledge, 2013.