LABEL WALKING NONNEGATIVE MATRIX FACTORIZATION

Long Lan, Naiyang Guan, Xiang Zhang, Xuhui Huang, Zhigang Luo**

Science and Technology on Parallel and Distributed Processing Laboratory,
College of Computer,
National University of Defense Technology, Changsha 410073, China

ABSTRACT

Semi-supervised learning (SSL) utilizes plenty of unlabeled examples to boost the performance of learning from limited labeled examples. Due to its great discriminant power, SSL has been widely applied to various real-world tasks such as information retrieval, pattern recognition, and speech separation. Label propagation (LP) is a popular SSL method which propagates labels through the dataset along high density areas defined by unlabeled examples, LP assumes nearby examples should share the same label, thus, it unavoidably pushes the labels to the wrong examples, especially when different labeled examples are not strictly separated. Seed K-means uses labeled examples to initialize class centers, and avoid getting stuck in poor local optima comparing to traditional K-means, however the hard constraint of each example's membership makes Seed K-means failed in many real world applications. This paper proposes a novel label walking nonnegative matrix factorization method (LWNMF) to handle labeled examples in SSL based on the framework of NMF. LWNMF decomposes the whole dataset into the product of a basis matrix and a coefficient matrix, and to travel labels to unlabeled examples, LWNMF regards the class indicators of labeled examples as their coefficients and iteratively updates both basis matrix and coefficients of unlabeled examples. Since LWNMF learns comprehensive class centroids, labels iteratively walk to unlabeled examples through these significant centroids.

***Index Terms*—** Nonnegative matrix factorization, Semi-supervised learning, Label propagation, K-means.

1. INTRODUCTION

Recently, semi-supervised learning (SSL) attracts significant attention [1], it has been widely studied and extends to various methods, among them, graph based semi-supervised learning method has shown its effectiveness in both theoretic and practical. Graph-based method constructs a graph to measure the similarity of examples, the vertices of the graph denote examples and the edges reflect the similarity of different examples. Label propagation (LP) [2] is an effective graph based semi-supervised learning method, it travels the labels under the assumption that the closely connected examples should share the same label, thus it is reasonable to use the graph to propagate the label to the unlabeled examples. Local and global consistency (LGC) [3] is another useful label propagation method, unlike standard LP, it propagates label in virtue of both similarity graph and initial label, that is, each example receives label from its neighbors and its initial state. Since standard LP and LGC depend heavily on the similarity graph, Wang *et al.* proposed Linear Neighborhood Propagation (LNP) [4] to construct graph, which assumes that each example can be linearly represented by its neighbors and similarities are measured via the reconstruction weights, LNP propagates the labels using the weight matrix. LNP avoids to the frustrating parameter selection of Euclidean space based graph, however, LNP also can not do well with the bridge points which connect different classes and misleads the labels propagating to the wrong direction.

K-means is the most widely used unsupervised clustering algorithm, whereas it fails in utilizing label information. Seeded K-means [5] and Constrained K-means [5] use labeled examples to initialize centers to improve clustering performance, these labeled examples are called seeds in the clustering. Seeded K-means updates labels of both seeds and unlabeled examples, while Constrained K-means keeps the labels of seeds fixed for each iteration round. with the help of labeled examples, both Seeded K-means and Constrained K-means achieve promising results. However, Seeded K-means and Constrained K-means strictly constrain the membership of examples, for each iteration round, every example is assigned to sole cluster, the exclusive assignment fails to find the multi-semantic of examples, for example, in document analysis, each document may simultaneously pertain to two or more topics, the hard assignment ignores the latent semantics of document.

Non-negative matrix factorization (NMF) [6], as a useful dimension reduction method, has attracted lots of attention, recently. NMF decomposes data matrix into two low-rank non-negative matrices, namely, the basis matrix and coefficient matrix. The solution of this factorization yields a natural parts-based representation for data. due to this favour decomposition, NMF has been widely used in information

*Zhigang Luo (Email: zgluo@nudt.edu.cn) and Naiyang Guan (Email: ny_guan@nudt.edu.cn) are the corresponding authors.

retrieval[7], pattern recognition [8] [9] [10] and speech separation [11]. Lee *et al.* [6] introduced NMF firstly in face representation and document clustering. The subsequent studies [7] [12] [13] show that the results of NMF can explain the clustering algorithm well, in clustering, the basis matrix can be considered as the clustering centers, and the elements of coefficient matrix are taken as the probabilities over the corresponding cluster centers, our previous work imposes normalization to NMF to get the representative clusters and achieve promising performance in image clustering [14] [15]. Ding *et al.* [16] have proved that constrained NMF equals to K-means theoretically, the objective function of K-means similar to NMF when impose orthogonal to rows of coefficient matrix.

In this paper, we propose label walking non-negative matrix factorization (LWNMF), LWNMF uses few label information to extract effective model. Label propagation pushes the labels based on the similarity graph while LWNMF travels labels using the basis matrix. LWNMF explains the label walking process in an innovative way. Specifically, LWNMF divides into two steps for each iteration round. In the first step, LWNMF learns basis matrix according to label of examples which corresponding to W updates of NMF. In the second step, LWNMF travels label to the unlabeled examples according to their distribution over basis matrix which corresponding to H updates. LWNMF holds the labels of labeled examples unchanged throughout the iteration rounds, which makes the obtained basis matrix more representative, thus labels walk smoothly. Also unlike K-means, LWNMF simultaneously learns the probabilities over different clusters for each example, it reveals the multi-topics of document example.

2. RELATED WORKS

Label propagation [2] is a simple graph based semi-supervised method, it assumes that close examples tend to have similar labels and examples labels propagate to neighboring examples according to their proximity. Assuming $V = [V_1, V_2, \cdots, V_n] \in R_+^{m \times n}$ is the given data matrix which can be taken as vertices of graph, and the followed matrix B is usually used to measure the similarity of every two examples, which is used to describe the edges of graph.

$$B_{ij} = exp(\frac{-\sum_{d=1}^{m}(v_i^d - v_j^d)^2}{\sigma^2}), \tag{1}$$

where σ is the control parameter, label propagate to unlabeled examples according to the weight of edges. Large edge weight propagate label easier while small edge weight propagate slowly. Labels travel through following iterations:

$$Y \leftarrow TY, \tag{2}$$

where T is the propagation matrix which is the normalization form of B, $Y \in R^{n \times c}$ is label matrix, where c denotes cluster number, it initializes by the labeled examples as following:

$$Y_{ij} = \begin{cases} 1 & if\ i\ is\ labeled\ with\ j \\ 0 & otherwise \end{cases}. \tag{3}$$

Formula (2) iterates until to convergence, the position of maximum of Y decides the examples label. Local and global consistency (LGC) [3] is also a popular label propagation method, it propagates label under local assumption and global assumption, LGC constructs similarity graph the same way with LP, however, LP clamps the labels of labeled examples to their initial label to propagate labels, while LGC directly receives label information from the initial label matrix. The propagation procedure of LGC can be summarized:

$$Y \leftarrow \alpha SY + (1 - \alpha)Y_0, \tag{4}$$

where $S = D^{(-1/2)}BD^{(-1/2)}$, B is defined as (1) and D is diagonal matrix whose element is the sum of each row of B, Y_0 is the initial label matrix, it is easily obtained by the definition of (3). α tradeoffs the label information receive from neighbors and initial state. LGC works well under the consistency assumption. Recently, Wang *et al.* proposed linear neighborhood propagation (LNP) [4], unlike LP, LNP assumes that every examples can be linearly reconstructed by its neighbors, and the reconstructed coefficients consist of the new weight matrix. LNP shows significant effectiveness and robustness to different datasets. However, similar to LP and LGC, it also faces to bridge point problem, bridge points lie between classes and usually lead to label propagate to wrong directions.

K-means is the most widely used unsupervised learning method, however, it initializes clustering centers randomly, and is unavoidable to get stuck in local optima and may achieve poor performance. Seeded K-means [5] and Constrained K-means [5] introduce labeled examples which called seeds to centers initialization and help reduce the chance of obtaining poor local optima. seeds are selected from all clusters evenly, after initializing the centers, Seeded K-means and Constrained K-means reassigned the label of each example according to its distance to the obtained centers. the center recalculated step and labels reassigned step iterate until to achieve stationary point. Seeded K-means and Constrained K-means share the same centers calculated step, however, they are different in labels assigned step, Seeded K-means reassigns the labels of all examples including the seeds, while the reassigned process of Constrained K-means only focuses on unlabeled examples, the labels of seeds keep their initial state throughout.

NMF is an popular unsupervised learning method, it decomposes an nonnegative data matrix into two low-rank nonnegative matrices [6]. To achieve this goal, NMF minimizes summation of the squared residues between the data matrix $V = [v_1, v_2, \cdots, v_n] \in R_+^{m \times n}$ and the product of basis matrix $W = [w_1, w_2, \cdots, w_c] \in R_+^{m \times c}$ and coefficient matrix

$H = [h_1, h_2, \cdots, h_n] \in R_+^{c \times n}$. c denotes a new dimension, generally $c \ll min(m, n)$.

$$f(W, H) = \min_{W,H} ||V - WH||_F^2, \quad s.t. W, H \geq 0, \quad (5)$$

where $||\bullet||_F^2$ denotes Frobenius norm. NMF has been proven theoretically equal K-means when imposing constraints of row orthogonality on coefficient matrix [16]. K-means strictly constrains the value of each label vector, allowing only one non-zero element. However, NMF relaxes this constraint, and just emphasizes its non-negativity of all elements. From this view, we can also see that K-means may easily fall into poor local optima since this discrete optimization. NMF shows good performance in clustering, when set c equal to the number of classes, it can explains the clustering directly. W signifies cluster centers whose columns are the centroids of every cluster, and H signifies the label matrix whose columns indicate the cluster membership of examples, as each column of H have many non-zero elements, NMF considers the index of maximal value as examples label.

3. LABEL WALKING NMF

In this section, we introduce label walking nonnegative matrix factorization, which discusses label walking in the view of NMF. LWNMF travels labels along with the optimization of NMF. Firstly, we regard label matrix as coefficient matrix, and the given label matrix guides the construction of basis matrix which can be considered as walking matrix, in the following updates, labels walk to the unlabeled part in virtue of the temporally learned basis matrix. Traditional label propagation method fixes its propagation matrix and label travels to neighbors according to proximity, whereas, LWNMF updates the walking matrix iteratively and label walks to unlabeled examples according to their distribution over the walking matrix. It is difficult for traditional label propagation to deal with the bridge examples, since local assumption do not considers the case of examples located in border, and bridge examples usually propagate label to the wrong direction and lead to a large crowd of errors [4]. However, LWNMF travels label according to the distribution of all examples, and well copy with bridge examples. We introduce the following objective for LWNMF:

$$\begin{aligned} f(W, H_u) = & \min_{W \geq 0, H_u \geq 0} \tfrac{1}{2}||V_l - WH_l||_F^2 \\ & + \tfrac{\lambda}{2}||V_u - WH_u||_F^2 \ + \tfrac{\sigma}{2}||W^T 1_m - 1_c||_F^2 \end{aligned}, \quad (6)$$

where λ tradeoffs the labeled part and unlabeled part, the first part and second part share the same W, this constrained basis matrix travels label information effectively. H_l is the label matrix of labeled examples, which is the transposing definition of (3), H_u can be regarded as soft label matrix of unlabeled examples, and each column is supposed to be the probabilities over different clusters for an example. The third term of (6) is used to column-normalized W, which makes sense for H_u, and will be explained in the following, σ is the balance parameter, $1_m \in R^m$ and$1_c \in R^c$ are vectors whose elements are all equal 1.

To keep the label probability interpretable, traditional LP must normalizes the label matrix in each iteration round. Fortunately, some simple normalization can implicitly make LWNMF interpretable to probability distribution. We column-normalize V in advance, this normalization shows favorable effects, as the following proposition:

$$\sum\nolimits_i V_{ij} = \sum\nolimits_i \sum\nolimits_k W_{ik} H_{kj} = \sum\nolimits_k (\sum\nolimits_i W_{ik}) H_{kj} = 1. \quad (7)$$

From (7) we know that for V, W, H, if $V = WH$, and both V and W are column-normalized, then H also column-normalized. The proposition explicitly explains the necessity of normalization of V. In practice, LWNMF normalizes each column of V at first, it is obvious that the minimization of the third term of (7) make W approximately column-normalized. And according to the proposition, the column-normalized W can be used to normalize H_u. The parameter λ and σ control the degree of normalization, Generally speaking, the larger of σ, the stronger of the normalization.

4. EXPERIMENT

In this section, we analyze labels travel results for standard LP [2], LGC [3], LNP [4], and semi-supervised K-means [5], and compare their performance in Reuters-21578 [1], TDT-2 [2]and WebKB[3] corpus. We introduce accuracy (AC) [17] to measure the clustering performance of LWNMF, AC compares the predicted label with the given label of unlabeled examples, for each example, if the propagated label is identical with its given label, the accuracy number adds 1, otherwise we ignore the incorrect propagation. AC is the ratio of accuracy number to number of all unlabeled examples u:

$$AC = \frac{sum(\delta(P, G))}{u}, \quad (8)$$

where $P \in R^u$ and $G \in R^u$ are the predicted label and ground truth respectively. $\sigma(P, G)$ denotes delta function that if P and G share the same value in the corresponding position then its value equals 1, and 0 otherwise. In practice, the predicted label vector P usually completely different from G, using above evaluation makes no sense. Hence before calculating AC, a map function should implement to match P to G, the P in (8) is the mapping result. It is obvious that AC increases with the prediction accuracy improving. Reuters-21578 consists of 21578 documents, and divides into 135 clusters. TDT-2 has 64527 documents, which grouped into 100 clusters. In our experiments, to meet the requirements

[1]http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html
[2]http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html
[3]http://www.cs.umd.edu/ sen/lbc-proj/LBC.html

of testing, we remove the documents which have multiple labels and leave these ones which have only one label, After this preprocessing, Reuters-21578 leaves 8213 documents and groups into 41 clusters. TDT-2 remains 10021 documents and can be categorized into 56 clusters. WebKB is collection of 842 web pages from 4 universities which can be divided into 5 clusters.

We compare the performance of LWNMF with standard label propagation algorithm (LP) [2], local and global consistency (LGC) [3], linear neighborhood propagation (LNP) [4], seeded K-means (SKM) [5] and constrained K-means (CKM) [5] in different cluster number, the number of cluster varies from 2 to 10 on both Reuters-21578 and TDT-2, WebKB varies the cluster number from 2 to 5. To valid the clustering results, we repeat the above algorithms on 50 different cluster sets for different cluster number situations and calculate the averages. In testing, five labeled examples are randomly selected for each cluster. On the other hand, we vary label size of each cluster to see the ability of LWNMF to utilize label information, label size varies from 1 to 10 means we select 1 to 10 labeled examples for each cluster in the experiments. In the label size experiment, we keep the number of clusters to 10 on Reuters-21578 and TDT-2, and 5 on WebKB, the label size results are also the average of 50 performances using different cluster sets. In experiment, we find the parameters of λ and σ can be set in a wide range, and experientially speaking, λ should less than the quantity ratio of labeled examples to unlabeled examples. NMF and K-means are also implemented as the benchmark. Figure 1 (a), (c) and (e) show the performance of LWNMF versus different cluster number on Reuters-21578, TDT-2 and WebKB respectively. In experiments, we randomly label five examples for each cluster. From the figures, the benchmark methods of NMF and K-means obtain the similar results on Reuters-21578, TDT-2 and WebKB, and both them are far poor than the rest of methods without the labeled examples, the few labeled examples are really helpful. It is obvious that LWNMF outperforms other methods on Reuters-21578 and WebKB for different cluster numbers, and is comparable with SKM, CKM, LP on TDT-2. The curves of AC suppose that LWNMF effectively travels labels to unlabeled examples. Seeded K-means and Constrained K-means almost achieve the same result from the pictures, which can be well explained since there is no wrongly labeled seeds. Figure 1 (b), (d), (f) are the results of varying label size, we can see that LWNMF shows the best performance comparing to the rest algorithms in different label size on both Reuters-21578 (Figure 1 (b)) and TDT-2 (Figure1 (d)). LWNMF on WebKB (Figure1 (f)) is worse than other algorithms in small label size, however, it is comparable to the best Semi-supervised K-means when labeled examples is enough. From all these three pictures, it is reasonable to believe that LWNMF can travel labels to unlabeled examples effectively, even there are few available labels.

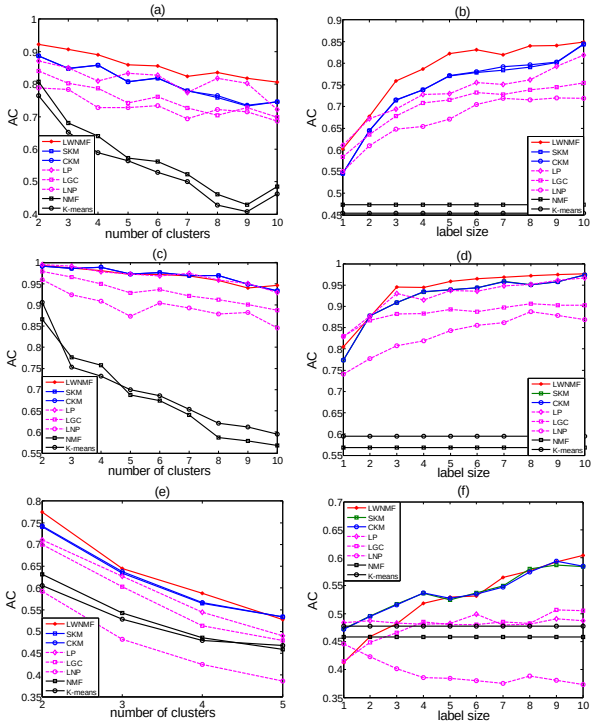

Fig. 1. clustering results on Reuters{(a),(b)}, TDT-2{(c),(d)} and WebKB{(e),(f)}

5. CONCLUSION

We have introduced a novel semi-supervised non-negative matrix factorization method in this paper, called label walking non-negative matrix factorization, which shows promising performance on document clustering. LWNMF considers coefficient matrix as label matrix, and learns the basis matrix under the constraints of label information and the whole examples, we regard the basis matrix as walking matrix and labels effectively walk to unlabeled examples according to the their distributions. The normalization of LWNMF makes label matrix probability interpretable and takes favorable effect in the label propagation. The experiments on three document corpora have demonstrated the effectiveness of our algorithm when comparing to the mentioned classical semi-supervised learning methods.

6. ACKNOWLEDGEMENT

This work is partially supported by Research Fund for Doctoral Program of Higher Education of China, SRFDP (under grant No. 20134307110017) and Plan for Innovative Graduate Student at National University of Defence Technology.

7. REFERENCES

[1] Xiaojin Zhu, "Semi-supervised learning literature survey," 2005.

[2] Xiaojin Zhu and Zoubin Ghahramani, "Learning from labeled and unlabeled data with label propagation," Tech. Rep., Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

[3] Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.

[4] Fei Wang and Changshui Zhang, "Label propagation through linear neighborhoods," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, no. 1, pp. 55–67, 2008.

[5] Sugato Basu, Arindam Banerjee, and Raymond Mooney, "Semi-supervised clustering by seeding," in *In Proceedings of 19th International Conference on Machine Learning*. Citeseer, 2002.

[6] Daniel D Lee and H Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[7] Wei Xu, Xin Liu, and Yihong Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 267–273.

[8] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han, "Non-negative matrix factorization on manifold," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 63–72.

[9] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan, "Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent," *Image Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 2030–2048, 2011.

[10] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan, "Non-negative patch alignment framework," *Neural Networks, IEEE Transactions on*, vol. 22, no. 8, pp. 1218–1230, 2011.

[11] Naiyang Guan, Long Lan, Dacheng Tao, Zhigang Luo, and Xuejun Yang, "Transductive nonnegative matrix factorization for semi-supervised high-performance speech separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 2534–2538.

[12] Yanhua Chen, Manjeet Rege, Ming Dong, and Jing Hua, "Non-negative matrix factorization for semi-supervised data clustering," *Knowledge and Information Systems*, vol. 17, no. 3, pp. 355–379, 2008.

[13] Naiyang Guan, Dacheng Tao, Zhigang Luo, and Bo Yuan, "Nenmf: an optimal gradient method for nonnegative matrix factorization," *Signal Processing, IEEE Transactions on*, vol. 60, no. 6, pp. 2882–2898, 2012.

[14] Long Lan, Naiyang Guan, Xiang Zhang, Dacheng Tao, and Zhigang Luo, "Soft-constrained nonnegative matrix factorization via normalization," in *Neural Networks (IJCNN), 2014 International Joint Conference on*. IEEE, 2014, pp. 3025–3030.

[15] Naiyang Guan, Dacheng Tao, Long Lan, Zhigang Luo, and Yang Xuejun, "Activity recognition in still images with transductive non-negative matrix factorization," in *The 6th International Workshop on Video Event Categorization, Tagging and Retrieval towards Big Data*, 2014.

[16] Chris HQ Ding, Xiaofeng He, and Horst D Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering.," in *SDM*. SIAM, 2005, vol. 5, pp. 606–610.

[17] Haifeng Liu, Zhaohui Wu, Xuelong Li, Deng Cai, and Thomas S Huang, "Constrained nonnegative matrix factorization for image representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1299–1311, 2012.