ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/332595797

Semi-supervised Clustering with Deep Metric Learning

Chapter · April 2019

DOI: 10.1007/978-3-030-18590-9_50

CITATIONS
0

READS
104

6 authors, including:

Hongzhi Yin
The University of Queensland
122 PUBLICATIONS **2,248** CITATIONS

SEE PROFILE

Ke Zhou
Huazhong University of Science and Technology
146 PUBLICATIONS **980** CITATIONS

SEE PROFILE

Hongxu Chen
The University of Queensland
12 PUBLICATIONS **145** CITATIONS

SEE PROFILE

Shazia Sadiq
The University of Queensland
199 PUBLICATIONS **4,642** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

A Framework for Socialisation of Work Practice for Improving Business Performance View project

AI for storage View project

All content following this page was uploaded by Ke Zhou on 09 August 2019.

The user has requested enhancement of the downloaded file.

Guoliang Li · Jun Yang ·
Joao Gama · Juggapong Natwichai ·
Yongxin Tong (Eds.)

LNCS 11448

Database Systems for Advanced Applications

DASFAA 2019 International Workshops: BDMS, BDQM, and GDMA
Chiang Mai, Thailand, April 22–25, 2019, Proceedings

Springer

Semi-supervised Clustering with Deep Metric Learning

Xiaocui Li[1], Hongzhi Yin[2(✉)], Ke Zhou[1(✉)], Hongxu Chen[2], Shazia Sadiq[2], and Xiaofang Zhou[2]

[1] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China
{LXC,k.zhou}@hust.edu.cn
[2] School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, QLD, Australia
{h.yin1,hongxu.chen}@uq.edu.au, {shazia,zxf}@itee.uq.edu.au

Abstract. Semi-supervised clustering has attracted lots of reserach interest due to its broad applications, and many methods have been presented. However there is still much space for improvement, (1) How to learn more discriminative feature representations to assist the traditional clustering methods; (2) How to make use of both the labeled and unlabelled data simultaneously and effectively during the process of clustering. To address these issues, we propose a novel semi-supervised clustering based on deep metric learning, namely SSCDML. By leveraging deep metric learning and semi-supervised learning effectively in a novel way, SSCDML dynamically update the unlabelled to labeled data through the limited labeled samples and obtain more meaningful data features, which make the classifier model more robust and the clustering results more accurate. Experimental results on Mnist, YaleB, and 20 Newsgroups databases demonstrate the high effectiveness of our proposed approach.

Keywords: Clustering · Semi-supervised learning · Deep metric learning

1 Introduction

Data mining has been a research hotspot for decades and many methods have been proposed [1–4]. However, how to extract useful features and learn an appropriate metric for high-dimensional data without any supervised information is a challenging task. Consequently, some supervised clustering algorithms were proposed to improve the clustering effect, and they indeed achieved limited achievements. These methods have great limitations in real practical applications, as it is almost impossible for all data to have labels, and tagging each data manually is a waste of human resources and time, and is unrealistic. In fact in most of real-world applications, we can only obtain quite limited labeled data. Based on the above problems, semi-supervised based clustering methods [7] emerged.

© Springer Nature Switzerland AG 2019
G. Li et al. (Eds.): DASFAA 2019, LNCS 11448, pp. 383–386, 2019.
https://doi.org/10.1007/978-3-030-18590-9_50

Although the existing semi-supervised clustering algorithms have achieved good results, there are still two important issues that will hinder the performance of clustering. (i) Most of these methods extract features or learn a distance metric through traditional SVM, neural networks or linear mapping, which limits its performance. (ii) They only use the labeled data to guide the process of the clustering, which can not be full used of the data especially unlabeled data. Motivated by the above analysis, we propose a semi-supervised clustering with deep metric learning (SSCDML), which can extract more discriminative features by using deep learning technique with nonlinear transformation, and simultaneously make full use of all data by combining semi-supervised learning.

2 Proposed Method

2.1 Semi-supervised Deep Metric Learning and Classification Network

We design a semi-supervised deep metric learning and classification network. The main training process of the network consists of the following three steps.

Step 1: First, extract discriminable features through CNNs, then use the features to train a classifier. We design a new loss function for semi-supervised deep metric learning and classification network as follows:

$$\min L = L_m + \lambda_1 L_c + \lambda_2 \|W\|_F^2, \tag{1}$$

where, λ_1 and λ_2 are a tunable positive parameter. $\|W\|_F^2$ is a regular term to prevent overfitting. L_m and L_c are metric learning loss and classification loss, respectively. They can be computed as follows:

$$L_m = \frac{1}{N}\sum_{i=1}^{N}(Y_i D(G(X_1), G(X_2)) + (1 - Y_i) max(\mu - D(G(X_1), G(X_2)), 0)), \tag{2}$$

where $D(,)$ is the Euclidean Distance function and $G(.)$ represent the outputs of the feature extracting network. $Y_i \in \{0, 1\}$ is the label of input pair samples.

$$L_c = -\sum_{G(X)} p(G(X)) \log q(G(X)), \tag{3}$$

where $p(.)$ is the expected outputs, and $q(.)$ is the actual outputs of the classification network.

Step 2: Encode the labeled and unlabeled data. Assume that $S_1 = \{(s_{1i}, l_{1i})|i = 1, 2, \ldots, N_1\}$ and $S_2 = \{(s_{2i}|i = 1, 2, \ldots, N_2\}$ separately represent the init labeled data and unlabeled data, where N_1 is the number of labeled samples, N_2 is the number of unlabeled samples, and $l_{1i} \in \{1, 2, \ldots, C\}$, where C is the number of classes. We use $S_1' = \{s_{1i}'|i = 1, 2, \ldots, N_1\}$ and $S_2' = \{s_{2i}'|i = 1, 2, \ldots, N_2\}$ represent the outputs of the S_1 and S_2 by CNNs.

Step 3: Tag the unlabeled data according to the classification network. Therefore, S_2 can be denoted as $S_2' = \{(s_{2i}', l_{2i}^1)|i = 1, 2, \ldots, N_2\}$, where l_{2i}^1 is the classification label of the s_{2i}'.

2.2 Semi-supervised Clustering Labeling Propagation Network

To acquire the strong label of the unlabeled data, we design a semi-supervised labeling propagation network. It includes two parts: semi-supervised clustering and labeling propagation.

In the process of the semi-supervised clustering, we applied the traditional k-means clustering algorithm to the data and mark the S_2' according to the clustering results, and record as $S_2' = \{(s_{2i}', l_{2i}^2)|i = 1, 2, \ldots, N_2\}$, where l_{2i}^2 is the clustering label of the s_{2i}'.

When both the classification label and clustering label of the unlabeled data S_1 are obtained, we can implement labeling propagation strategy.

3 Experiments

3.1 Datasets and Compared methods

We implement experiments on three publicly available datasets including: Mnist, YaleB [5] and 20 Newsgroups [6] and compare our approaches with some state-of-the-art related methods including: FSLSC [7], DFCM [8]. To evaluate the performance of our proposed methods and compared methods, we use clustering accuracy (AC):

$$AC = \frac{1}{N} \sum_{i=1}^{K} \max(C_i|L_i), \tag{4}$$

Table 1. Clutering results of proposed methods and three semi-supervised clustring methods with different percentages of labeled data on Mnist, YaleB and 20 newsgroups datasets.

Datasets	Percentages	FSLSC	DFCM	SSCDML
Mnist	5%	69.0 ± 1.1	87.4 ± 1.6	$\mathbf{89.6 \pm 1.3}$
	10%	75.2 ± 0.9	90.4 ± 1.6	$\mathbf{92.3 \pm 1.3}$
	20%	85.6 ± 1	93.2 ± 0.8	$\mathbf{94.2 \pm 1.1}$
YaleB	5%	52.8 ± 1.3	68.8 ± 1.5	$\mathbf{75.7 \pm 1.5}$
	10%	58.7 ± 1.8	73.8 ± 2.3	$\mathbf{78.3 \pm 1.3}$
	20%	66.4 ± 1	77.9 ± 1.8	$\mathbf{81.2 \pm 0.9}$
20 Newsgroups	5%	$29.1 \pm 0.$	50.3 ± 2.2	$\mathbf{53.5 \pm 1.1}$
	10%	38.6 ± 1.6	52.7 ± 2.2	$\mathbf{57.1 \pm 1.4}$
	20%	46.2 ± 2.3	56.2 ± 1.5	$\mathbf{60.2 \pm 1.2}$

3.2 Results and Analysis

In this subsection, we conduct experiment to evaluate the clustering performance of our proposed SSCDML approach. To evaluate the clustering performance of our proposed approach, we increase the percentage of labeled data from 5% to 20%. Table 1 report the AC results of our proposed method and three semi-supervised clustering methods on three datasets. We can obviously see that our SSCDML approach performs better than compared semi-supervised clustering methods.

4 Conclusion

In this paper, we propose a novel semi-supervised clustering with deep metric learning approach named SSCDML. SSCDML comprises a semi-supervised deep metric learning network and a labeling propagation network. Semi-supervised deep metric learning network can extract more powerful features, and then learn a more discriminative metric. After that, labeling propagation network is used to label new data. Experimental results on Mnist, YaleB and 20 Newsgroups datasets have shown the high performance and effectiveness of our SSCDML approach.

Acknowledgement. This work was supported by ARC Discovery Early Career Researcher Award (DE160100308) and ARC Discovery Project (DP170103954; DP190101985).

References

1. Yin, H. Zou, L. et al.: Joint event-partner recommendation in event-based social networks. In: 34th International Conference on Data Engineering (2018)
2. Yin, H. Wang, Q. et al.: Social influence-based group representation learning for group recommendation. In: 35th ICDE (2019)
3. Chen, H. Yin, H. et al.: PME: projected metric embedding on heterogeneous networks for link prediction. In: The 2018 ACM SIGKDD(2018)
4. Xie, M. Yin, H. et al.: Learning graph-based POI embedding for location-based recommendation. In: The 25th ACM CIKM (2016)
5. Cui, G., Li, X., Dong, Y.: Subspace clustering guided convex nonnegative matrix factorization. Neurocomputing **292**, 38–48 (2018)
6. Chen, G.: Deep learning with nonparametric clustering. arXiv preprint arXiv:1501.03084 (2015)
7. Guan, R., Wang, X. et al.: A feature space learning model based on semi-supervised clustering. In: IEEE International Conference on CSE (2017)
8. Arshad, A., Riaz, S., et al.: Semi-supervised deep fuzzy c-mean clustering for software fault prediction. IEEE Access **6**, 25675–25685 (2018)

View publication stats