An Exemplar-Based Clustering Model with Loose Constraints in Social Network

Bi Anqi(✉) and Ying Wenhao

Changshu Institute of Technology, Changshu, Jiangsu, China
anqi_b@cslg.edu.cn

Abstract. Loose constraints have great effects on the study of message passing through social networks. This paper proposes a novel EEM-LC model who joints the pairwise loose constraints existing in social networks and the exemplar-based clustering model together, and also observes the application prospects of this model. Exemplar-based clustering model directly selects cluster centers from actual samples, so the structure and semantics of the comments on social networks would be preserved accordingly. Besides, EEM-LC unifies the two pairwise link constraints by one mathematical definition, and looses the restrictions of strong constraints. Moreover, on the basis of the Bayesian probability framework, EEM-LC implants loose pairwise constraints into its target function. That is to say, enhanced α-expansion move algorithm is capable of optimizing this new model. Experimental results based on several real-world data sets have shown very convincing performance of the proposed EEM-LC model.

Keywords: Loose constraints · Exemplar-based clustering model · Message passing · Social networks

1 Introduction

Sociologists have found that weak ties [8,11,13] have significant effects on the message passing on social network. Weak ties is a more extensive but superficial social cognition of social relationships. Although weak ties are not as straightforward as classical connections, it potentially has extremely fast, low-cost and high efficient propagation efficiency. Artificial intelligence techniques are widely used to study such weak ties. Generally speaking, procedure for artificial intelligence technology to process information on social networks contains 3 steps. Firstly, extract keywords from the comments published on the social networks. Then, analyze these keywords by machine learning models. Thirdly, series the

Supported by the Humanities and Social Sciences Foundation of the Ministry of Education under grant no.18YJCZH229 and the Natural Science Foundation of Jiangsu Province under grant no. BK20161268.

W. Fu et al. (Eds.): ICMTEL 2021, LNICST 387, pp. 248–254, 2021.
https://doi.org/10.1007/978-3-030-82562-1_22

evolution of public opinion. In this paper, we focus on the second step. As there have been many works on incorporating this weak constraints into typical mathematical models in social network processing [1,2,4,5,7,9], including Exemplar-Based Clustering model, Latent Dirichlet Allocation(LDA), Support Vector Machine(SVM) [6], etc. We focus on dealing with this weak constraints start from the exemplar-based clustering framework in this paper.

Unlike strong constraints, the loose constraints are pairwise. Two kinds of this pairwise constraints should be considered, that is must-link(ML) and cannot-link(CL). Clearly, must-link means the linked two samples should be assigned in one cluster, while cannot-link separates these two samples in different clusters. Furthermore by loose, we relax the restriction on sample's constraints. Namely, loose constraint only requires that a sample may have must-link and/or cannot-link, or just have no link.

In summary, in this paper, we derive an extended version of exemplar-based clustering model for loose constraints existing in social networks, called EEM-LC in short. As exemplar-based framework directly selects cluster centers from actual samples, the obtained exemplars would preserve the structure and semantics of keywords naturally. Moreover, EEM-LC also unifies the two pairwise link constraints together, thus we loose the restrictions of the constraints. Obviously, such pattern is more consistent with the broad but superficial nature of weak ties. Besides, Bayesian probability framework is introduced, which can naturally helps us to implant loose pairwise constraints into the exemplar-based clustering model and improve generalization performance of the algorithm. We would deeply discuss the scenario and EEM-LC mechanism in the next sections.

2 Exemplar-Based Clustering Model with Loose Constraints

In this section, we derive a novel exemplar-based clustering model dealing with this pairwise loose constraints exist in social networks. The procedure of this proposed EEM-LC is described in Fig. 1.

Assume $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N\} \in \mathbb{R}^{N*D}$, N is the total number of D-dimensional data points. E is the output exemplar set, whereas the element $E(i)$ refers to the exemplar for sample $\boldsymbol{x}_i$. According to the discussion above, we give the mathematical definition of the involved loose pairwise constraints in Eq. (1) below:

$$C(\boldsymbol{x}_i, \boldsymbol{x}_j) = \begin{cases} 1, & \boldsymbol{x}_i \; must-link \; with \; \boldsymbol{x}_j \\ 0, & \boldsymbol{x}_i \; no-link \; with \; \boldsymbol{x}_j \\ -1, & \boldsymbol{x}_i \; cannot-link \; with \; \boldsymbol{x}_j \end{cases} \tag{1}$$

Therefore, for current data $\boldsymbol{x}_i$, subset $\mathbf{ML}_{\boldsymbol{x}_i} = \{\boldsymbol{x}_j | C(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1\}$ defines all samples must-link with $\boldsymbol{x}_i$, and subset $\mathbf{CL}_{\boldsymbol{x}_i} = \{\boldsymbol{x}_j | C(\boldsymbol{x}_i, \boldsymbol{x}_j) = -1\}$ defines all samples cannot-link with $\boldsymbol{x}_i$. So far, start from the theory of machine learning, we have given several definitions related to the loose pairwise constraints. This

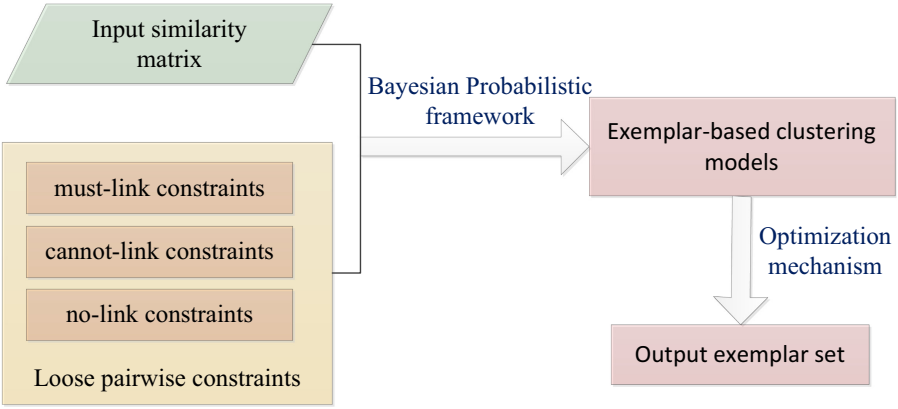

Fig. 1. The procedure of EEM-LC framework. We implant the loose pairwise constraints into the exemplar-based models.

work would naturally helps us to implant loose pairwise constraints into the basic machine learning models.

Theoretically based on the descriptions of exemplar-based clustering model and Bayesian probabilistic framework, the rough target function of EEM-LC equals to Eq. (2).

$$\max_{E} \ln \prod_{i=1}^{N} p(\boldsymbol{x}_i) p(E) \tag{2}$$

where $p(E)$ is the probabilistic information based on the exemplar set, shown in Eq. (3), and $p(\boldsymbol{x}_i)$ represents probability relationship for a single sample which is listed in Eq. (4).

$$p(E) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\sum_{i=1}^{N}\sum_{j=1}^{N} \theta_{i,j}(E(i), E(j))/2\sigma^2\right) \tag{3}$$

$$\begin{aligned} p(\boldsymbol{x}_i) = {} & p(\boldsymbol{x}_i, \boldsymbol{x}_{E(i)}) \cdot \\ & \prod_{\boldsymbol{x}_j \in \mathbf{ML}_{\boldsymbol{x}_i}} p(\boldsymbol{x}_j, \boldsymbol{x}_{E(j)}) p(\boldsymbol{x}_{E(i)}, \boldsymbol{x}_{E(j)}) \cdot \prod_{\boldsymbol{x}_j \in \mathbf{CL}_{\boldsymbol{x}_i}} p(\boldsymbol{x}_j, \boldsymbol{x}_{E(j)}) p(\boldsymbol{x}_{E(i)}, \boldsymbol{x}_{E(j)}) \end{aligned} \tag{4}$$

Carefully analyze Eq. (4), note that if $\mathbf{ML}_{\boldsymbol{x}_i}$ or $\mathbf{CL}_{\boldsymbol{x}_i}$ is empty, the corresponding value is set to be 1, which means the EEM-LC model is inspective of loose link constraints here. The framework will degrade into classical exemplar-based model EEM in [12]. On the other hand, notations $p(\boldsymbol{x}_i, \boldsymbol{x}_{E(i)})$ and $p(\boldsymbol{x}_{E(i)}, \boldsymbol{x}_{E(j)})$ are defined as Eqs. (5 and 6).

$$p(\boldsymbol{x}_i, \boldsymbol{x}_{E(i)}) = \frac{1}{\sigma\sqrt{2\pi}} \exp(s(\boldsymbol{x}_i, \boldsymbol{x}_{E(i)})/2\sigma^2) \tag{5}$$

$$p(\boldsymbol{x}_{E(i)}, \boldsymbol{x}_{E(j)}) = \frac{1}{\sigma\sqrt{2\pi}} \cdot \exp\left(-\sum_{i=1}^{N}\sum_{j=1}^{N} \eta_{i,j}(E(i), E(j))/2\sigma^2\right) \tag{6}$$

$s(\boldsymbol{x}_i, \boldsymbol{x}_{E(i)})$ is the similarity relationship between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{E(i)}$. Usually we set $s(\boldsymbol{x}_i, \boldsymbol{x}_{E(i)}) = -d(\boldsymbol{x}_i, \boldsymbol{x}_{E(i)})$ where $d(\boldsymbol{x}_i, \boldsymbol{x}_{E(i)})$ is the Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_{E(i)}$.

$\theta_{i,j}(E(i), E(j))$ and $\eta_{i,j}(E(i), E(j))$ in Eqs. (3 and 6)are both set to guarantee the validity of the exemplar set. The definitions are below in Eqs. (7 and 8), where M, L, L' are set to be big here. See [1,12] for detail discussion.

$$\theta_{i,j}(E(i), E(j)) = \begin{cases} M, & E(i) = j, E(j) \neq j, or E(j) = i, E(i) \neq j \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$\eta_{i,j}(E(i), E(j)) = \begin{cases} L, & E(i) \neq E(j), C(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 \\ L', & E(i) = E(j), C(\boldsymbol{x}_i, \boldsymbol{x}_j) = -1 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Put the definitions of $p(E)$ and $p(\boldsymbol{x}_i)$ in Eqs. (3 and 4) into the rough target function Eq. (2), and after some mathematical simplifications, the final target function becomes

$$\begin{aligned} \min_{E} & \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \boldsymbol{X}} d(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \sum_{\boldsymbol{x}_j \in \mathbf{ML}_{\boldsymbol{x}_i}} \eta_{i,j}(E(i), E(j)) \\ & + \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \sum_{\boldsymbol{x}_j \in \mathbf{CL}_{\boldsymbol{x}_i}} \eta_{i,j}(E(i), E(j)) + \sum_{\boldsymbol{x}_i \in \boldsymbol{X}} \sum_{\boldsymbol{x}_j \in \boldsymbol{X}} \theta_{i,j}(E(i), E(j)) \end{aligned} \tag{9}$$

The target function in Eq. (9) can be approximately optimized by α-expansion move algorithm with s-t graph cut [1,10,12]. Hence we utilize the enhanced α-expansion move algorithm to optimize the proposed EEM-LC model. Specifically speaking, the optimization mechanism regards the target function as the energy defined by the Markov random field, thus the energy reduction is introduced to detect the change in the value of the target function. We traverse the possible exemplars, and compare the corresponding values of the reduction in energy. On this basis, we gradually get the minimum energy value, when the current exemplar set is the optimal solution of the target function as well. Besides, the optimization mechanism here also expands the search space for possible exemplars in iteration, that is, sub-optimal exemplar is considered as well. Experiments of correlation algorithms have proved that the trick improves the optimization efficiency of this model.

3 Experimental Analysis

3.1 Setup

Just to be clear, the experiments in this section are implemented in 2010a Matlab on a PC with 64 bit Microsoft Window 10, an Intel(R) Core(TM) i7-4712MQ

and 8GB memory. We first utilize Diabetes dataset from UCI Machine Learning Repository[1]. Diabetes dataset has 768 8-dimensional samples, and contains 2 classes. Also to better observe the characteristics of social network, we took "movies" as key word, crawled 1000 comments on Sina Weibo[2] with Python 3.7.3. Then we use the Jieba toolbox for word segmentation, and obtain relevant high-dimensional vector by word2vex toolbox. Finally, PCA is also introduced to reduce the dimension to 20 dimensions, and accordingly establish the dataset called D1 here. Thus, ignore some nonsense words, dataset D1 has 3500 20-dimensional samples. Furthermore, considering that the loose pairwise constraints is extensive and low-cost, for both Diabetes and D1, we randomly assign either must-link or cannot-link constraints to 75% samples.

We compare our model EEM-LC with classical EEM [12] model and AP [3] model, to observe the important role of loose pairwise constraints in the cluster procedure. On the other hand, for Diabetes dataset, true labels for samples are available, so we evaluate the performances of these models by indices RI, which is defined below:

$$RI = \frac{f_{00} + f_{11}}{N(N-1)/2} \tag{10}$$

where f_{00} is the number of data whose cluster is in line with its class, while f_{11} is the number of those data whose cluster is inconsistent with its class. For D1, as true labels are not available, we take the number of clusters as performance indices.

3.2 Results Analysis

Each algorithm is repeatedly executed 10 times, we record the average performance. For the involved 3 models, multiply the median value of similarities by the optimal value in $\{0.01, 0.1, 1, 10, 50, 100\}$, then we get the value of self similarity $d(\boldsymbol{x}_i, \boldsymbol{x}_i)$. Grid search is used to choose parameter for Diabetes dataset. As to D1 dataset, considering that true labels are not available, we set $d(\boldsymbol{x}_i, \boldsymbol{x}_i)$ equals to 10 times of the median value of similarities.

Figure 2 shows the average comparison of EEM, AP and EEM-LC on Diabetes dataset by RI, meanwhile Fig. 3 describes the average number of clusters obtained by the 3 models on D1 dataset. Deeply observing Figs. 2 and 3, we can conclude that when real labels are available, the clustering results of the EEM-LC model are reliable. Though the real labels are absent, EEM-LC still can effectively deal with the information by incorporating with loose pairwise constraints. The performance of EEM-LC is very promising.

In the experiment procedure, we found that compared to both EEM and AP, the EEM-LC model has the longest running time, which is because the model deals with the loose pairwise constraints involved in the sample while processing the input similarity matrix.

[1] http://archive.ics.uci.edu/ml/datasets/Diabetes.
[2] http://m.weibo.cn/.

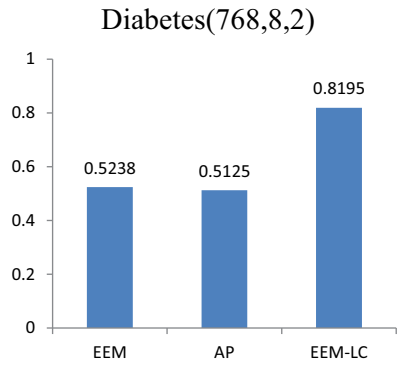

Fig. 2. Comparison of EEM, AP and EEM-LC on Diabetes by *RI*

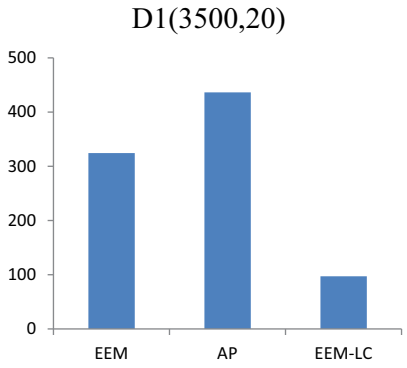

Fig. 3. Average number of clusters obtained by EEM, AP and EEM-LC on D1

4 Conclusion

This paper focuses on incorporating loose pairwise constraints with classical exemplar-based clustering model, and proposes EEM-LC algorithm. Our experimental results have shown promising performance of EEM-LC. In the future, such a model will help us further study the public opinion transmission, event evolution, public opinion monitoring, etc. However, the time complexity of the EEM-LC algorithm needs to be reduced.

References

1. Bi, A., Fulai Chung, S.W.: Bayesian enhanced α-expansion move clustering with loose link constraints. Neurocomputing **194**, 288–300 (2016)
2. Arzeno Natalia M.V.H.: Semi-supervised affinity propagation with soft instance-level constraints. IEEE Trans. Pattern Anal. Mach. Intell. **37**(5), 1041–1052 (2015)
3. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**, 972–976 (2017)

4. Blei, D.M., Ng, A., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 903–1022 (2003)
5. Bugaychenko, D., Dzuba, A.: Musical recommendations and personalization in a social network (2013). https://doi.org/10.1145/2507157.2507192
6. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
7. Givoni, I., Frey, B.: Semi-supervised affinity propagation with instance-level constraints. J. Mach. Learn. Res. Proc. Track **5**, 161–168 (2009)
8. Granovetter, M.: The strength of weak ties. Amer. J. Sociol. **78**, 1360–1380 (1973)
9. Luo, D., Huang, H.: Link prediction of multimedia social network via unsupervised face recognition. In: Proceedings of the 17th International Conference on Multimedia 2009, Vancouver, British Columbia, Canada, 19–24 October 2009 (2009)
10. Tappen, M.F., Freeman, W.T.: Comparison of graph cuts with belief propagation for stereo, using identical MRF parameters. In: 9th IEEE International Conference Computer Vision, pp. 900–906 (2003)
11. Weenig, M.W.H.: The strength of weak and strong communication ties in a community information program1. J. Appl. Social Psychol. **23**(20), 1712–1731 (2006)
12. Zheng, Y.,Chang, P.: Clustering based on enhanced -expansion move. IEEE Trans. Knowl. Data Eng. (TKDE) **25** (2013)
13. Zhao, J., Wu, J., Xu, K.: Weak ties: a subtle role in the information diffusion of online social networks. Phys. Rev. E **82**(1), 1–1348 (2010)