IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Special Issue, 6, March 2019 E-ISSN: 2347-2693

A Review on Multi-Task clustering with self-adaptive and Model Relation Learning

C.Rupakumar[1*], S. Girinath[2]

[1,2]Dept. of MCA, Sri Padmavathi College of Computer Sciences And Technology,Tiruchanoor-Tirupati, India

[*]*Corresponding Author: Roopkumar.sirigili@gmail.com, Tel.: +91-9133852098*

DOI: https://doi.org/10.26438/ijcse/v7si6.106108 | Available online at: www.ijcseonline.org

Abstract - Multi-task clustering improves the clustering performance of each task by transferring knowledge among the related tasks. An important aspect of multi-task clustering is to assess the task relatedness. However, to our knowledge, only two previous works have assessed the task relatedness, but they both have limitations. In this paper, we propose two multi-task clustering methods for partially related tasks: the self-adapted multi-task clustering (SAMTC) method and the manifold regularized coding multi-task clustering (MRCMTC) method, which can automatically identify and transfer related instances among the tasks, thus avoiding negative transfer. Both SAMTC and MRCMTC construct the similarity matrix for each target task by exploiting useful information from the source tasks through related instances transfer, and adopt spectral clustering to get the final clustering results. But they learn the related instances from the source tasks in different ways.

Key words**: -** Multi-task Clustering, Partially Related Tasks, Negative Transfer, Instance Transfer.

I. INTRODUCTION

Along with the progression of information technology, tremendous number of unlabeled data are been generated each day. It is time-consuming and expensive if the data are labeled manually, hence we step on to clustering algorithms for dredging the unexplored knowledge in the data. Clustering is a task of allocating a group of samples to a set of clusters such that tasks belonging to one cluster are similar than tasks in other clusters. This technique is adopted for statistical data analysis in enormous fields such as information retrieval, pattern recognition, object recognition, image analysis and so on. One among the traditional simple task clustering technique is k-means clustering algorithm. It aims to find a mean vector for each one of the C clusters, and partitions a given sample into a specific cluster with the least nearest mean. However, the samples tend to scatter within the cluster, and in such cases k-means may achieve poor performance, because the vector cannot fully represent the scattering observations within the cluster.

There are enormous amount of related data set, for instance, given the web pages from two schools, i.e., P and Q, the ultimate aim is to cluster web pages of each school into three categories, e.g., teacher, project and subject. MTC have been receiving increasing interest over the past decade. Clustering the web pages of each school is referred as a task. Instinctively, these two tasks are related, since the web page contents and label spaces are similar. It is possible to use such relationship between tasks and clustering simultaneously and that leads to better performance and robust solution. One direct approach is to combine the web pages of two schools together, followed by classical simple task clustering such as k-means algorithm. Although, it results in poor performance, because school P and school Q outputs distinct characteristics and hence the distributions of the web pages are different.

In this paper, we propose two multi-task clustering methods for partially related tasks: the self-adapted multitask clustering (SAMTC) method and the manifold regularized coding multi-task clustering (MRCMTC) method, which can automatically identify and transfer related instances among the tasks, thus avoiding negative transfer when the tasks are partially related. In the multi-task setting, each task can be seen as a target task, and the other tasks are source tasks. If the given tasks are related, there are certain parts of instances from the source tasks that can be reused for clustering each target task. The intention of SAMTC and MRCMTC is to identify such parts and transfer knowledge among them.

SAMTC begins with an initialization by performing single-task clustering on each task, and then executes the following three steps. 1) Reusable instances finding: it computes the distance of any two clusters between each pair of source and target tasks with a common distribution and considers only the instances in any two clusters whose distance is smaller

than a boundary value and which are closest to each other reusable. Then it constructs a pair of source and target subtasks with the obtained reusable instances. 2) Subtask relatedness learning: it calculates the weights of the instances in the source subtask by performing kernel mean matching between each pair of source subtask and target subtask. 3) Clustering through instance transfer: it constructs a similarity matrix for each target task by exploiting the reusable instances in the source tasks, then it adopts spectral clustering to get the final clustering results.

MRCMTC learns the related instances from the source tasks in a different way. It consists of two steps. 1) Related instances learning: it alternatively learns a lower dimensional feature space which can reduce the domain divergence for each pair of source and target tasks, and calculates which part of instances in the source tasks can represent each data point in the target task under the lower dimensional feature space, and the two procedures boost each other. 2) Clustering through instance transfer: it constructs a similarity matrix for each target task by utilizing the instances with higher representative coefficients in the source tasks, and then it adopts spectral clustering to get the final clustering results.

II. RELATED WORK

In this paper, Xiaotong Zhang, Xianchao Zhang, Han Liu, and Xinyue Liu proposed multi-task clustering methods for partially related tasks: the self-adapted multitask clustering (SAMTC) method and the manifold regularized coding multi-task clustering (MRCMTC) method, which can automatically identify and transfer related instances among the tasks, thus avoiding negative transfer when the tasks are partially related. In the multi-task setting, each task can be seen as a target task, and the other tasks are source tasks. If the given tasks are related, there are certain parts of instances from the source tasks that can be reused for clustering each target task [1]. The intention of SAMTC and MRCMTC is to identify such parts and transfer knowledge among them

III. ALGORITHM SELF-ADAPTEMULTI-TASK CLUSTERING

SAMTC begins with an initialization by performing single task clustering on each task, e.g., the Normalized Cut spectral clustering method with the Shared Nearest Neighbor similarity then it executes the following three steps.

1) Reusable instances finding: this step is to find reusable instances in the source tasks for each target task. Generally if the tasks are related, there must exist some instances belonging to the same topic among these tasks. Therefore, for each target task, if we find the related clusters in its source tasks, the instances in the related clusters are considered reusable to the target task. To find the related clusters, we compute the distance of any two clusters between the source and target tasks through a commonly used symmetric and bounded distribution measure, i.e., the Jensen Shannon divergence. Then based on the assumption that only the clusters whose distance is smaller than a boundary value and which are closest to each other are possibly related, a pair of source and target subtasks which consist of the possibly related clusters are obtained, and instance knowledge is only transferred between the source and target subtasks.

2) Subtask relatedness learning: this step is to further explore the relatedness of the constructed subtasks based on the first step. We use kernel mean matching to estimate the relatedness between them. Kernel mean matching is a method for estimating the mathematical expectation of a particular distribution by a weighted average of instances from another distribution. In other words, given a pair of source subtask and target subtask, kernel mean matching actively selects positive weighted instances from the source subtask to estimate the expectation of the target subtask distribution, which means that these selected positive weighted instances in the source subtask are potentially relevant to the target subtask. Thus the relatedness of the source subtask to the target subtask can be estimated by the ratio of positive weighted instances in the source subtask.

3) Clustering through instance transfer: this step is to cluster each target task by exploiting useful instance knowledge from its source tasks. We construct a similarity matrix for each target task; with the similarity between any two data points to be the normalized weighted number of shared nearest neighbors from the target task itself and its source tasks. For each data point in the target task, if it is in the target subtask, its nearest neighbors are computed in both the target task and the corresponding source subtasks; otherwise, its nearest neighbors are computed only in the target task. Then we adopt the Normalized Cut spectral clustering to cluster each target task based on the learned similarity matrix. Through such a similarity construction, not only the useful instance knowledge from all tasks is used, but the independence of the clustering for each task is also ensured.

MANIFOLD REGULARIZED CODING MULTI-TASK CLUSTERING:-

There are two steps in MRCMTC. 1) Related instances learning: this step is to learn the related instances from the source tasks for each data point in the target task. It is motivated by sparse coding which discovers sets of basis vectors to represent data. More specifically, each data point in the target task can be represented by a linear combination of the data points in the source task, and the more related the data points are, the higher representative coefficients they have. But considering that there often exists domain divergence among different tasks (i.e., the source and target tasks usually have task specific features, the related data

points in the source and target tasks are not close enough), which is unfavorable to use the data points in the source tasks to represent the data points in the target task. Therefore, for each pair of source and target tasks, we first map the data points to a lower dimensional feature space such that the domain divergence between the source and target tasks can be reduced, i.e., any two data points with a higher representative coefficient between the source and target tasks should have a smaller distance in the lower dimensional feature space. The process of learning the related instances from the source tasks is encoded into an objective function, it alternatively learns a lower dimensional feature space which can reduce the domain divergence for each pair of source and target tasks, and calculates which part of instances in the source tasks can represent each data point in the target task under the lower dimensional feature space, and the two procedures boost each other. Clustering through instance transfer: this step is to cluster each target task by utilizing the instances with higher representative coefficients from its source tasks. We construct a similarity matrix for each target task, with the similarity between any two data points to be the number of shared related instances from the target task itself and its source tasks. For each data point in the target task, its related instances from the target task itself are the nearest neighbors; its related instances from the source tasks are the data points with higher representative coefficients, which are also its nearest neighbors from the source tasks. Then we adopt the Normalized Cut spectral clustering to cluster each target task according to the learned similarity matrix.

IV. CONCLUSION AND FUTURE SCOPE

In this paper, we have proposed two multi-task clustering methods for partially related tasks: the self-adapted multitask clustering (SAMTC) method and the manifold regularized coding multi-task clustering (MRCMTC) method. They first identify the related instances from the source tasks for each target task, then construct the similarity matrix for each target task by exploiting the related instances from the source tasks based on the Shared Nearest Neighbor similarity, finally perform the spectral clustering on the constructed similarity matrix. Both of them can exploit the positive relationship among the tasks and avoid negative transfer by identifying the related instances between each pair of tasks. But they learn the related instances from the source tasks in different ways. SAMTC reconstructs a pair of source and target subtasks that contain the possibly related clusters, and only the instances in the source subtask are considered reusable to the data points in the target subtask. MRCMTC calculates the representative instances from the source task for each data point in the target task under a lower dimensional feature space which can reduce the domain divergence for each pair of source and target tasks, and it is more stable than SAMTC by avoiding performing single-task clustering in the initialization.

REFERENCES

[1] S. J. Pan and Q. Yang, "A survey on transfer learning," IEEE Trans. Knowl. Data Eng., vol. 22, no. 10, pp. 1345–1359, 2010.

[2] J. Zhang and C. Zhang, "Multitask Bregman clustering," in Proc. 24th AAAI Conf. Artif. Intell., 2010, pp. 655–660.

[3] X. Zhang and X. Zhang, "Smart multi-task Bregman clustering and multi-task Kernel clustering," in Proc. 27th AAAI Conf. Artif. Intell., 2013, pp. 1034–1040.

[4] X. Zhang, "Convex discriminative multitask clustering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 1, pp. 28–40, 2015.

[5] J. Lin, "Divergence measures based on the shannon entropy," IEEE Trans. Inform. Theory, vol. 37, no. 1, pp. 145–151, 1991.

[6] J. Huang, A. J. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," ¨ in Proc. 20th Adv. Neural Inform. Process. Syst., 2006, pp. 601–608.

[7] X. Zhang, X. Zhang, and H. Liu, "Self-adapted multi-task clustering," in Proc. 25th Int. Joint Conf. Artif. Intell., 2016, pp. 2357–2363.

[8] R. Caruana, "Multitask learning," Mach. Learn., vol. 28, no. 1, pp. 41–75, 1997.

[9] R. K. Ando and T. Zhang, "A framework for learning predictive structures from multiple tasks and unlabeled data," J. Mach. Learn. Res., vol. 6, pp. 1817–1853, 2005.

[10] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in Proc. 20th Adv. Neural Inform. Process. Syst., 2006, pp. 41–48.

[11] J. Chen, L. Tang, J. Liu, and J. Ye, "A convex formulation for learning shared structures from multiple tasks," in Proc. 26th Int. Conf. Mach. Learn., 2009, pp. 137–144.

[12] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in Proc. 10th ACM SIGKDD Int. Conf. Knowl. Disc. Data Min., 2004, pp. 109–117.

[13] C. A. Micchelli and M. Pontil, "Kernels for multi–task learning," in Proc. 18th Adv. Neural Inform. Process. Syst., 2004.

[14] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," J. Mach. Learn. Res., vol. 6, pp. 615–637, 2005.

[15] A. Barzilai and K. Crammer, "Convex multi-task learning by clustering," in Proc. 18th Int. Conf. Artif. Intell. and Stat., 2015, pp. 65–73.

[16] N. D. Lawrence and J. C. Platt, "Learning to learn with the informative vector machine," in Proc. 21st Int. Conf. Mach. Learn., 2004.

[17] E. V. Bonilla, K. M. A. Chai, and C. K. I. Williams, "Multitask gaussian process prediction," in Proc. 21st Adv. Neural Inform. Process. Syst., 2007, pp. 153–160.

[18] B. Zadrozny, "Learning and evaluating classifiers under sample selection bias," in Proc. 21st Int. Conf. Mach. Learn., 2004, pp. 114–121.

[19] W. Dai, Q. Yang, G. Xue, and Y. Yu, "Boosting for transfer learning," in Proc. 24th Int. Conf. Mach. Learn., 2007, pp. 193–200.

[20] W. Dai, G. Xue, Q. Yang, and Y. Yu, "Transferring naive bayes classifiers for text classification," in Proc. AAAI Conf. on Artif. Intell., 2007, pp. 540–545.

Author's Profile

Mr. Cirigili Rupakumar has received his graduation degree in BSc. Bachelor of Science from gayatri Degree & PG College, Affiliated to S.V University , Chittoor , AP in the year of 2013 – 2016. At Present he is Pursuing Post graduate degree MCA, Master of Computer Applications from Sri Padmavathi College of Computer Sciences and Technology Affiliated to SriVenkateswara University , Tirupati, AP, India.

