ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

An overview of Multi Task Clustering Techniques: a Survey

M. Arul Jothi, Amritha Chandran

Assistant Professor, Dept. of CSE., Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu

Student, Dept. of CSE., Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu

ABSTRACT: Clustering is the process of grouping data points into a bunch such that data points that belong to one cluster are more similar to each other than data points that belong to other cluster. Lots of data is being collected and warehoused and due to emanation of data on web, many issue s had been hurled on classical clustering algorithms such as the correlations that exist among many clustering tasks. However traditional clustering algorithms fail to deal with related clustering tasks. Combining such multiple clustering tasks that are related to each other are simply termed as Multi Task Clustering (MTC). It is one such promising way to reinforce the clustering performance. In this paper an effort is made to address various multi task clustering techniques in the data mining environment.

KEYWORDS: Multi Task Clustering, Correlations, co-Cluster, Multi Task Learning

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.

The clustering algorithms can further be subdivided into two classes – generative clustering and discriminative clustering. Generative clustering will build models to define clusters and use these models to classify data points. Discriminative clustering will involve in grouping data points based on the similarity measures. Examples of Discriminative clustering are *k*-means and hierarchical clustering and of Generative clustering are finite guassian mixture model. The essential problem in multi-task learning is to find a way to categorize the task relation. Some of the Representative methods in data mining that involves in characterizing task relation include learning a shared subspace, by using the common prior of model parameters and the kernel methods.

II. LITERATURE SURVEY

Q. Gu and J. Zhou proposed a novel approach which performs clustering of multiple related tasks simultaneously. The underlying relation between tasks can be used to complement the clustering performance and this is considered to be the pristine work addressing multi task clustering. The ultimate aim is to find a subspace shared by all clustering

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

tasks and the subspace is treated as a new feature representation. When the labels of the data points are predicted in advance, then the knowledge can further be used to predict the data in target task. This is simply termed as domain adaption or transductive transfer classification. The proposed approach not only makes use of the knowledge about a individual task but also utilizes the knowledge shared among multiple tasks. The transductive transfer classification surpass many classical clustering algorithms and this progress is because the knowledge obtained from one source task is transferred to the target task.

Q. Gu, Z. Li, and J. Han proposed a method called Kernel Learning and the ultimate goal of this approach is to describe a Reproducing Kernel Hillbert Space(RKHS) where the physical structure of the data point is concealed. Generally there are two types of kernel learning such as Non Parametric Kernel Learning and Spectral Kernel Learning. The above kernel learning can be implemented by linear programming. Learning a kernel provides better generalization ability provided if the relations between the tasks are well utilized. The objective is to find a RKHS in which the data distribution of various tasks are of close proximity. A kernel k must be sought such that the mapped data in the RKHS space is not only smooth with respect to geometric structure but also to the data distribution. According to this approach, it is importunate to guard the geometric structure of the data and need not require data distribution of one task to be close to another task. Once the kernel mean matching is performed, the distribution of all tasks are similar and this is followed by single task clustering algorithm such as *k*-means. Experimental results demonstrate that the proposed approach out performs the existing clustering algorithms.

S. Kong and D. Wang proposed an unsupervised cluster method via a multi-task learning strategy which is known as Mt Cluster. To represent sample signals and a shared common pattern pool (the commonality) Mt cluster learns the cluster specific dictionary for the essentially complemental representation. By treating learning the cluster-specific dictionary as a single task, Mt Cluster works in a multi-task learning manner, in which all the tasks are connected by simultaneously learning the commonality. However, some atoms of some learned dictionaries can be very similar or coherent, thus they can be used for representing signals from different clusters. Empirically, it is observed that different classes of signals (*e.g.* images) always share some common patterns, and these shared patterns are not promotive to discrimination of them but are essential for representing them. Inspired by this cognition, the common patterns are separated and the most discriminative cluster specific features are described to achieve better performance. The cluster-specific dictionary pretends to be the representative for the samples belonging to the corresponding cluster, *i.e.* a few bases of the dictionary can be linearly combined to well represent the samples in a exiguous manner.

Kurt Hornik demonstrated Cluster Ensemble(CE) which is a technique similar to MTC that is widely employed in machine learning and data mining. CE contains various clustering of a list of data points onto a single compact clustering, which is often interpolated as Consensus Solution. Consensus clustering provides robust and stable clustering when compared with traditional task clustering algorithms. The CE integrates multiple clustering using consensus function to improve the accuracy and stability. The cl_ensemble() function with reference to [6] tends to create CE from multiple clustering which is given as input, either as one by one or as a list. The objective of cl_ensemble() function is to check whether that all clustering are of similar type and all have similar number of objects.

S. Xie, H. Lu, and Y. He developed a multitask learning approach by modeling the task-feature relationships. Specifically, instead of assuming that similar tasks have similar weights on all the features, an assumption is made that the tasks must be related in terms of subsets of features, which implies a co-cluster structure. That is, the clustering process should be conducted simultaneously either on tasks and features, or on the rows and columns of the matrix, and the aim is to divide the tasks and features into k groups A novel regularization term to capture this task-feature co-cluster structure is designed. The problem of learning multitasks with the assumption that tasks can be classified into various groups which are not known in prior is considered. With this task clustering structure, the problems of negative information transfer among dissimilar and outlier tasks can be avoided. Besides global similarities, we would like to include a second regularization term that encodes clustering of tasks. Therefore a natural advantage of the task clustering approach is improved robustness when learning from multiple tasks and goal is to present a novel Co-Clustered Multi-Task Learning method (CoCMTL). The proposed approach is formulated as a decomposition model which separates the global similarities and group-specific similarities. To capture the group specific similarities, unlike the traditional task clustering approaches with only one-way clustering, we impose a novel regularization term which leads to a block structure.

Tao Li and Chris Ding studied an unique framework with the ability to cluster multiple related tasks based on the inter task and intra task relationships through geometric affine transformation of distance between an intra task and inter task instance. The geometric transformation improves inter task clustering without devastating the individual tasks with the bias acquired from other tasks. Recent results had empirically proved that, given several related tasks with

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

different data distributions and an algorithm that can utilize both task-specific and cross-task knowledge can significantly enhance the clustering performance. The main focus is on clustering multi tasks via a 3-factor non negative matrix factorization. The objective comprises mainly of two parts (1) Intra task co-clustering: to cluster the data points in the given space separately (2) Inter task regularization: to learn and define the relations of feature spaces among different tasks. An intra task orthogonal constraint is induced to a symmetric non negative matrix factorization form to acquire vectors that are close orthogonal within each task, which induces prior learning that a task must have independent clusters and the results proved that it outperforms other traditional single task clustering methods.

J. Zhou, J. Chen, and J. Ye proposed a technique called Alternating Structure Optimization(ASO) which is a well-known MTL approach that makes an attempt to learn shared low dimensional structure of various related tasks on the hypothesis space. As an alternate MTL technique, Clustered Multi Task Learning(CMTL) ascertain that many tasks tends to follow a clustered structure such that input tasks are classified into a set of clusters. The overall objective of CMTL and ASO differs in one aspect i.e., how the tasks are related. The equivalence relationship between CMTL and ASO provides new vision to ASO and CMTL and to their basic relationship. One major limitation of ASO/CMTL formulation is that it involves a non-convex optimization as the negative Ky Fan k-norm is concave. A convex relaxation of CMTL is proposed and the equivalence relationship between the proposed convex relaxation of CMTL and ASO is also established. The proposed approach is widely employed in various real time applications.

Jianwen Zhang and Changshui Zhang demonstrated that Multi task clustering can be specified as cutting down the loss that comprises of both task loss and task regularization. Each loss related to the task can be defined as the mean Bregman-divergence from the given input sample and to the centroid of the cluster. And the task regularization is a type of average divergence between partitions of any two tasks. An attempt is made to gain knowledge about the relationships that exists among different clusters of the task. Clustering is regularized as a problem of reducing the mean divergence between sample data and the cluster centroid. Therefore divergence assumption is important for a clustering algorithm and a large number of divergences can simply be termed as Bregman-divergence. The proposed approach is based on the general Bregman-divergence and hence its suitable for assumptions made on the divergence and distribution of the data and also it has proven to provide uniform solution to all divergence assumption and to the data distribution.

Comparison:

MTC Techniques	Methodology	NMI Measure
Shared Sub space Learning	Transfers knowledge to the target task	48%
Kernel Learning	Preserves Geometric structure of data	50%
Dictionary Learning	Separates the most common patterns	97%
Cluster Ensemble	Consensus function to combine multiple clustering	
Co-Cluster MTL	Regularization term used to capture co-cluster structure	75%
Non Negative Matrix Factorization	Performs hard and soft clustering simultaneously	93.2%
Alternating Structure Optimization	Learns shared low dimensional structure on tasks on hypothesis space	
Bregman Clustering	Reduces mean divergence between the sample and cluster centroid	61%

III. CONCLUSION

Clustering is one among the most popular techniques explored in many real applications such as statistics, image processing and object recognition. And Multi Task clustering involves grouping related tasks together that will help to acquire comparatively better performance than grouping the tasks separately. In this paper, we have introduced the theoretical background and procedures of Multi Task clustering. The main focus was to review on the variety of clustering approaches.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 2, February 2016

REFERENCES

1. K. Wagstaff, C. Cardie, S. Rogers, and S. Schrodl, "Constrained k-means clustering with background knowledge," in ICML, 2001, pp. 577–584.
2. I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in KDD, 2001, pp. 269–274.
3. C. A. Micchelli and M. Pontil, "Kernels for multi–task learning," in NIPS, 2004.
4. T. Li, C. Ding, and M. I. Jordan, "Solving consensus and semi-supervised clustering problems using nonnegative matrix factorization," in ICDM, 2007, pp. 577–582.
5. Wu and B. Scholkopf, "A local learning approach for clustering," in Advances in Neural Information Processing Systems 19, B. Scholkopf, J. C. Platt, and T. Hoffman, Eds., vol. 19. Cambridge, MA, USA: MIT Press, Dec. 2007 pp. 1529–1536.
6. Kurt Hornik "A CLUE for CLUster Ensembles" in 2007
7. W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, "Self-taught clustering," in Proc. 25th Int. Conf. Mach. Learn. (ICML), Helsinki, Finland, 2008, pp. 200–207.
8. F. Wang, C. Zhang, and T. Li, "Clustering with local and global regularization," IEEE Trans. Knowl. Data Eng., vol. 21, no. 12, pp. 1665–1678, Dec. 2009.
9. Q. Gu and J. Zhou, "Learning the shared subspace for multi-task clustering and transductive transfer classification," in Proc. 9th IEEE Int. Conf. Data Min. (ICDM), Miami, FL, USA, 2009, pp. 159–168.
10. J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in Proc. 24th Conf. Adv. Neural Inf. Process.Syst. (NIPS), 2011, pp. 702–710.
11. Q. Gu, Z. Li, and J. Han, "Learning a kernel for multi-task clustering," in Proc. 25th AAAI Conf. Artif. Intell., San Francisco, CA, USA, Aug. 2011.
12. S. Kong and D. Wang, "A multi-task learning strategy for unsupervised clustering via explicitly separating the commonality," in Proc. 21St Int. Conf. Pattern Recognit. (ICPR), Tsukuba, Japan, Nov. 2012, pp. 771–774.
13. Z. Zhang and J. Zhou, "Multi-task clustering via domain adaptation," Pattern Recognit., vol. 45, no. 1, pp. 465–473, 2012.
14. S. Xie, H. Lu, and Y. He, "Multi-task co-clustering via nonnegative matrix factorization," in Proc. 21st Int. Conf. Pattern Recognit. (ICPR), Tsukuba, Japan, Nov. 2012, pp. 2954–2958.
15. Xiao-Lei Zhang, Member, IEEE "Convex Discriminative Multitask Clustering" in arXiv:1303.2130v2 [cs.LG] 21 Oct 2013.