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. 3, Issue 10, October 2015

Sequential Multi Task Spectral Clustering Scheme with Active Learning paradigm

Dhanabagyam D[1], Manikandan P [2]

Assistant Professor, Dept. of Computer Applications, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore, Tamil Nadu, India [1]

M.Phil Scholar, Dept. of Computer Science, Sri Jayendra Saraswathy Maha Vidyalaya College of Arts and Science, Coimbatore, Tamil Nadu, India[2]

ABSTRACT: Clustering is one of the most classical research problems in pattern recognition and data mining and it has been widely explored and applied to various applications. In MTSC (Multitask Spectral Clustering) there were the inter task correlations are identified in the unsupervised way in the random matched correlations, so that the clusters labels are most accurate. In this proposed study Sequential clustering is going to perform with the priority based correlation clustering. This proposed idea provides to a most accurate spectral clustering results. This sequential prediction finds the important labels between the multi tasks. There are many similar inter tasks are there but all the inter tasks are not important to make spectral clusters so that the prediction is performed by finding the most connected inter task labels. Since this proposed system find the most important labels this study gives more accuracy than the existing system and also it takes less time than the MTSC scheme. The experimental results and study shows that the proposed system outperforms the previous multi task clustering schemes.

KEYWORDS: Clustering, Multi task, spectral, Sequence, Priority

I. INTRODUCTION

Clustering is a process of partitioning a set of data or objects in a set of meaningful sub-classes, called Clusters. It is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Data clustering is a method in which we make cluster of objects that are somehow similar in characteristics.

The definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A Cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. The users can show this with a simple graphical example:

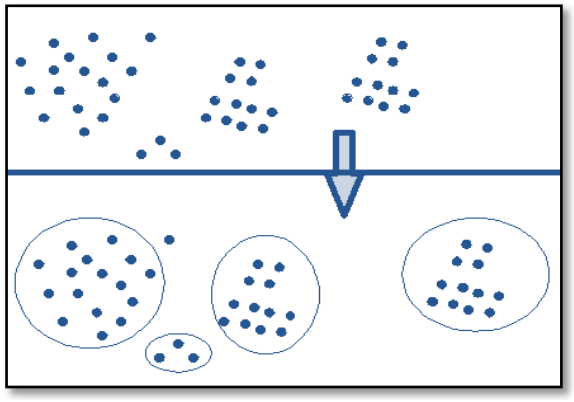

Fig: 1.1 Examples for Clustering

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. 3, Issue 10, October 2015

Spectral clustering:
Spectral clustering has many applications in machine learning, exploratory data analysis, computer vision and speech processing. Most techniques explicitly or implicitly assume a metric or a similarity structure over the space of configurations, which is then used by clustering algorithms. The success of such algorithms depends heavily on the choice of the metric, but this choice is generally not treated as part of the learning problem. Thus, time-consuming manual feature selection and weighting is often a necessary precursor to the use of spectral methods. Several recent papers have considered ways to alleviate this burden by incorporating prior knowledge into the metric, either in the setting of K-means clustering [1, 2] or spectral clustering [3, 4]. In this paper, we consider a complementary approach, providing a general framework for learning the similarity matrix for spectral clustering from examples. We assume that we are given sample data with known partitions and are asked to build similarity matrices that will lead to these partitions when spectral clustering is performed. This problem is motivated by the availability of such datasets for at least two domains of application: in vision and image segmentation, a hand-segmented dataset is now available, while for the blind separation of speech signals via partitioning of the time-frequency plane, training examples can be created by mixing previously captured signals.

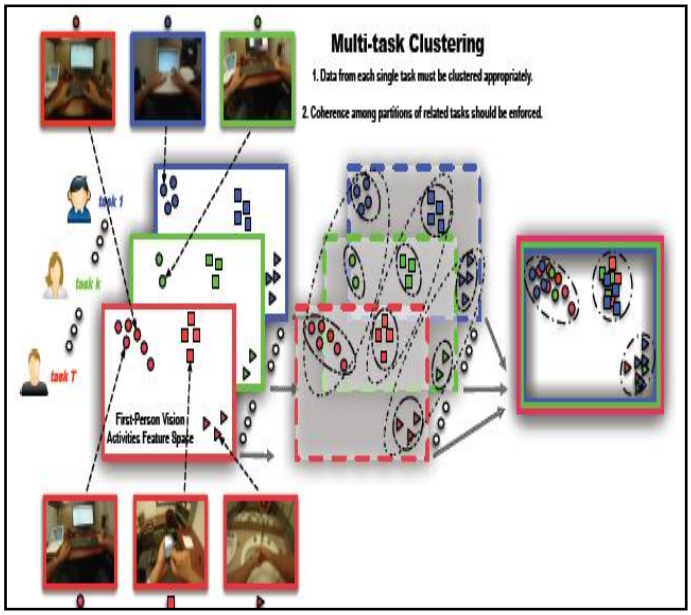

Fig: 1.2 Multi Task Clustering

1.2.2. Use of Clustering in Data Mining
Clustering is often one of the first steps in data mining analysis. It identifies groups of related records that can be used as a starting point for exploring further relationships. This technique supports the development of population segmentation models, such as demographic-based customer segmentation. Additional analyses using standard analytical and other data mining techniques can determine the characteristics of these segments with respect to some desired outcome. For example, the buying habits of multiple population segments might be compared to determine which segments to target for a new sales campaign. For example, a company that sale a variety of products may need to know about the sale of all of their products in order to check that what product is giving a wide scope and which is lacking. This is done by data mining techniques. But if the system clusters the products that are giving less sales then only the cluster of such products would have to be checked rather than comparing the sales value of all the products. This is actually to facilitate the mining process.

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. 3, Issue 10, October 2015

Components of a Clustering Task

Typical pattern clustering activity involves the following steps:

(1) Pattern representation (optionally including feature extraction and/or selection)
(2) Definition of a pattern proximity measure appropriate to the data domain
(3) Clustering or grouping
(4) Data abstraction (if needed) and
(5) Assessment of output (if needed).

II. LITERATURE SURVEY

The normalized cut criterion measures [5] both the total dissimilarity between the different groups as well as the total similarity within the groups. We show that an efficient computational technique based on a generalized eigen value problem can be used to optimize this criterion. We have applied this approach to segmenting static images, as well as motion sequences, and found the results to be very encouraging. [6] The method is based on a recently introduced information-theoretic principle, the information bottleneck (IB) principle. Images are clustered such that the mutual information between the clusters and the image content is maximally preserved. [7] The discretization is efficiently computed in an iterative fashion using singular value decomposition and non maximum suppression. The resulting discrete solutions are nearly global-optimal. [8] Spectral clustering arise from concepts in spectral graph theory and the clustering problem is configured as a graph cut problem where an appropriate objective function has to be optimized. An explicit proof of the fact that these two paradigms have the same objective is reported since it has been proven that these two seemingly different approaches have the same mathematical foundation. Besides, fuzzy kernel clustering methods are presented as extensions of kernel K-means clustering algorithm. [9] Focus on a semi-supervised framework that incorporates labeled and unlabeled data in a general-purpose learner. Some transductive graph learning algorithms and standard methods including support vector machines and regularized least squares can be obtained as special cases.

III. PROPOSED SYSTEM

Sequential inter task prediction algorithm is proposed in this study. The inter tasks between several tasks are represented in the following architecture diagram. In the Fig 3.1 there are three tasks are correlated and four inter tasks are represented. All the four inter tasks can be used to make clusters but some inter task only connected with other tasks or more important which means that task present in many task.

3.1 Basic sequential algorithmic scheme

The basic sequential algorithmic scheme (BSAS) is a very basic clustering algorithm that is easy to understand. In the basic form vectors are presented only once and the number of clusters is not known at priori. What is needed is the dissimilarity measured as the distance $d(x, C)$ between a vector point x and a cluster C, threshold of dissimilarity Θ and the number of maximum clusters allowed q. The idea is to assign every newly presented vector to an existing cluster or create a new cluster for this sample, depending on the distance to the already defined clusters. As can be seen the algorithm is simple but still quite efficient. Different choices for the distance function lead to different results and unfortunately the order in which the samples are presented can also have a great effect to the final result. What's also very important is a correct value for Θ. This value has a direct effect on the number of formed clusters.

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. 3, Issue 10, October 2015

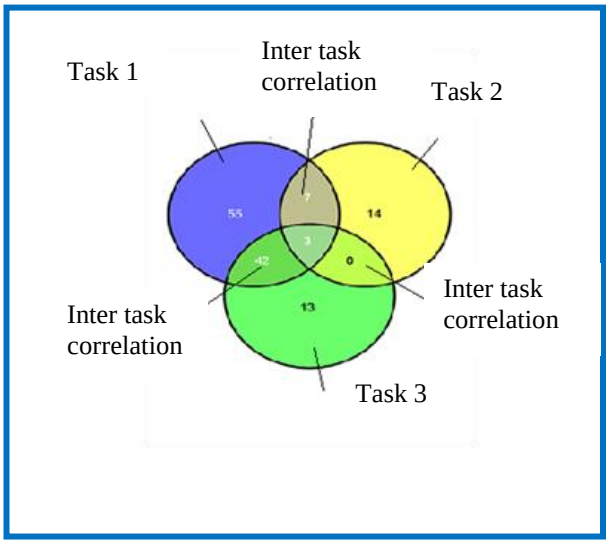

Fig 3.1 Architecture Diagram

If Θ is too small unnecessary clusters are created and if too large a value is chosen less than required number of clusters are formed. One detail is that if q is not defined the algorithm 'decides' the number of clusters on its own. This might be wanted under some circumstances but when dealing with limited resources a limited q is usually chosen. Also, BSAS can be used with a similarity function simply by replacing the min function with max.
There exists a modification to BSAS called modified BSAS (MBSAS), which runs twice through the samples. It overcomes the drawback that a final cluster for a single sample is decided before all the clusters have been created. The first phase of the algorithm creates the clusters (just like 2b in BSAS) and assigns only a single sample to each cluster.

IV. EXPERIMENT AND ANALYSIS

Time Comparison:
In this analysis we compare the running time of proposed method with the existing MTSC scheme and comparison result which achieve relatively better clustering performance than the existing algorithm. The following table 4.1 represents the time taken to make clusters of proposed and existing system and the chart shows the performance.

Clustering Time	**Multitask Spectral Clustering (MTSC)**	**Sequential Multi Task Spectral Clustering (BSAS)**
Dataset 1	30	25
Dataset 2	28	24
Dataset 3	36	29
Dataset 4	29	23

Table 4.1

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. 3, Issue 10, October 2015

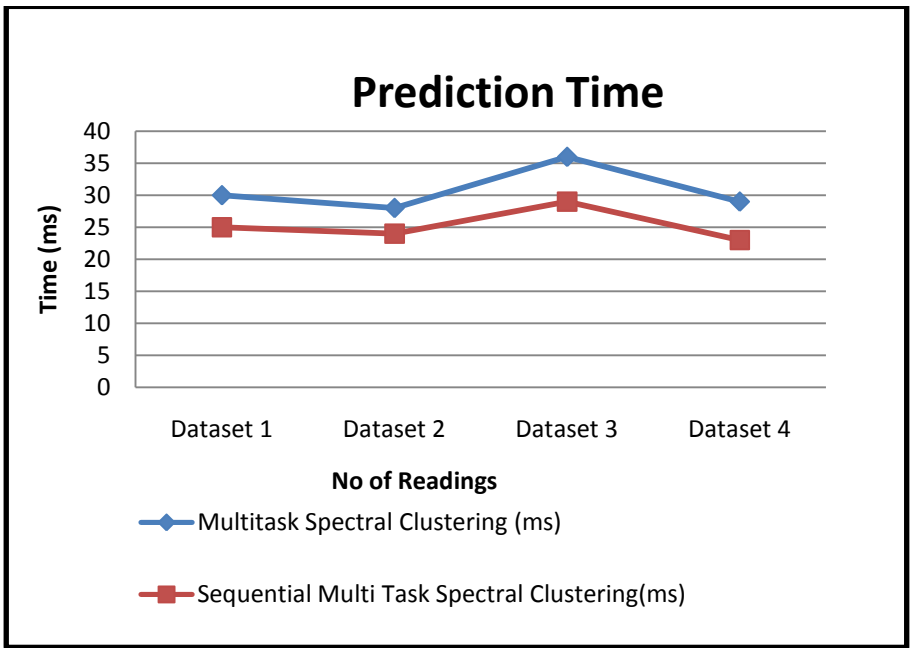

Fig 4.2 Prediction time chart

Accuracy Comparison:

Accuracy of sequential inter task clustering is compared with the existing MTSC scheme and the results show that the proposed scheme gives better accuracy. The following table 4.3 and fig 4.4 shows the implemented result.

Label Prediction Accuracy	**Multitask Spectral Clustering %**	**Sequential Multi Task Spectral Clustering %**
	70	90

Table 4.3

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. 3, Issue 10, October 2015

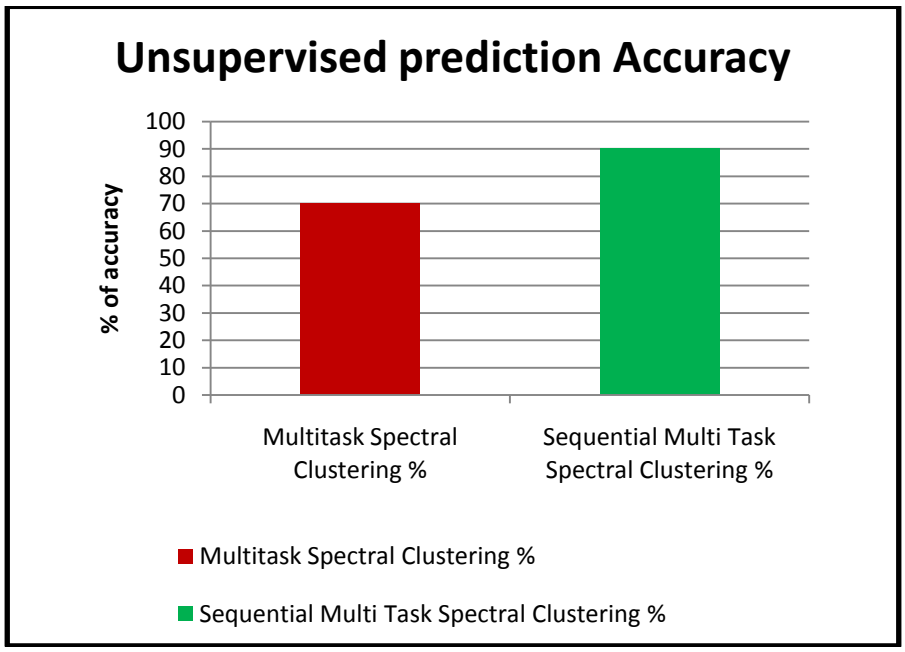

Fig 4.4 Accuracy chart

V. CONCLUSION

In this study, we proposed a new clustering model that is sequential inter task clustering namely BSAS, to cope with the emerging challenges faced by traditional clustering approaches. In the proposed prediction of clusters is known at priority based order. The dissimilarity and similarity of inter task relation is measured with the distance d (x, C) of a different tasks C and threshold of dissimilarity Θ and the number of maximum clusters allowed q. The prediction strategy is to identify every task that interconnected with different multi task in the existing scheme depending on the distance to the already defined clusters. Moreover, for each individual task, an explicit mapping function was learnt by mapping features to the cluster label matrix. We discussed the connections between our proposed model and several representative clustering techniques, including spectral clustering, *k*-means and DKM. Extensive experiments on various datasets illustrated the advantage of the proposed BSAS model over the state-of the-art clustering approaches. In the future, we intend to explore more properties of the cluster label matrix, to enhance the performance of the current proposal.

REFERENCES

[1] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrodl. Constrained K-means clustering with back- ¨ ground knowledge. In ICML, 2001.
[2] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In NIPS 15, 2003.
[3] S. X. Yu and J. Shi. Grouping with bias. In NIPS 14, 2002.
[4] S. D. Kamvar, D. Klein, and C. D. Manning. Spectral learning. In IJCAI, 2003.
[5] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[6] S. Gordon, H. Greenspan, and J. Goldberger, "Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations," in *Proc. 9th IEEE Int. Conf. Comput. Vis. (CVPR)*, Nice, France, Oct. 2003, pp. 370–377.
[7] S. Yu and J. Shi, "Multiclass spectral clustering," in *Proc. 9th IEEE Int. Conf. Comput. Vis. (ICCV)*, Nice, France, Oct. 2003, pp. 313–319.
[8] M. Filippone, F. Camastra, F. Masulli, and S. Rovetta, "A survey of kernel and spectral methods for clustering," Pattern Recognit., vol. 41, no. 1, pp. 176–190, Jan. 2008.
[9] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Jan. 2006.