A Novel Multi View Clustering Technique for Group the Data Objects in Process Mining

Panila Lokanadham[1], Jayanthi Rao Madina[2]

Final M.Tech Student[1], Asst.professor[2]
[1,2]Dept of CSE, Sarada Institute of Science, Technology and Management (SISTAM), Srikakulam, Andhra Pradesh

Abstract:

Clustering is the process of grouping objects based on some notion of similarity. It is commonly applied for exploratory analysis, segmentation, pre-processing and data summarization. Similarity is dependent on the features describing data. Clustering ensembles are a common approach to clustering problem, which combine a collection of clustering into a superior solution. The key issues are how to generate different candidate solutions and how to combine them. Common approach for generating candidate clustering solutions ignores the multiple representations of the data and the standard approach of simply selecting the best solution from candidate clustering solutions ignores the fact that there may be a set of clusters from different candidate clustering solutions which can form a better clustering solution. Multi view clustering can be applied at various stages of the clustering paradigm. This paper proposes a novel multi-view clustering algorithm that combines different ensemble techniques via various similarity metrics have been used to measure the similarity between data objects. In the novel multi view clustering algorithm contains mainly two techniques; the first technique is used to generate multiple partitions from each of the single view of a multi-view dataset. After completion of multi view of data set we can perform the clusterization process on multi view data set. In this paper we are implementing clustering process we are using K Means algorithm. After completion of clustering process take those clusters and combine those clusters will get an efficient cluster groups. By performing combining the clusters groups we are using cluster based similarity matrix. By implementing those concepts we can improve efficiency for performing the clustering process and also the cluster groups will contains most relevant datasets.

Keywords: *data mining, cluster, cluster based similarity matrix, process mining, dataset, data objects.*

I. INTRODUCTION

Clustering is a key issue in intelligence science and is widely used in the field of artificial intelligence. The technique has been studied for several decades in areas of pattern recognition, machine learning, applied statistics, communications and information theory. It is applied to numerous fields of applications including data mining, text mining, bio-informatics, image analysis and segmentation, data compression, and data classification. Clustering is an unsupervised learning technique for organizing similar objects into different groups. Since it is hard to define the similarity especially in high-dimensional data, thousands of clustering algorithms have been proposed in the last 50 years [1]. As no single clustering algorithm is suitable for all types of problems, researchers have been trying different techniques for combining different clustering algorithms (clustering ensembles) [2-5].

Multi-view clustering explores and exploits multiple views simultaneously in order to obtain a more accurate and robust partitioning of the data than single view clustering. There exist two methods in multi-view clustering: centralized and distributed [6]. Centralized algorithms simultaneously use all views to cluster the data while distributed algorithms cluster each view independently from others, using a single view algorithm, and then combine the individual clustering to obtain a final partitioning. During the past decade, Bickel and Scheffer [7] developed a two-view EM and a two-view spherical k-means algorithm under the assumption that the views are independent. De Sa [8] proposed a two-view spectral clustering algorithm that creates a bipartite graph and is based on the ―minimizing-disagreement‖ idea. Kumar et al. [9] proposed a co-training approach for multi- view spectral clustering, co-regularized multi-view spectral clustering [10] and kernel-based weighted multi-view clustering [11].

The main goal of clustering ensembles is to solve the problem of producing superior clustering solution from given set of clustering solutions. This problem was previously approached by researchers from different angles and so far the best known approach for clustering ensembles is median partition based approach in which a single candidate clustering solution that has the maximum similarity from all candidate clustering solutions is selected as

the final clustering solution. The clustering ensembles methods include two important steps:

1) Generating a set of candidate clustering solutions
2) Combining the set of candidate clustering solutions to generate final clustering solution.

In our evolutionary based clustering approach, step 1 corresponds to an initialization phase in which a set of initial candidate clustering solutions is generated, and step 2 is the evolutionary phase in which the final solution is evolved from the initial candidates. However, by using ensemble clustering can produce a more consistent and more accurate solution. In this paper, propose a novel multi-view clustering framework based on ensemble clustering. It first generates multiple partitions from each of the single view of a multi-view dataset. Clustering algorithms are applied on the different data matrices to obtain partitions of the data.

II. RELATED WORK

Clustering has been extensively studied in the literature in many domains, such as in information retrieval to cluster documents, in bio-informatics [12] to cluster genes, in social network analysis [13] to find communities, etc. The basic idea of merging the clustering results from different algorithms evolved as a different field of study for improvement of clustering results. Combining the clustering results of different clustering algorithms is a new clustering framework that is more robust and less susceptible to the adverse effects of each of the single view clustering algorithm. Most of these clustering frameworks consist of finding similarity/distance matrix on the original dataset and using this to combine data samples into groups or clusters. In recent times a number of authors have proposed different multi-view clustering algorithms [13].

Janssens et al [14] proposed a hybrid clustering method which is based on statistical Meta-analysis using Fisher's inverse chi- Square method. In this technique, the distances of data sources are converted into *p*-value by using CDF. These values computed against a randomized data having similar statistical characteristics. The p-values are then converted into a unified p-value using a logarithmic function, which is then used for clustering. Weighted Hybrid Clustering algorithm [13] proposes two steps for combining multiple similarity matrices: Weighted Kernel Fusion clustering and Weighted Ensemble Clustering.

In the kernel fusion technique, kernel functions are used to compute the similarity matrices in higher dimensions for each of the data view. Hong et.al [15] proposes a novel clustering ensembles method, termed as resampling-based selective clustering ensembles method. The proposed selective clustering ensembles method works by evaluating the qualities of all obtained clustering results through resampling technique and selectively choosing part of promising clustering results to build the ensemble committee. Recently, Azimi and Fern [16] proposed an adaptive cluster ensemble method.

In contrast, some studies also indicated that medium diversity leads to the best performing ensembles. First generates a diverse set of solutions and combines them into a consensus partition P*. Based on the diversity between the ensemble members and P*, a subset of ensemble members is selected and combined to obtain the final output. Strehl and Ghosh [17] have developed three different consensus functions based on hyper graph for ensemble learning: cluster-based similarity partitioning algorithm, hyper graph-partitioning algorithm, and Meta clustering algorithm. Topchy et al. [3] designed a consensus function based on a finite mixture model. The final partition is found as a solution to a maximum likelihood problem for a given clustering ensembles. Ensemble method, SELective Spectral Clustering Ensemble (SELSCE), is proposed [18].. After the generation of component clustering, the bagging technique, usually applied in supervised learning. Randomly pick part of the available clustering's to get a consensus result and then compute normalized mutual information (NMI) between the consensus result and the component clustering.

III. PROPOSED SYSTEM

In this paper we are proposed a novel multi view clustering algorithms for getting efficient cluster result. Before apply the novel multi view clustering algorithm, we are represent data object into multiple view. The representing multi view of data objects depending upon number of data object in the input dataset. In this paper we are convert the total input dataset into three multiple view data object. After completion of multi view process apply the clustering process on the multi view data objects. By performing clustering process we are get related data objects in to groups. But those groups are not efficient groups for related data object. So that we are again apply the technique for cluster based similarity matrix by getting most relevant documents into groups. In this paper we are implementing K Means Algorithm for clustering multi view data object and grouping related cluster groups by using cluster based similarity matrix technique. The implementation procedure of proposed system is as follows.

Represent Data Object In Multi View:

In this module we are read the input dataset from the database and perform the multi view of data object. In this paper the input dataset can be split into three multi view data object. Before

performing the multi view process we are count the number of dataset in the input dataset. By taking that we can divide all the input dataset into three multi view partition. Take the count of number of data objects in the data set and divide by three we can get all equal partition of data object. By performing this process we can easily perform the clustering process and also get most relevant data object into groups.

K Means Clustering Algorithm:

Clustering is a technique to categorize the data into groups. Distance metrics plays a very important role in the clustering process. The more the similarity among the data in clusters, more the chances of particular data-items to belong to particular group. There are number of algorithms which are available for clustering. In general, K-means is a heuristic algorithm that partitions a data set into K clusters by minimizing the sum of squared distance in each cluster. The algorithm consists of three main steps: a) initialization by setting center points (or initial centroids) with a given K, b) Dividing all data points into K clusters based on K current centroids, and c) updating K centroids based on newly formed clusters. It is clear that the algorithm always converges after several iterations of repeating steps b) and c). In this paper, the simulation of basic k-means algorithm is done, which is implemented using Manhattan distance metric.

Let X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1. Select ‘c’ cluster centers randomly.

2. Calculate the distance between each data point and cluster centers using the Manhattan distance metric as follows

$$D=X_{ik}- Y_{jk}$$

3. Data point is assigned to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4. New cluster center is calculated using:

$$Vi=1/Ci\sum_{1}^{Ci}(Xi)$$

where, ‘ci’ denotes the number of data points in ith cluster.

5. The distance between each data point and new obtained cluster centers is recalculated.

6. If no data point was reassigned then stop, otherwise repeat steps from 3 to 5.

After completion of clustering process we are get individual cluster groups of multi view data object. So that each view contains specified number of clusters groups and take those cluster groups and perform cluster base similarity matrix we are getting most relevant data object of specified cluster groups are generated. The implementation procedure of cluster based similarity matrix is as follows.

Cluster based Similarity Matrix:

In this module we can generate all relevant data objects of specified number of clusters. Before apply this process we can read all view of cluster groups and apply cluster based similarity matrix technique for getting most relevant data object. The implementation procedure of cluster based similarity matrix is as follows.

1. Read the all views cluster based groups and randomly choose the centers as specified number of clusters.

2. Take the each data object from the each view oriented cluster groups and count the numbers of fields are related to each centroid.

3. By calculating number of equal attribute and take most number of matches to respect centroid.

4. The data object will be stored into respect centroid and take another other data object, perform the above process.

5. Repeated steps 2 to four until all data objects are completed and those groups are contains most related data object.

Based on these similarity matrices on the individual datasets and aggregates these to form a combined similarity matrix, which is then used to obtain the final clustering. By applying those techniques we can get more related data into groups and also improve the efficiency by performing the clusterization process. The results show that our method significantly outperforms other methods.

IV. CONCLUSIONS

In this paper we are proposed an ensembles novel multi view algorithms for generating relevant cluster groups of data object. Before getting this result we can take input dataset and represent that data set into multi view data objects. After completion of multi view process take those multi view oriented data set and apply clustering process each view data object. By performing clusterization process we are using k means clustering algorithm. After completion of clusterization process we are get multiple clusters in multiple view orient. Each view contains group of data objects specified in the number of clusters. Take those clusters of multi view based data object and apply the cluster based similarity matrix we can get specified number of groups with contains more relevant data object. So

that by applying those methods we can perform an efficient clusterization process and also most relevant data object in specified cluster groups.

REFERENCES

[1] K. Jain, ―Data clustering: 50 years beyond k-means‖, *Pattern Recognition Letters, vol. 31, no. 8, pp. 651– 66, 2010.*

[2] Strehl and J. Ghosh, ―Cluster ensembles—a knowledge reuse framework for combining multiple partitions‖ ,*The Journal of Machine Learning Research, vol. 3, pp. 583–617, 2003.*

[3] A. Topchy, A. K. Jain, and W. Punch, ―A mixture model of clustering ensembles‖, in *Proc. SIAM Intl. Conf. on Data Mining. Citeseer, 2004.*

[4] A. Goder and V. Filkov, ―Consensus clustering algorithms: Comparison and refinement‖, in *ALENEX, vol. 8, 2008, pp. 109–117.*

[5] X. Wang, C. Yang, and J. Zhou, ―Clustering aggregation by probability accumulation‖, Pattern Recognition, *vol. 42, no. 5, pp. 668–675, 2009.*

[6] S. Vega-Pons, J. Correa-Morris, and J. Ruiz-Shulcloper, ―Weighted partition consensus via kernels‖, Pattern Recognition, *vol. 43, no. 8, pp. 2712–2724, 2010*

[7] S. Bickel and T. Scheffer, ―Multi-view clustering‖, Proceedings of the 4th*IEEE International Conference on Data Mining, pp. 19-26, 2004*

[8] V.R. De Sa,― Spectral clustering with two views‖, Proceedings of the 22th*IEEE International Conference on Machine Learning, pp. 20-27, 2005.*

[9] A. Kumar and H. Daume, ―A co-training approach for multiview spectral clustering‖, Proceedings of the 28th*IEEE International Conference on Machine Learning, pp. 393-400, 2011*

[10] A. Kumar, P. Rai and H. Daume, ―Co-regularized Multi-view Spectral Clustering‖, Proceedings of the 12th IEEE International Conference on Data Mining, pp. 675-684, 2012.

[11] G. Tzortzis and A. Likas, ―Kernel-based Weighted Multi-view Clustering‖, Proceedings of the 12th IEEE International Conference on Data Mining, pp. 675-684, 2012.

[12] Frings, O., Alexeyenko, A., Sonnhammer, E.L. (2013), ―MGclus: network clustering employing shared neighbours‖, Molecular BioSystems.

[13] Tang, L., Wang, X., Liu, H. (2010), ―Community detection in multi-dimensional networks‖, *Technical Report, Defense Technical Information Center.*

[14] Janssens, F., Gl¨anzel,W., De Moor, B. (2007), ―Dynamic hybrid clustering of bioinformatics by Incorporating text mining and citation analysis‖, In Proceedings of the 13th *ACM SIGKDD international conference on knowledge discovery and data mining (pp. 360–369). New York.*

[15] Hong, Yi, Sam Kwong, Hanli Wang, and QingshengRen, ―Resampling-based selective clustering ensembles‖, *Pattern recognition letters 30, no. 3 (2009): 298-305.*

[16] Azimi, J., & Fern, X. (2009, July), ―Adaptive Cluster Ensemble Selection‖, *InIJCAI (Vol. 9, pp. 992-997)*

[17] A. Strehl and J. Ghosh, ―Cluster Ensembles-A Knowledge Reuse Framework for Combining Multiple Partitions‖, Journal of Machine Learning Research, pp. 583- 617, 2002.

[18]. Jia, Jianhua, Xuan Xiao, Bingxiang Liu, and Licheng Jiao, ―Bagging-based spectral clustering ensemble selection‖, Pattern Recognition Letters 32, no. 10 (2011): 1456-1467.

BIOGRAPHIES:

Panila Lokanadham is student in M.tech (CSE) in Sarada Institute of Science Technology and Management, Srikakulam. He has received her B.tech (CSE) from Prajna Institute of technology and management, RamakrishnaPuram, Palasa. His interesting areas are data mining and network security

JayanthiRao Madina is working as a HOD in Sarada Institute of Science, Technology and Management (SISTAM), Srikakulam, Andhra Pradesh. He Pursuing his Ph.D. from Krishna University, Machilipatnam. Andhra Pradesh. His research areas include Image Processing, Computer Networks, Data Mining, and Distributed Systems. He published six papers in international journals and he attended for three conferences.