Auto-Encoder Based K-Means Clustering Algorithm

VenubabuRachapudi, S VenkataSuryanarayana, T.SubhaMastan Rao

Abstract*: - Identifying network, or clusters, in graphs is a task of great importance when analyzing network structure.A telecommunications provider, for instance, would like to identify communities of customers that place a large amount of calls to each other, in order to create more effective, directed marketing campaigns. Linear or non-linear data transformations are widely used processing techniques in clustering.There are wide range of algorithms and methods that can be applied, in order to identify communities in network structures, such as Spectral Clustering, Modularity Maximization, and Hierarchical Clustering etc. Spectral Clustering is one of the traditional algorithms that is suitable for network clustering because it first map the original data points into space such that graph tend to be much more evident, allowing for a subsequent application of standard clustering techniques such as K-Means. Despite its good results, Spectral Clustering presents some computational challenges when applied to very large networks.Recent literature demonstrates that systems and calculations from Deep Learning, for example, layered neural systems based auto-encoders, are appropriate for the task of mapping information focuses into lower-dimensional spaces, which can be helpful for an assortment of further undertakings. In this proposed paper we will make another handling pipeline for non-covering network identification in system structures dependent on K-Means. We will demonstrate that this methodology is like a conventional profound learning auto-encoder in its capacity to obtain reasonable portrayals of the remarkable information in a lower-dimensional space, making it less demanding to play out the information bunching undertaking. We will at that point test its appropriateness for the particular difficulties of network recognition in systems and contrast its execution and the conventional Spectral Clustering approach.

Index Terms*: Auto Encoder, *Deep learning, Graph Cluster, K-means Clustering, Spectral Clustering.

I. INTRODUCTION

Clustering plays a major role in the field of data science and decision support systems.It has been used in various applications since its inception for achieving better and optimal solutions. Clustering is a learning technique which is unsupervised, i.e. learning takes place without the help of any training dataset [1].Clustering separates the data objects into meaningful groups or clusters based on proximity measures among the data objects [2]. This technique of grouping objects, based on similarity measures, is used for grouping documents, separating images and for many other such tasks. For this reason, clustering plays an important role in the field. of pattern recognition. There is not a universal, unique definition of what a community is in the context of network analysis. A very comprehensive study on this topic was performed in [3]. The exact definition will depend on the specific system and/or application considered.It is obvious that results will change from application to application and hence it shall be kept in mind that all the clustering methods cannot be used for all sorts of applications. Nevertheless, a general intuition is that a community is constituted by a group of cohesive nodes, in the sense that there are so many links inside that cohesive group than links connecting that group to nodes outside it.

Consider for example the network depicted by the graph shown in Fig.1 with a set of vertices (Nodes) and a set of edges (or links). We would like to find One ormore community defined as a subset of nodes connected which are denser than the nodes connected with the rest of the network. We consider the internal and external degrees of a node.

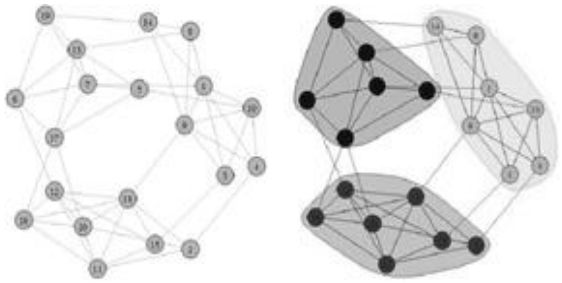

Fig. 1: Graph to represent good communities

A. K-MEANS CLUSTERING

K-means [4] is one of the modest unsupervised learning algorithms and trails the clustering partitioning method. This is most used in applications where clustering of objects in a specified manner is required. Here clustering i.e. grouping of objects is based on k as number of clusters and it depends on a specified and customized algorithm as per requirement. Here k objects that represent the initial cluster center or mean begin with randomly selected objects. Next, the cluster is

Revised Manuscript Received on March 10, 2019.

VenubabuRachapudi,Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,Vaddeswaram, Guntur, Andhra Pradesh, India.

S VenkataSuryanarayana, Department of Information Technology, CVR College of Engineering, Hyderabad,India.

T.SubhaMastan Rao, Department of Computer Science and Engineering, CMRTechnical Campus, Hyderabad, India.

*Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication*

assigned to each object grounded on the familiarity of the object to the cluster centre. In order to allocate the object to the nearest centre, an amount of proximity specifically

Euclidean distance is used to quantify the concept of neighboring. Euclidean distance is a standard and old distance measure used to calculate the distance between two pair of objects or group of objects. It is a simple mathematical tool utilized in numerous applications to arrive at distance between to co-ordinate points with ease and accuracy. After all the objects have been disseminated to k clusters, the new k cluster centers are originated by the mean of the cluster objects in each cluster. The procedure is repetitive until no alteration occurs in cluster centers or maximum number of iterations have reached. The objective of the K-means algorithm is to reduce the objective function, explicitly the sum of the squared error (SSE). SSE is defined as

$$SSE = \sum_{i=1}^{k} \sum_{e \in C_i} \left| e - r_i \right|^2$$

Here SSE represents the sum of the squares of the distance between cluster centre and the objects in the cluster in all k clusters.Here ri represents cluster centre of ith cluster.

This Algorithm has time complexity which is proportional to knt, where k indicate number of clusters , n represents the number of objects and t represents the maximum number of iterations i.e., O (t*k*n) represents the time complexity of this algorithm.

Although K-means is efficient, It has many limitations. Firstly, The number of clusters k must know in advance which is more problematic for some applications. Secondly, the algorithm seeding with k initial cluster centres is also problematic because inappropriate cluster centre lead the algorithm by local optima. Thirdly, the random selection of initial cluster centre is also considerable problem because each run of the algorithm gives different clusters. Fourthly, the shape of the clusters formed by k-means is always spherical because of distance measure. Fifthly, the algorithm cannot handle missing data and susceptible to outliers.

B. SPECTRAL CLUSTERING

Spectral Clustering [3],[5-7] is one of the traditional and popular method of data clustering that is well suited for the problem of finding communities in networks. When applying Spectral Clustering to an arbitrary dataset, the idea is to represent that dataset by its similarity matrix (or a derivation of it).It has to be taken care that certain prerequisite cautionary measures have to be met such that there won’t be any ambiguity while finding the communities in networks.

Consider for example the graph of Fig. 1. Then a Similarity matrix would be constructed by applying a pairwise Similarity function to every pair of nodes with being symmetric and nonnegative. Spectral Clustering divides the data according to a lower-dimensional representation obtained from the calculation of own vectors of the graph Laplacian matrix obtained from this data, the number of clusters desired. When computing eigenvectors, we consider those corresponding to the smallest non-zero eigenvalues. After embedding the original data into a lower-dimensional representation, the final clustering result is obtained by running K-Means on that representation.

The procedure depicted as follows

1. Compute similarity graph.
2. Compute Laplacian Matrix.
3. Compute first k Eign values of the Laplacian and corresponding Eigen vectors.
4. Construct a matrix containing those eigen vectors as columns which is reduced data.
5. Cluster the data points where each row in the above matrix represents a data point.
6. Project these data points to the original data space for getting resultant clusters

C. AUTOENCODER

Autoencoder is an unsupervised neural network algorithm that uses machine learning to do compression. The aim is to learn compressed version of data typically for data reduction. Even though, we have Principal Component Analysis for the same purpose, Autoencoder can learn liner and nonlinear transformations unlike Principal component Analysis, with nonlinear activation function and having multiple layers. It does not have to learn dense layers. It is more efficient with modern parameters with several layers and gives representation as the output at each layer. It applies back propagation setting the target values to be equal to the given inputs.

The components of Autoencoder include encoder, code and decoder. Encoder is the part of the network that compresses the input into latent space representation having reduced dimensionality. Code is part of the network that represents compressed input fed into the decoder. Decoder is the part of network which aims to construct the input from the latent space representation. The components of Autoencoder are depicted in Fig2. The output is the lossy representation of the original input.

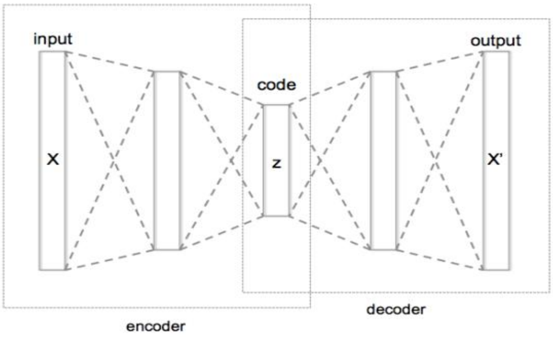

Fig2.Components of the Autoencoder

The remainder of the paper is organized as follows. Section II presents related work to auto-encoder based clustering. Proposed algorithm is presented in section III. Experimental results elaborated in section IV. At the end, the conclusion of the paper is detailed in section V.

II. RELATED WORK

We found a lot of literature related to Autoencoder based clustering for many applications includes segmentation of images.C.Song et.al.

[8] proposed Auto-encoder based data clustering which learn a high nonlinear mapping function and used for hand written digits images and faces images.Song Chunfeng et.al.[11] proposed Deep auto-encoder based clustering algorithm for hand written digits images and faces images and achieved good results.FeiTian et.al.[12] proposed Learning Deep Representations for Graph Clustering and tested with several benchmarked algorithms on several data sets.Yimin Duet.al.[13] proposed Auto-encoder Based Clustering Algorithms for Intuitionistic Fuzzy Sets and evaluated on well-known benchmarked data sets Iris and Wine and proved the proposed method outperforms baseline algorithms.

III. PROPOSED ALGORITHM

Proposed Implementation for Auto-encoder based K-Means

1. Consider network data represented wirh a graph G=(V,E), where V is the set of nodes and E is the set of edges
2. Compute pairwise similarity matrix S from G, using similarity metric.
3. Compute a diagonal matrix D
4. Consider the Autoencoder with few hidden layers and each layer uses K-Means algorithm with random centroid assignment.
5. Extract k cluster communities at decoder.

IV. RESULTS

we have applied our algorithm on a large set of artificial, computer-generated graphs as depicted in Fig. 1 to check its performance.

we generated synthetic data from the approach proposed in [9]. Table 1 below summarizes some of the main characteristics of the tested networks which incorporates important features like the heterogeneity of node degree distributions and community sizes. We also applied our algorithm on the Zachary's Karate Club Study dataset and Football dataset as described in [10].

Graph ID	nodes	edges	max-degree	min-degree	avg path length	communities	largest community	smallest community
1	1000	7468	45	6	2.9	25	39	15
2	2000	12987	53	8	3.2	35	47	22
3	3000	19563	57	9	3.7	45	54	23
4	4000	26342	64	7	4.1	56	63	19
5	5000	35378	76	8	3.4	74	68	21

Table 1.Description of the synthetic networks

Table2. Comparison between Spectral Clustering and Autoencoder based Clustering algorithm

Dataset	Spectral Clustering		Auto-encoder based	
	NMI	Accuracy	NMI	Accuracy
synthetic(Graph ID-1)	0.56	0.53	0.59	0.57
synthetic(Graph ID-2)	0.68	0.65	0.71	0.69
synthetic(Graph ID-3)	0.71	0.67	0.73	0.7
synthetic(Graph ID-4)	0.64	0.62	0.69	0.65
synthetic(Graph ID-5)	0.65	0.63	0.72	0.69
Zachary's study	0.69	0.66	0.77	0.73
Football	0.68	0.65	0.76	0.72

In all experiments, we considered two parameters for measuring the performance of the algorithm. One is accuracy and another one is the Normalized Mutual Information(NMI) between the communities detected by the algorithm.We ran both Auto encoder based K-Means and Spectral Clustering 10 times for each network with the same setup and computed the average of the elapsed running time. We have given the NMI and corresponding accuracy for each dataset as shown in the following table Table 2.

From the above Table 2 it is clear that our Auto encoder based clustering algorithm outperforms the traditional Spectral Clustering algorithm. We also performed empirical analysis of the time complexity of Auto encoder based K-Means compared to Spectral Clustering.

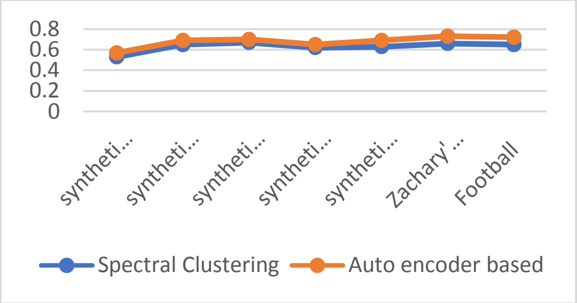

Fig3.Comparison between the Spectral clustering and Auto encoder based K-Means algorithm

The comparison between the Spectral clustering and Auto encoder based K-Means algorithm for different datasets as mentioned in table 1 with respect to accuracy is as shown in the above Fig3.

V. CONCLUSION

In this work, we proposed an algorithm for identification of communities within the given network based on Deep Learning features for embedding data in lower-dimensional spaces. From our experiments, it is clear that the proposed Auto encoder based clustering algorithm outperforms traditional Spectral Clustering in accuracy measured by the Normalized Mutual Information between the communities discovered by the algorithm and the ground-truth communities associated with the respective datasets. It is also observed that for small sized datasets Spectral Clustering algorithm works well compared to Auto encoder based K-Means algorithm. But, for large sized networks our proposed algorithm performs better than the Spectral Clustering algorithm.

The way the algorithm was implemented demands the prior knowledge of the number of communities to be found. This is not a problem for networks with ground truth communities, such as those tested in this work, but it is not the case for most of the practical problems. For future work, we are going to adopt meta-heuristic methods and probability based methods for assigning data to clusters , by which we can avoid the possibility of local optimal solution.

REFERENCES

1. Liu B., Xia Y., Yu P.,"Clustering Via Decision Tree Construction", In: Chu W., Young Lin T. (eds) Foundations and Advances in Data Mining. Studies in Fuzziness and Soft Computing, vol 180. Springer, Berlin, Heidelberg, 2005, pp 97-124.
2. SriparnaSaha et. al."Performance Evaluation of Some Symmetry-Based Cluster Validity Indexes", IEEE Transaction on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol-39, No-4, 2009,pp-420-425.
3. White, Scott & Smyth, Padhraic," A Spectral Clustering Approach To Finding Communities in Graph", Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005. 5. 10.1137/1.9781611972757.25.
4. MacQueen, J,"Some methods for classification and analysis of multivariate observations", Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, University of California Press, Berkeley, Calif., 1967,pp.281—297.
5. Stella X. Yu and Jianbo Shi, "Multiclass Spectral Clustering", Ninth IEEE International Conference on Computer Vision, 2003, pp.1 – 7.
6. U. von Luxburg,"A Tutorial on Spectral Clustering",Statistics and Computing 17(4): 395-416, 2007.
7. Jing Qiu, Jing Peng, Ying Zhai," Network community detection based on spectral clustering", 2014 International Conference on Machine Learning and Cybernetics,2014, DOI: 10.1109/ICMLC.2014.7009685.
8. Song C., Liu F., Huang Y., Wang L., Tan T. (2013) ,"Auto-encoder Based Data Clustering", In: Ruiz-Shulcloper J., Sanniti di Baja G. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2013. Lecture Notes in Computer Science, vol 8258. Springer, Berlin, Heidelberg,pp 117-124.
9. Lancichinetti, Andrea, Santo Fortunato, and FilippoRadicchi. "Benchmark graphs for testing community detection algorithms." Physical Review E 78.4 (2008): 046110.
10. Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." Proceedings of the National Academy of Sciences 99.12 (2002): 7821-7826.
11. Song, Chunfeng& Huang, Yongzhen& Liu, Feng& Wang, Zhenyu& Wang, Liang. (2014). Deep auto-encoder based clustering. Intelligent Data Analysis. 18. S65-S76. 10.3233/IDA-140709.
12. FeiTian, Bin Gao, Qing Cui, Enhong Chen, Tie-Yan Liu," Learning Deep Representations for Graph Clustering" Proceedings of the Twenty-EighthAAAI Conference on Artificial Intelligence,2014,pp.1293-1299.
13. Yimin Du, Guixing Wu, Guolin Tang, "Auto-encoder Based Clustering Algorithms for Intuitionistic Fuzzy Sets",12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE),2017,pp1-6.