CS 6140 Data Mining Project Intermediate Report

Padmashree Teeka (u0880562), Roshani Nagmote (u0941394), Anirudh Narasimhamurthy (u0941400)

March 23, 2015

Comparing performances of different Clustering Techniques on the dataset.

1 What progress you have made towards your proposed goal?

Our proposed goal was to perform different clustering techniques on social network dataset. We chose LiveJournal dataset as our input data set to work on. We plan on performing 3 clustering techniques on our dataset. They are:

1. Hierarchical Clustering
2. Spectral Clustering
3. K-means Clustering

- **Hierarchical Clustering**
  We started working on the code for the clustering algorithm. We realized that the data set needs to be in the form of an adjacency matrix. But this would mean we will run into space constraints. So we generated Sparse Matrix of the data using Matlab. Our dataset contains 'From' and 'To' nodes. The number of edges is about 34681189. Edges refer to the number of connections among the users. Initially we started working on smaller data sets to see if our algorithm works correctly. We had success with very small data sets consisting of about 10 nodes and 12 edges between them. But when we tried working on slightly larger data set with 1049866 nodes, we are running into 'Running out of memory' error.
  So we have a basic algorithm working on a small set of nodes. We are working on scaling this to larger data sets and resolving Memory issues. ze

- **Spectral Clustering**
  We started working on the Spectral Clustering Algorithm in Matlab. We chose a smaller data set and found the adjacency sparse matrix. When we run our algorithm using adjacency matrix, we got out of memory error. So, we built sparse matrix. Then we created a full symmetric matrix from this. Our program runs for 1049866 nodes dataset but our system crashes when we try to run for our dataset which has 34681189 nodes.

2 If you tried some basic approaches: what worked well and what did not

Our dataset basically has from nodes and to nodes and we are supposed to find communities in this dataset. So, we referred this paper which discusses about community

detection problems with different clustering techniques.

http://snap.stanford.edu/class/cs224w-readings/fortunato10community.pdf

Hierarchical Clustering

To start with the method, we need similarity measure between the nodes of the graph. To find out similarity measure, there are different methods like the cosine similarity, the Jaccard index, and the Hamming distance between rows of the adjacency matrix. We discussed these methods and decided that the best technique to find the similarity between nodes is hamming distance.

$d^{HAD}(i,j) = \sum_{k=0}^{n-1}[y_{i,k} \neq y_{j,k}]$

where, d^{HAD} is the Hamming distance between rows of our Adjacency Matrix.

Hamming distance was found by finding difference between each of the row elements of our adjacency matrix by finding the percentage of the nodes that differ. This distance vector was reformed into a Square matrix for easier computation. This is a symmetric Matrix with diagonal elements as zero.

Initially we grouped nodes which are close together into clusters then we grouped the obtained clusters with other nodes till we got the larger cluster until all the nodes came under one cluster. Using dendrogram we plotted the clusters for our sample data set.

When we were trying to create the adjacency matrix, we tried using the sparse function. This function takes the maximum out of the number of nodes and edges. This function worked for our sample data set of 8 nodes but when we tried on a bigger data set of 1049866 nodes it failed. So we explored other options and used Accumarray function to create the sparse matrix. This was used to calculate the symmetric matrix from which Hamming distance was calculated.

– **Spectral Clustering**

We tried to implement the Spectral Clustering to a graph dataset to start with and found interesting results during the experiment. .Since our dataset is a graph containing from and to nodes describing the realtionship among different people in the network, performing a *top-down* approach to clustering looked atleast easier to start with. But we soon ran into different issues right from building our adjacency and degree matrix for our given graph in order to apply the Spectral clustering. While the dataset is decribed as an undirected graph, for the sake of simplicity and to reduce the size of an already big data file, thwe rom and two nodes for any given undirectedoundge ws represented by only a single edge list pair. Hence after loading the datatset into an vector corresponding to the second soperations to ensure the adjacency matrix was constructed taking i to be a very good descriptor for the gwas an undirected graph.
ly bigger dataset, we also found that constructing a full adjacency matrix might not be possible due to the limit on mand array dimensions and hence we constructed a sparse adjacency matrix.
Since we want the Volume for each cluster to be as large as possible and cut to be smaller

between a pair of clusters, we tried finding the cluster which has the minimum nornamlized cut. Based on the number of clusters 'k' provided as input, we found the k-clusters. To obtain the clusters, we constructed the degree matrix for the given dataset and then constructed the un-normalized Laplacian matrix.

Among the several different eigen vectors of the Laplacian, we found the Fiedler vector, which is eigen vector corresponding to the second smallest eigenvalue of the Laplacian matrix to be a very good descriptor for the graph. We used simple commands in MATLAB to obtain the Laplacian and the fiedler vector. Our code rerturns both the fiedler vector and the *k partitioned groups of nodes*. This gives us an indication about where to make the cut and which vertices belong to which cluster.

3 What could be done to improve the basic approaches?

We have chosen following Clustering techniques for the above dataset:

- K Means Clustering
- Hierarchical/Agglomerative Clustering
- Spectral Clustering

4 What experiments have you run and are you planning to run to demonstrate the effectiveness

We will use clustering techniques for finding communities in the dataset. Communities will consist of user defined groups such that We found our spectral and hierarchical clusters who have joined a specific group will maton is important for finding groups of usters having similar interests. This would help us in predicting which community the new user might like to join. many different clustering techniques available, we will compare their performance wigraphs, we are planning to run it on large graph datasets of social networks and would be interested to see the respect to the above dataset.