A STUDY ON ANALYSIS OF GRAPH CLUSTERING TECHNIQUES IN FEATURE SELECTION

S.DeepaLakshmi
Research Scholar, Bharathiar University, Coimbatore, India
T.Velmurugan
Associate Professor, PG and Research Department of Computer Science,
D.G. Vaishnav College, Chennai, India

Abstract: Graph Clustering is the task of grouping the vertices of a graph into clusters. The grouping is based on similarity measure defined for the data elements. The field of graph clustering has become popular nowadays. Feature selection or extraction is a technique that transforms and simplifies the data to make data mining tasks easier. Feature selection removes the irrelevant and redundant features and selects the relevant and useful features that provide an enhanced classification results as the original data. This research work presents about the application of graph clustering in feature selection of high dimensional data. Also, this work aims at clustering the features using graph theoretical concepts. The irrelevant features are removed and the relevant features are grouped into clusters using minimum spanning tree in this work. The main contribution of this work is to select a representative feature from each resulting cluster to form the set of relevant features.

Keywords –Feature Selection, Graph Clustering, Minimum Spanning Tree, Mutual Information.

Introduction : Data Mining is the task of discovering interesting patterns from large amounts of data. Mining High Dimensional data has some challenges including the curse of dimensionality and the meaningfulness of the similarity measure in the high dimensional space[1]. Feature selection or attribute selection is the process of selecting relevant features from a large number of features. Feature Selection also known as Attribute Selection or Variable Subset Selection is the process of selecting the most relevant subset of attributes from large set of attributes according to some selection criteria[2]. Some of the benefits of Feature Selection are facilitating data visualization and data understanding, reducing the measurement and storage requirements, reducing training and utilization times of the final model, defying the curse of dimensionality to improve prediction and performance improvement.

Many algorithms exist that select optimal features from high dimensional dataset. The use of graph theory to feature selection has gained momentum in the recent research works. This research work aims at reducing the feature set from a large and high dimensional dataset using the concepts of graph theory. In graph based clustering methods, similar data are represented in a graph. The highly connected sub graph forms the clusters. The elements in a cluster are highly similar to each other. In this research work, irrelevant features are removed using filter method. The relevant features are grouped into clusters using graph based clustering methods in the second step. Also, from each cluster, one strong representative feature is selected. Thus, the numbers of features are reduced and relevant features are selected. The rest of the paper is organized as follows: in section 2, the application of graph theory in data mining is discussed. In section 3, graph clustering algorithms are presented and in section 4, experimental results are discussed and section 5 summarizes the research contribution and concludes.

Graph Theory in Data Mining : Graph is an ordered pair of vertices and edges. A graph structure can be extended by assigning weight to each edge of the graph called weighted graphs[3]. Most of the data in data science can be modelled as graphs. Graphs can be mined using algorithms in graph theory to understand them better. Graphs can be used to model relations and processes in biological, social and information systems. Graphs are used to represent networks of communication, flow of computation etc. It is highly used by computer science applications like data mining, image segmentation, networking etc. Numerous graph algorithms are used to solve graph theoretic concepts which solve

computer science applications problems[4]. Some of the algorithms are shortest path algorithm in network, finding a minimum spanning tree, finding graph planarity, algorithms to find adjacency matrices etc.

In machine learning, classification is the task of classifying the objects into classes and graph classification is the task of predicting the label of an input graph[5]. Graph classification also refers to graph clustering as it is the task of grouping objects into classes. The approach to graph classification is to reduce it by feature extraction to the task of classifying or grouping vectors of attribute-value pairs. Classifiers require a similarity measure and clustering algorithms are able to work on distance or similarity matrix as input[6]. The task of Label prediction is to classify the nodes of a partly labelled graph that has application in image processing, fraud detection etc. Many methods for label prediction are proximity based. The task of frequent item set mining and frequent subgraph mining tries to find subgraphs that are contained in a large graph database.

Graph Clustering : The aim of clustering is to divide the data set into clusters in which the elements within a cluster are similar. Graph Clustering involves the task of grouping the vertices of a graph into clusters in such way that there are many edges within each cluster and few edges between the clusters[7]. It is concerned with finding the densely connected group of nodes in a graph. The goal of graph clustering is to infer groups of closely related nodes given the similarity or dissimilarity observations encoded in the graph[8]. The algorithms used for Graph Clustering are discussed in this chapter.

Minimum Spanning Tree : Minimum spanning tree is a subset of the edges of a connected graph that connects all vertices without any cycle and has a minimum sum of weights over all the included edges[9]. It has applications in cluster analysis, computational biology, broadcasting in computer networks, image segmentation and circuit design. The weight can represent either distance or similarity of the two vertices. Two algorithms are used to find the minimum spanning tree namely Kruskal's algorithm and Prim's algorithm. Kruskal's Algorithm builds the spanning tree by adding the edge with smallest weight. Only the edge which doesn't form a cycle is added. Prim's algorithm grow the spanning tree from a starting position adding the cheapest vertex to the spanning tree[10].

Shared Nearest Neighbour : A proximity or similarity graph is a graph obtained by connecting two objects that are similar to each other in some sense. Proximity measures are used to determine the extent to which two vertices belong to a group. Proximity forms the basis for shared nearest neighbour measure. The principle is that if two vertices have more than k neighbours in common, they are considered similar to one another. Shared nearest neighbour denotes the number of neighbour nodes common between any given pair of nodes[11]. One of the main advantages of shared nearest neighbour is that it can find similarities between vertices that are not adjacent.

Betweenness Centrality Based : Betweenness centrality is a measure of centrality in a graph based on shortest path. It measures the extent to which a vertex lies on paths between other vertices[12]. It finds wide application in network theory, social networks, biology and scientific cooperation. There are two types: Vertex betweenness and edge betweenness. Vertex betweenness is defined as the number of shortest paths in the graph that pass through a given node. Edge betweenness is the number of shortest paths in the graph that pass through given edge.

Highly Connected Components : Highly Connected Subgraphs known as HCS algorithm is an algorithm based on graph connectivity for cluster analysis. It represents the similarity data in a similarity graph and finds all highly connected subgraphs as clusters. Minimum cut is the minimum set of edges whose removal disconnects a graph. HCS algorithm finds the subgraphs with n vertices such that the minimum cut of the subgraphs contain more than n/2 edges and it is identified as clusters[13].

Maximal Clique Enumeration : A Clique is a fully connected subgraph in a finite, simple graph. A clique is maximal if it cannot be augmented by adding additional vertices[14]. Bron-Kerbosch algorithm is an efficient method for finding maximal cliques in an undirected graph. Maximal cliques are

important in graph theoretic applications, including graph coloring and fractional graph coloring.

Results and Discussion : The irrelevant features are removed by using mutual information, a filter method. The redundant features are removed by constructing a minimum spanning tree and retrieving a forest from the minimum spanning tree. Each tree is a cluster of features in the forest. A representative feature from each cluster is selected to form the set of relevant features. The irrelevant feature removal obtains features relevant to the class by eliminating the features which are irrelevant to the target class. Feature relevance is measured in terms of feature correlation. Relevant features have a strong correlation with the target class. Mutual Information measures the mutual dependence between two variables and can handle both linear and non-linear relationship. If two features are independent, the mutual information between them is zero, and if the two features are highly dependent, the mutual information is large[15]. So, mutual information is chosen as the measure of correlation between the feature and the class variable. The mutual information is defined as follows:

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)} \right) \quad \text{....(1)}$$

Where X and Y are two features, p(x,y) is the joint probability distribution function of X and Y, p(x) and p(y) are the probability distribution functions of X and Y. Mutual Information is always greater than or equal to zero. The features whose mutual information values are greater than a particular threshold value comprise the relevant feature subset.

Redundant feature removal removes features that are redundant in 3 steps: Constructing a minimum spanning tree from the relevant features, grouping the features in the forest into clusters and selecting the representative feature from each cluster. Mutual information between each pair of features f_i' and f_j' is calculated as $MI(f_i',f_j')$. A complete Graph G= (V, E) is constructed where V is the set of features from relevant feature subset and E is the mutual information $MI(f_i',f_j')(i \neq j)$ which is the weight of the edge between the vertices. For a high dimensional data, the graph G is heavily dense, and thus a minimum spanning tree is built. The mutual information between each pair of features $MI(f_i',f_j')$ is compared with the mutual information between each feature and the class variable. If $MI(f_i',f_j')$ is less than the mutual information $MI(f_i',C)$ and $MI(f_j',C)$ where C is the target class, then the edge (f_i',f_j') is removed.

Each deletion results in a disconnected tree and a forest is obtained. A biograph is constructed for visualization of the forest. Each collection of disconnected trees in the forest represents a cluster. The forest is then traversed, and the mutual information of the features in each cluster with the class variable is determined. The feature that has the maximum $MI(f_i',C)$ is selected as the representative feature from each cluster. The set of the representative feature from each cluster forms the subset of features which is strongly relevant to the class variable.

Data set	No. of features	No. of instances	No. of Classes	Domain
Arcene	10001	200	2	Microarray
SMS spam	1833	5574	2	Text

Publicly available data set is used in this work. Arcene dataset is a microarray dataset with 10000 features, 200 instances and the number of classes is two. SMSSpam dataset is a text dataset with 1833 features and 5574 instances and it is a two class classification problem.
Table 1: Data Set Description

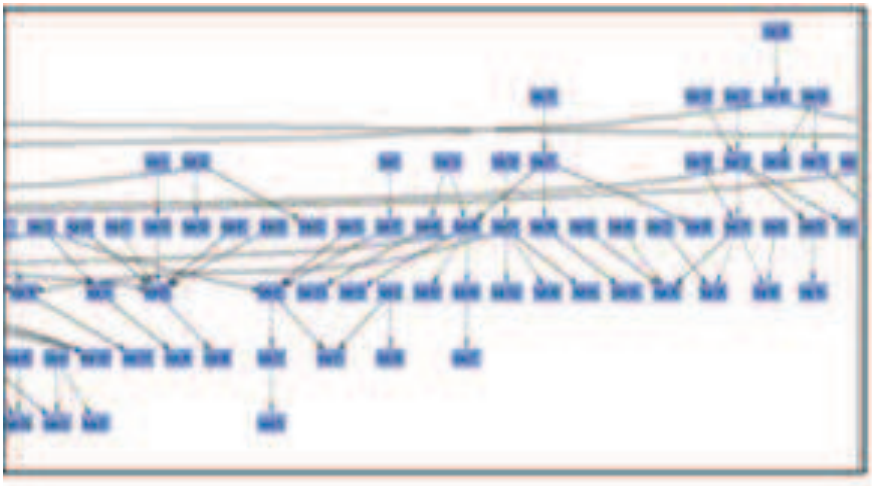

Figure 1: Biograph Image Showing The Minimum Spanning Tree Of The Arcene Dataset

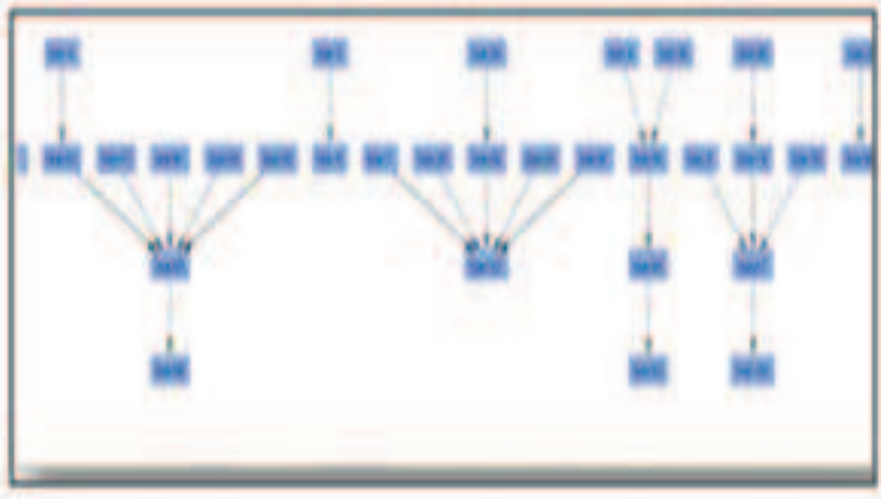

Figure 2: Biograph Image Showing The Forest Of The Arcene Dataset

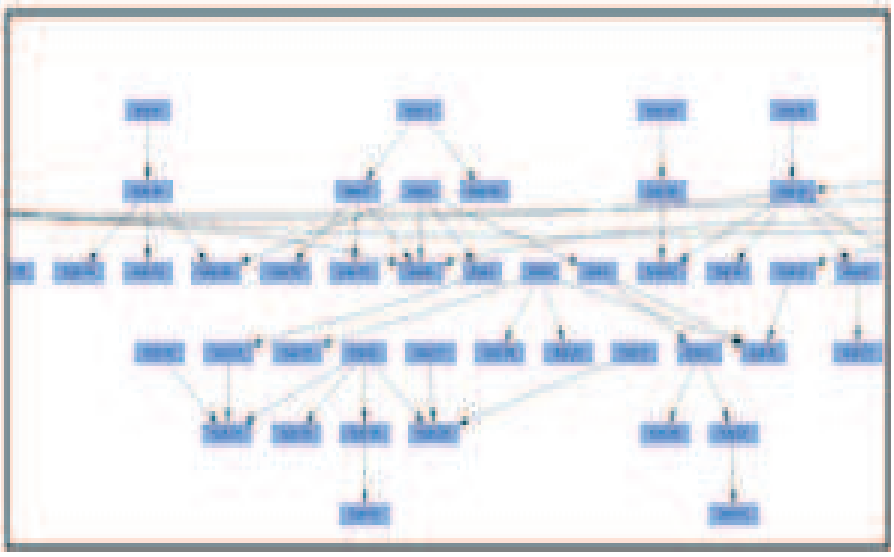

Figure 3: Biograph Image Showing the Minimum Spanning Tree of the SMSSpam Dataset

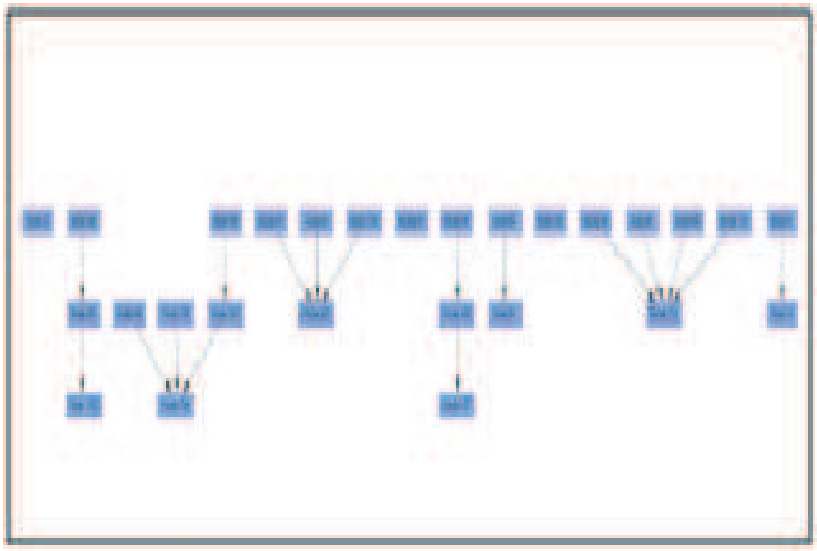

Figure 4: Biograph Image Showing the Forest of the SMSSpam Dataset

The irrelevant features are removed using mutual information. The feature whose mutual information is lesser than the threshold value is removed. A graph is constructed with the features as the vertices and the mutual information between the features as the edges. Minimum spanning tree of the graph G is constructed. Biograph is a tool used in MATLAB to display the minimum spanning tree as shown in figure 1 and figure 3. The arcene dataset has 10001 features and 200 instances. The irrelevant features are removed, and a minimum spanning tree with relevant features is constructed. Figure 1 shows the minimum spanning tree constructed for the arcane dataset with 1067 nodes and 1045 edges. The nodes indicate the features and the edges indicate the mutual information between the features. A part of the minimum spanning tree is shown in Figure 1, in which the node numbers specify the feature number. From the minimum spanning tree, the redundant nodes are removed and a forest is constructed for Arcene dataset and SMSSpam dataset as shown in the figure 2 and figure 4. Each tree in the forest is a cluster. From each cluster, a representative feature is selected. The feature which has the highest mutual information value of the feature and the class is selected as the representative feature from each cluster. Thus, the number of features is reduced and the relevant features are retrieved from a high dimensional dataset which has huge number of features.

Conclusion : Feature Selection of high dimensional data poses a serious problem in machine learning due to the curse of dimensionality. Graph theoretic concepts are utilized for clustering and selecting the relevant features from a large set of features. Graph Clustering is the clustering of elements in a graph into groups. A Graph is constructed with the features as vertices and the mutual information between the features as edges. A minimum spanning tree is constructed and a forest is deduced eliminating the redundant features. The features are thus grouped into clusters and a representative feature from each cluster is selected that forms the set of relevant features.

References:

1. L. Yu and H. Liu, “Efficiently handling feature redundancy in high-dimensional data,” Proceedings of the ninth ACM SIGKDD International Conference Knowledge Discovery data Mining - KDD ’03, pp. 685, 2003.
2. Guyon, I., & Elisseeff, A,” An Introduction to Variable and Feature Selection. Journal of Machine Learning Research (JMLR)”,vol.3 issue 3, pp.1157–1182, 2003
3. M. E. J. Newman, “Analysis of weighted networks", Physical review E 70, no. 5 , 2004.
4. S. G. Shirinivas, S. Vetrivel, and N. M. Elango, “Applications Of Graph Theory In Computer Science An Overview,” International Journal of Engineering Science and Technology, Vol. 2, no. 9, pp. 4610–4621, 2010.
5. L. Getoor and C. P. Diehl, “Link Mining: A Survey", Acm Sigkdd Explorations Newsletter, Vol.7, no.2,

pp.3-12, 2005.
6. E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell, "Distance Metric Learning, with Application to Clustering with Side-Information", In Advances in neural information processing systems, pp. 521-528. 2003.
7. S. E. Schaeffer, "Graph clustering," Comput. Science Review, Vol. 1, no. 1, pp. 27–64, 2007.
8. Y. Chen, S. Hong Lim, and H. Xu, "Weighted Graph Clustering with Non-Uniform Uncertainties", In International Conference on Machine Learning, pp. 1566-1574, 2014.
9. O. Grygorash, Y. Zhou, and Z. Jorgensen, "Minimum Spanning Tree Based Clustering Algorithms", In Tools with Artificial Intelligence, 18th IEEE International Conference, pp. 73-81. IEEE, 2006.
10. Q. Song, J. Ni, and G. Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data", IEEE transactions on knowledge and data engineering, vol.25, no.1, pp.1-14, 2011.
11. K. A. Wilson et al., "Graph-based Proximity Measures", Practical Graph Mining with R, p.135.
12. W. Cukierski, B. Hamner, and B. Yang, "Graph-based Features for Supervised Link Prediction", In Neural Networks (IJCNN), The 2011 International Joint Conference, pp. 1237-1244, 2011.
13. P. Williams, "Clustering Using Graph Connectivity," 2010.
14. J. D. Eblen, C. A. Phillips, G. L. Rogers, and M. A. Langston, "The maximum clique enumeration problem: algorithms, applications, and implementations.," BMC Bioinformatics, vol. 13 Suppl 10, no. Suppl 10, p. S5, Jun. 2012.
15. W. Li, "Mutual information functions versus correlation functions," Journal of Statistical Physics, vol. 60, no. 5–6, pp. 823–837, 1990.

***