JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-3, Issue-6 E-ISSN: 2347-2693

Correlated Probabilistic Graph with Clustering

Sawant Ashlesha G.

Department Of Computer Engineering,
JSPM's Imperial College Of Engineering and Research, Wagholi.

Email Id:sawantashlesha3@gmail.com

www.ijcseonline.org

Received: May /30/2015 Revised: June/09/2015 Accepted: June/22/2015 Published: June/30/ 2015

***Abstract*—** Recently, probabilistic graph have more interest in the data mining. After some result it is found that correlations may exist among adjacent edges in various probabilistic graphs. As one of the basic mining techniques, graph clustering is widely used. Different Clustering methods are used. But, when correlations are considered, it becomes more challenging to efficiently cluster probabilistic graphs. Here, we define the problem of clustering correlated probabilistic graphs and its techniques. To solve the challenging problem the PEEDR and the DPTC clustering algorithm are defined for each of the proposed algorithms, with some several pruning techniques and Different Similarity measures.

***Keywords*—** Clustering; Correlated; Probabilistic Graph; Graph Clustering; Pruning

I. INTRODUCTION

The process of identifying structure in terms of grouping the data elements is called Clustering. The groups generated are called clusters. The grouping is usually based on some similarity measure defined for the data elements. Clustering is having closely relation to unsupervised learning in pattern recognition systems [1]. Graphs are structures formed by a set of vertices (also called nodes) and a set of edges that are connections between pairs of vertices. Graph clustering is the task of grouping the vertices of the graph into clusters by considering the edge structure of the graph in such a way that there should be many edges within each cluster and relatively few between the clusters.

As one of the basic data mining techniques, clustering is widely used in various graph analysis applications [3], such as community detection, index construction, etc. This paper focuses on clustering correlated probabilistic graphs which aims to partition the vertices into several disconnected clusters with high intra-cluster and low inter-cluster similarity.

K-means algorithm is an algorithm based on partition, the algorithm assumes that there is a database consisting of n objects and k is known as the number of clustering. We can make use of the partition method to build k partitions ($k \leq n$). Each partition denotes a cluster. Clustering is also based on the similarity.
between objects. Usually the distance such as Euclidean distance and cosine distance [4].

In Protein-Protein Interaction (PPI) networks, the interaction between two proteins is generally established with a probability property due to the limitation of observation methods [3]. In addition, it has been verified that the interaction between proteins *A* and *B* can influence the interaction between protein *A* and another protein *C*, if *A*, *B* and *C* have some common features. It has been verified that the probability of pair wise interaction and correlation among edges can be derived from statistical models [6]. Clustering applied to such correlated probabilistic protein-protein interaction network data is helpful in finding complexes to analyze the structure properties of the PPI Network.

II.RELATED WORK

A.Graph Theory

A graph G is a pair of sets G = (V, E). V is the set of vertices and the number of vertices n = |V| is the order of the graph. The set E contains the edges of the graph. In an undirected graph, each edge is an unordered pair {v,w}. In a directed graph (also called a digraph in much literature), edges are ordered pairs. The vertices v and w are called the endpoints of the edge. The edge count |E| = m is the size of the graph. In a weighted graph ,a weight function w : E → R is defined that assigns a weight on each edge. A graph is planar if it can be drawn in a plane without any of the edges crossing.

B. Social Network

Social network is a collection of individuals or organizations as well as the links between them, in which each node represents an individual and each link between two nodes denotes their relationship. Social network analysis has emerged as a key technique in many areas, such as biology, economics and etc. A key task of social network analysis is to find community structure, which is quite common in real networks, and being able to identify communities within a network can provide insight into how network function and topology affect each other.

C. Unsupervised Learning

Data modeling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others.

D. Probabilistic Graph Mining

Clustering and partitioning of deterministic graphs has importance in research [6], [7], [8]. These algorithms can be used to handle probabilistic graphs by considering the:

a. Edge probabilities as weights

The disadvantage of this approach is that once probabilities are converts eights, then no other weights can be considered unless the probabilities are multiplied with edge weights. In these cases this constituent weight has no use.

b. By setting a threshold value to the probabilities of the edges and ignoring any edge with probability below this threshold

The disadvantage of the second approach is that there is no rule of deciding what the right value of the threshold is. Since both the above methodologies would result in an algorithm that would output some node clustering would not have specific objective defined over all possible worlds of the input probabilistic graph.

D. Data mining on uncertain data

Data mining of uncertain data have lot of importance. Several classical data-mining problems are there which includes clustering of relational data [10], [11], [12], [13], frequent-pattern mining and evaluating spatial queries and then new idea is proposed.

E. Querying and Mining the Probabilistic Data with Correlations:

Recently, correlations among uncertain data are having more interest. It proposed a framework to represent the correlations among probabilistic tuples. The problem of probabilistic path queries in correlated probabilistic networks is defines and evaluated [13]. They addressed three effective heuristic evaluation functions to in advance estimate the conditional probability of each edge.[4] proposed a method for sub graph similarity search over correlated probabilistic graphs based on possible world semantics. Tight lower and upper bounds of the sub graph similarity probability were developed to prune the search space. Compared to these queries, clustering over correlated probabilistic graphs is more complicated.

III.PROPOSED WORK

A. PEEDR Algorithm

This PEEDR (Partially Expected Edit Distance Reduction) for finding adjusted vertex PEEDR is used[14]. Initialized a cluster with one vertex, then initialized for all vertex in cluster, vertex removed from cluster when it reduce the expected edit distance from graph to current cluster graph. This step is repeated until cluster cannot expand. Then next choose a vertex from the unclustered vertices and repeat this procedure to generate another cluster. Will get final cluster until procedure is not repeated for all vertices but, the problem is which vertex is choose in each repeat step. The solution is find maximum degree vertex which is mostly in centres of cluster, vertices sort in descending order of their degree. Prioritize the vertices with higher degree. Then initialize virtual cluster which keeps all the unclustered vertices. To check each vertex that is adjusted to cluster Distance-Probability-Threshold Clique DPTC is used [9] for which isReduceEdit algorithm is used .Then, pruning by loose bound and pruning by upper bound these techniques are used. Then it is redefined according to joint existence state.

B. DPTC Clustering Algorithm

By correlated probabilistic graph and a cluster number reduce the number of objects by establishing DPTCs (Distance Pint Threshold cliques) first and represent these DPTCs as the objects to be clustered. Second, define the similarity between pair wise adjacent DPTCs to find the K-NN of each DPTC[14]. Third, a Laplacian matrix can be obtained according to the K-NN results, and propose a new approach to calculating the eigenvectors of the Laplacian matrix. Then eigenvectors will be represented in a K dimensional space, and these points are iteratively clustered with a K-means algorithm, such that we get the final cluster graph.

B. Pruning Technique

Pruning, dissimilarity and similarity measure variations lead to many algorithms in literature. We used this point to propose our new algorithm. It becomes easier to compare too.

Similarity measure is a quantification of similarity function like Euclidean distance , distance, cosine. Pruning is thinning the cluster further and thereby improve the prediction accuracy. Pruning identifies the most similar classifiers. A diversity measure indicates predictors disagree with each other. A pair wise distance between two trees can be considered as a measure. K statistics is a measure used in genetic algorithms which complements similarity. Thus a pruning strategy first computes the κ, origin, and pair wise measures and then chooses the predictors to eliminate the dissimilarity.

I borrowed idea of pruning techniques using traditional similarity measure which does not include the node or edge-based similarity parameters.

Formulae for walk starting at C_i first hits one of its adjacent DPTC C_j is defined as

$$sim(C_i,C_j)=\sum_{ep\in S(C_i,C_j)} p^k(X_D(ep))$$

Where, $S(C_i.C_j)$ is the set of edges between C_i and C_j in the correlated probabilistic graph, and $X_D(e_p)$ is the existence state of e_p in the DPTC graph G_D.

The similarity measure considers only the number of edges connected in between the cluster.

Our proposed algorithm considers density ratio (number of vertices connected in every cluster) and we modified the formulae for similarity as

$$new\_sim(C_i,C_j)=\frac{1}{2}\left(\Sigma_{ep\in S_e}(C_i,C_j) + \Sigma_{ep\in S_v}(C_i,C_j)\right)$$

IV .System Architecture

Graph data is collection of a nodes which stores the graph node.Which is input of the system that given as input for probabilistic graph construction where,nodes are formed by using probability. Then grouping of vertices is done according to probability.All the nodes are checked with the similarity and then pass to the probabilistic graph.Which is given as input for PEEDR Algorithm.This algorithm is find the Highest degree order nodes.

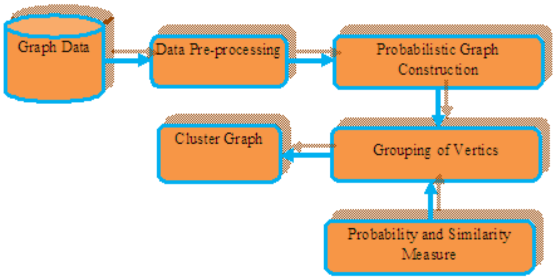
Graph Data
Data Pre-processing
Probabilistic Graph Construction
Cluster Graph
Grouping of Vertics
Probability and Similarity Measure

Fig 4.a System Architecture

Then, DPTC algorithm is performed, which gives the output generated is given to the PEEDR algorithm. a new approach to calculating the eigenvectors of the Laplacian matrix. Then eigenvectors will be represented in a K dimensional space, and these points are iteratively clustered with a K-means algorithm, such that we get the final cluster graph.

V.Result

The effect of the optimizations for PEEDR in terms of running time for that we comparing it with previous method with You-Tube Data Set. PWE is Prununing with Edges And PWED is pruning with Edges and Density ratio.

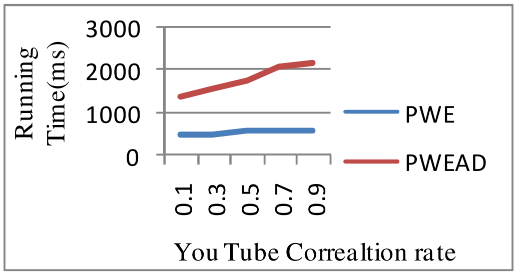

Fig.4.b. PEEDR Efficiency vs Θ:YouTubeCorrelation rate

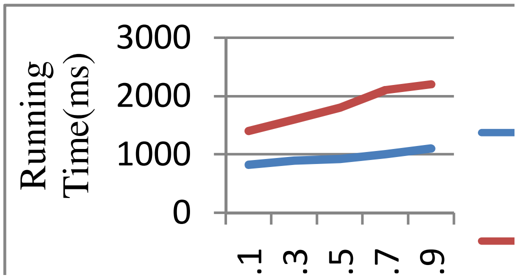

Fig.4.C. PEEDR Efficiency vs Θ:YouTubeCorrelation rate

VI.Conclusion

In this paper we define probabilistic graphs containing correlated adjacent edges as correlated probabilistic graphs which is one of the important and basic technique in data mining. Algorithm used for finding adjusted vertex to cluster PEEDR. To check each vertex that is adjusted to cluster Distance-Probability-Threshold Clique DPTC is used, Pruning techniques introduced with this the efficiency of the PEEDR clustering algorithm improved.

References

[1] “Graph Clustering” Satu Elisa Schaeffer_C Computer Science Review 2007 .

[2] A.K. JAIN, M.N. MURTY, P.J. FLYNN, “Data Clustering: A Review "

[3] C. C. Aggarwal and H. Wang, “Managing and Mining Graph Data”, New York, NY, USA: Springer, 2010.

[4] Ye Yuan, Guoren Wang, Lei Chen, Haixun Wang,"Efficient Subgraph Similarity Search on Large Probabilistic Graph Databases"

[5] Wang, W. and Demsetz" Model for Evaluating Networks under Correlated Uncertainty",—NETCOR.” J. Constr. Eng. Manage., 126(6), 458–466.

[6] U. Brandes, M. Gaertler, and D. Wagner, “Engineering Graph Clustering: Models and Experimental Evaluation,” ACM J. Experimental Algorithmics, vol. 12, article 1.1, pp. 1-26, 2007.

[7] G. Karypis and V. Kumar, “Parallel Multilevel K-Way Partitioning for Irregular Graphs,” SIAM Rev., vol. 41, pp. 278-300, 1999.

[8] M. Newman, “Modularity and Community Structure in Networks,” Proc. Nat’l Academy of Sciences USA, vol. 103, pp. 8577-8582, 2006.

[9] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[10] C.C. Aggarwal and P.S. Yu, “A Framework for Clustering Uncertain Data Streams,” Proc. IEEE 24th Int’l Conf. Data Eng.(ICDE), pp. 150-159, 2008

[11] G. Cormode and A. McGregor, “Approximation Algorithms for Clustering Uncertain Data,” Proc. 27th ACM SIGMOD-SIGACTSIGART Symp.Principles of Database Systems (PODS), pp. 191-200, 2008.

[12] S. Gu¨nnemann, H. Kremer, and T. Seidl, “Subspace Clustering for Uncertain Data,” Proc. SIAM Int’l Conf. Data Mining (SDM),pp. 385-396, 2010

[13] S. Guha and K. Munagala, “Exceeding Expectations and Clustering Uncertain Data,” Proc. 28th ACM SIGMOD-SIGACT-SIGARTSymp. Principles of Database Systems (PODS), pp. 269-278, 2009.

[14] Yu Gu,Chunpeng Gao, Gao Cong, and Ge Yu, "Effective and Efficicient Clustering Methods for correlated probabulistic graph" IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 5, May 2014.