International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

Survey of Correlated Probabilistic Graph

Sawant Ashlesha G.[1], Gadekar Devendra P[2]

[1]Pune University, Pune, Maharashtra, India
[2]Assistant Professor, Pune University, Pune, Maharashtra, India

Abstract: ***Now days probabilistic graph have more interest in the data mining community. After observation it is find that correlations may exist among adjacent edges in various probabilistic graphs. As one of the basic mining techniques, graph clustering is widely used in data analysis where a problem that has not been clearly defined, such as data compression, information retrieval, image segmentation, etc. Graph clustering is used to divide data into clusters according to their similarities, and a number of algorithms have been proposed for clustering graphs, such as the pKwik Cluster algorithm, spectral clustering, k-path clustering, etc. In this way, little research has been performed to develop efficient clustering algorithms for probabilistic graphs. But, it becomes more challenging to efficiently cluster probabilistic graphs when correlations are considered. In this paper, we define the problem of clustering correlated probabilistic graphs and its techniques which are used before and its problem. To solve the challenging problem two algorithms, namely the PEEDR and the CPGS clustering algorithm are defined for each of the proposed algorithms, and then also define some several pruning techniques to further improve their efficiency.***

Keywords: Clustering, Graph Mining, Correlated, Probabilistic Graph, Spectral Clustering

1. Introduction

In recent years, Graph Mining has achieved a lot of importance. Graph is a diagram showing the relation between variable quantities and graph mining is a special case of structured data mining where Structure mining is the growth of the use of semi-structured data which created new opportunities for data mining, which has traditionally been concerned with tabular data sets, reflecting the strong association between data mining and relational databases .It is the process of finding and extracting useful information from semi structured data sets. Graphs become increasingly important in modeling complicated structures, such as circuits, images, chemical compounds, protein structures, biological networks, social networks, the Web, workflows, and XML document. It has various applications such as social network, protein-protein interaction networks etc. As Social network is consisting of nodes and link, nodes are used as people and link are used as communication. A method for determining the clustering structure with the Eigen-structure of the linkage matrix is determine the community structure which is proposed in managing and mining graph. Large network is managed by sub graphs. Which is important that handle nature of sub graph for large graph network .Communication is captured in terms of graph and such a application is very challenging. Therefore for their purpose of structural analysis all data cannot be localized on disk therefore new techniques need to summarize. This data displays an inherent property of uncertainty and they modeled as probabilistic graph. Similar to the problem of similarity search in standard graphs, a fundamental problem for probabilistic graphs is to efficiently answer k-nearest neighbor queries (k-NN), which is the problem of computing the k closest nodes to some specific node that extend well-known graph concepts, such as shortest paths for that sampling based algorithm is used [12]. Querying and mining uncertain graphs has become increasingly important nowadays. The distance-constraint reach ability (DCR) problem is given two vertices what is the probability that the distance from two vertices is less than or equal to a user-defined threshold in the uncertain graph. Since this problem is #P-Complete [3].

Similarly, $e1$ and $e2$ are also conditionally dependent on each other due to a coexistence constraint. In this case if correlations are ignored then it gives incorrect result. According to many scenarios, the correlations among edges not consider mutex or coexistence and more complicated dependency exists. In order to model such correlation joint probability table having joint probability among adjacent edges.

This paper defines probabilistic graphs containing correlated adjacent edges as correlated probabilistic graphs. As one of the important and basic technique of data mining clustering is used, for various graph analysis applications [1] . Clustering is the unsupervised classification of observations, data items, or feature vectors into groups. It is important to understand the difference between clustering i.e. Unsupervised classification and discriminate analysis i.e. supervised classification. Supervised classification is a collection of labeled patterns. Clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. Labels are associated with clusters also, but these category labels are data driven that is, they are obtained anything from data. Such as community detection, index construction, etc. This paper projected on clustering correlated probabilistic graphs. Which includes partitioning the vertices into several disconnected clusters with high intra-cluster and low inter-cluster also motivate the problem of clustering correlated probabilistic graphs using several applications. In Protein-Protein Interaction (PPI) networks, Due to limitation of observation methods, the interaction between two proteins is generally existed. Probability of pair wise interaction and correlation between edges can be derived from statistical model. Where if main nodes are divided into sub nodes then that nodes also fragmented to another sub node; in this case correlation is captured by sampling from the same-condition as child node is gives iteration of simulation.[5] In social network there is correlation for the link. To detect effective user communities it is necessary to consider the potential

probabilities and correlation. As compare to clustering probabilistic graphs, clustering correlated probabilistic graph has more rules.

2. Background

To make cluster of a correlated probabilistic graph a possible world graph can be mode led as a deterministic from the correlated probabilistic graph which considers the joint probability distribution. The edit distance to the cluster graphic defined as the number of edges that need to be added or removed to transform graph into new graph. By evaluating all the possible world graph instances, the expected edit distance can be obtained and viewed as a measurement for getting value of the deviation from a correlated probabilistic graph to the cluster graph. Therefore, a smaller deviation indicate a more accurate result, and this paper objective turns to the goal of finding a cluster graph that can minimize a new edit distance cluster. As, it is more time-consuming if we calculate the expected edit distance by considering all possible world graphs. To re solve this problem, this paper proposes a novel estimation model which has the dynamical generation of an edge access order when calculating conditional probabilities. The estimation model has errors.

2.1 Graph and 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, either by considering the edge probabilities as weights, or by setting a threshold value to the probabilities of the edges and ignoring any edge with probability below this threshold.

- The disadvantage of the first 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.
- 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.

Hence, various graph mining problems have been studied recently assuming uncertain graphs [9], [10], [11], Potamias[12] proposed new robust distance functions between nodes in probabilistic graphs that extend shortest path distances from deterministic graphs and proposed methods to compute them efficiently. The problem of finding shortest paths in probabilistic graphs based on transportation networks has also been considered [13], [14]. The intersection between the methods and all of them regarding probabilistic graphs .But, the graph-clustering task under the possible world's semantics has not been addressed by researchers in probabilistic graph mining.

2.2 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 [19], [20], [21], [22], [23], [24], [25], frequent-pattern mining [26], [27], [28] and evaluating spatial queries [29].And then new idea is proposed [30].

It may be trying and use the same definitions for probabilistic graphs, particularly since standard clustering objectives (e.g., k-center or k-median) can be optimized in deterministic graphs. As, there is a difficulty with such clustering definitions in the probabilistic-graph setting since there are many worlds where parts of the graph are disconnected, the distance of a node to any of the existing clustering centers can be infinity And then, for nontrivial probabilistic graphs, there is always a nonzero probability of having a node within finite distance to all the cluster centers. In that case, the optimization function becomes infinity. Therefore, new definitions of the clustering problem in probabilistic graphs proposed [31].

2.3 Probabilistic Database:

Probabilistic databases is another active research area, mostly includes methods for storing, managing, and querying probabilistic data [32].A probabilistic database management system, is a system that stores large volumes of probabilistic data and supports complex queries. It may also need to perform some additional tasks, such as updates or recovery, but these do not differ from those in conventional database management systems.. The challenge in this database is that it needs both to scale to large data volumes, a core competence of database management systems, and to do probabilistic inference. While many scalable data management systems exists, probabilistic inference is in general a hard problem , and current systems do not scale to the same extent as data management systems do. To address this challenge, researchers have focused on the specific nature of relational probabilistic data, and exploited the special form of probabilistic inference that, there exists fundamental work on the complexity of query evaluation on such data.

2.4 Probabilistic-Graph Model:

Similar to deterministic graphs, probabilistic graphs may be undirected or directed and carry additional labels on the edges such as weights [32] model assumes independence among edges it focuses on independent probabilistic graph. It represents a probabilistic graph using tuple. Unweighted probabilistic graphs represent the probabilistic Graph. One can think of a probabilistic graph as a generative model for deterministic graphs. A deterministic graph is generated by connecting two nodes via an edge with probability. Deterministic graphs are an instance of probabilistic graphs for which random graphs are an instance of probabilistic graphs where all edge probabilities are the same and equal. Then there are distinct graphs that can be generated .They use the term possible world to refer to each such graph.

2.5 Deterministic Cluster Graphs:

In this section, formulate the problem of clustering in probabilistic graphs as an optimization problem. First, define the edit distance between two graphs. Then, generalize definition for probabilistic graphs and use it to define graph-clustering problem. The edit distance is the number of edges that need to be added or deleted from graph in order to be transformed. The edit distance between a probabilistic graph and a deterministic graph is defined as the expected edit distance between every graph. The expected edit distance between a probabilistic graph and a deterministic graph can be computed in polynomial time. Using this observation and linearity of expectation is gained. As unweight graph to cluster. A central notion analysis is the cluster graph. A cluster graph is a special deterministic graph that consists of vertex-disjoint disconnected cliques. A probabilistic graph finds the cluster graph. The number of output clusters is not part of the input. In fact, the objective function itself dictates the number of clusters that are appropriate for every input. The input graph is deterministic called CLUSTEREDIT problem.

3. Related Work

3.1 Algorithms for Clustering Deterministic Graphs:

In data mining research and no. of clustering algorithm deterministic graph clustering has been briefly studied. Survey of graph clustering method is provided in[1]. They discussed the different categories of clustering algorithms and recent efforts to design clustering methods for various kinds of graph data. Clustering techniques and some important applications of clustering algorithms is discussed in [3]. As one of the most widely used graph clustering algorithms, spectral clustering is issue of researchers. Spectral clustering depend on the eigen structure of a graph Laplacian matrix to partition vertices into disjoint clusters, with points in the same cluster having high similarity and points in different clusters having low similarity[15]. The rationality of the spectral clustering method was analyzed [16]. For spectral clustering they derived new cost functions which are based on measures of error between a given partition and a solution of the spectral relaxation. Then a number of optimizations for spectral clustering were proposed in [17][18].Spectral clustering refers to a flexible class of clustering procedures that can produce high-quality clustering on small data sets but which has limited applicability to large-scale problems due to its computational complexity. Furthermore, most of the existing algorithms are applied in clustering deterministic graphs. Particularly, as correlations exist among edges, it is inappropriate to directly apply these algorithms to clustering correlated probabilistic graphs.

3.2 Querying and Mining of Probabilistic Graphs:

Now a days, querying and mining of probabilistic graphs have more attraction part of researchers. Many classical data mining problems have been redefined in probabilistic graphs, such as the reach ability query, shortest path query, *K*-NN query, etc. The Distance-constraint reach ability query and presented sampling algorithm to answer the NP-hard problem [12] introduced an efficient algorithm for *K*NN queries in probabilistic graphs based on the random walk method. As an important preliminary work, [12] advanced the state of the art by exploring the problem of clustering probabilistic graphs. They proposed efficient algorithms to find a cluster graph, such as the pKwik-Cluster algorithm, the furthest algorithm, etc. These algorithms do not consider the correlations among edges, and thus are not applicable for clustering correlated probabilistic graphs.

3.3 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 .An efficient strategy was developed for query evaluation over such probabilistic databases by selecting the query processing problem as inference problem in a properly constructed probabilistic graphical model. Then, the nearest neighbor query on uncertain data with local correlations is investigated. After that, a novel filtering technique by offline pre-computations was developed to reduce the query search space. There also exist studies on evaluating correlated probabilistic graphs. 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.

3.4 PEEDER Algorithm:

This algorithm is used for finding adjusted vertex to cluster. After initialization of a cluster with one vertex, initialized for all vertexes in cluster, vertex removed from cluster when it reduces 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 uncluster vertices and repeat this procedure to generate another cluster. This procedure is repeated until all vertices of first graph are grouped into cluster and then will get final cluster. One problem is in this clustering is which vertex is choose in each repeat step. The solution find on this maximum degree vertex is mostly in centers 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, 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.

3.5 CPGS Clustering Algorithm:

CPGS Correlated Probabilistic Graphs Spectral to cluster correlated probabilistic graphs. By correlated probabilistic

graph and a cluster number reduce the number of objects by establishing DPTCs 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. 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.

4. 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. Clustering is used for various graph analysis applications. 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, for which isReduceEdit algorithm is used .Pruning techniques introduced with this the efficiency of the *PEEDR* clustering algorithm improved. To get better effectiveness of clustering, we also addressed another clustering algorithm CPGS.

References

[1] A.K. JAIN, M.N. MURTY, P.J. FLYNN, "Data Clustering: A Review"
[2] C. C. Aggarwal and H. Wang, Managing and Mining Graph Data, New York, NY, USA: Springer, 2010.
[3] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," ACM Comput. Surv, vol. 31, no. 3, pp. 264–323, Sept. 1999.
[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. Emek, A. Korman, and Y. Shavitt, "Approximating the Statistics of Various Properties in Randomly Weighted Graphs," Proc. 22ndAnn. ACM-SIAM Symp. Discrete Algorithms (SODA), pp. 1455-1467. 2011,
[10] P. Hintsanen and H. Toivonen, "Finding Reliable Subgraphs from Large Probabilistic Graphs," Data Mining Knowledge Discovery, vol. 17, no. 1, pp. 3-23, 2008.
[11] X. Lian and L. Chen, "Efficient Query Answering in Probabilistic Rdf Graphs," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD '11), pp. 157-168, 2011.
[12] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios, "K-Nearest Neighbors in Uncertain Graphs," Proc. VLDB Endowment, vol. 3, nos. 1/2, pp. 997-1008, 2010.
[13] M. Hua and J. Pei, "Probabilistic Path Queries in Road Networks: Traffic Uncertainty Aware Path Selection," Proc. 13th Int'l Conf.Extending Database Technology (EDBT), pp. 347-358, 2010.
[14] Y. Yuan, L. Chen, and G. Wang, "Efficiently Answering Probability Threshold-Based Shortest Path Queries over UncertainGraphs," Proc. 15th Int'l Conf. Database Systems for Advanced Applications (DASFAA), pp. 155-170, 2010.
[15] U. von Luxburg, "A tutorial on spectral clustering, Statist. Comput" vol. 17, no. 4, pp. 395–416, Dec. 2007.[16]
[16] F. R. Bach and M. I. Jordan, "Learning spectral clustering, with application to speech separation," J. Mach. Learn. Res., vol. 7, pp. 1963–2001, Oct. 2006.
[17] D. Yan, L. Huang, and M. I. Jordan, "Fast approximate spectral clustering," in Proc. 15th KDD, Paris, France, 2009, pp. 907–916.
[18] R. Kannan, S. Vempala, and A. Vetta, "On clusterings: Good, bad and spectral," J. ACM, vol. 51, no. 3, pp. 497–515, 2004.
[19] 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.
[20] 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.
[21] 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.
[22] 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.
[23] B. Kao, S.D. Lee, F.K.F. Lee, D.W.-L.Cheung, and W.-S. Ho, "Clustering Uncertain Data Using Voronoi Diagrams and R-TreeIndex," IEEE Trans. Knowledge Data Eng., vol. 22, no. 9, pp. 1219- 1233, Sept. 2010.
[24] H.-P. Kriegel and M. Pfeifle, "Density-Based Clustering of Uncertain Data," Proc. 11th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), pp. 672-677, 2005.
[25] W.K. Ngai, B. Kao, C.K. Chui, R. Cheng, M. Chau, and K.Y. Yip,"Efficient Clustering of Uncertain Data," Proc. Sixth Int'l Conf. Data Mining (ICDM), pp. 436-445, 2006.
[26] T. Bernecker, H.-P. Kriegel, M. Renz, F. Verhein, and A. Zuefle, "Probabilistic Frequent Itemset Mining in ncertain databases,"Proc. 15th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), 2009.
[27] L. Sun, R. Cheng, D.W. Cheung, and J. Cheng, "Mining Uncertain Data with Probabilistic Guarantees," Proc. 16th ACM SIGKDD Int'l Conf. Knowledge Discovery in Data Mining (KDD), pp. 273-282, 2010.

[28] Q. Zhang, F. Li, and K. Yi, "Finding Frequent Items in Probabilistic Data," Proc. ACM SIGMOD Int'l Conf. Management of Data (SIGMOD), pp. 819-832, 2008.
[29] M.L. Yiu, N. Mamoulis, X. Dai, Y. Tao, and M. Vaitis, "Efficient Evaluation of Probabilistic Advanced Spatial Queries on Existentially Uncertain Data," IEEE Trans. Knowledge and Data Eng.,vol. 21, no. 1, pp. 108-122, Jan. 2009.
[30] 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
[31] G. Kollios, M. Potamias, and E. Terzi, "Clustering large probabilistic graphs," IEEE Trans. Knowl. Data Eng., vol. 25, no. 2, pp. 325–336, Feb. 2013.
[32] N.N. Dalvi, C. Re´, and D. Suciu, "Probabilistic Databases: Diamonds in the Dirt," Comm. ACM, vol. 52, no. 7, pp. 86-94, 2009.
[33] Ruoming Jin ,Lin Liu, Bolin Ding, Haixun Wang, Distance-Constraint Reachability Computation in Uncertain Graphs