Information Sciences 317 (2015) 78–95

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Clustering large probabilistic graphs using multi-population evolutionary algorithm

Zahid Halim *, Muhammad Waqas, Syed Fawad Hussain

Faculty of Computer Science and Engineering, Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan

ARTICLE INFO

Article history:
Received 6 October 2014
Received in revised form 7 April 2015
Accepted 18 April 2015
Available online 24 April 2015

Keywords:
Probabilistic graphs
Clustering
Multi-population evolutionary algorithm
Graph mining

ABSTRACT

Determining valid clustering is an important research problem. This problem becomes complex if the underlying data has inherent uncertainties. The work presented in this paper deals with clustering large probabilistic graphs using multi-population evolutionary algorithm. The evolutionary algorithm (EA) initializes its multiple populations, each representing a deterministic version of the same probabilistic graph given to it as an input. Multiple deterministic versions of the same input graph are generated by applying different thresholds to the edges. Each chromosome of the multiple populations represents one complete clustering solution. For the purpose of clustering, EA is employed which is guided by pKwikCluster algorithm. The proposed approach is tested on two natively probabilistic graphs and nine synthetically converted probabilistic graphs using cluster validity indices of Davies–Bouldin index, Dunn index, and Silhouette coefficient. The proposed approach is also compared with two baseline clustering algorithms for uncertain data, Fuzzy-DBSCAN and uncertain K-mean and two state-of-the-art approaches for clustering probabilistic graphs. The results obtained suggest that the proposed solution gives better performance than the baseline methods and the state-of-the-art algorithms.

1. Introduction

There are many real world problems that are cluttered with inherent uncertainties. These uncertainties may be attributed to multiple reasons, including: equipment's limitations to observe or record data, limited resources to collect and store data or due to incorrect data entry. Analyzing such datasets and extracting useful information therefrom is mostly influenced by the underlying uncertainty. This problem is prevalent in domains such as social networks, biological networks, and mobile ad hoc networks [31] to name a few. However, most of these datasets can be modeled as a graph converting the problem of uncertain data into a probabilistic graph. Later, the graph problem can be solved for uncertain data. Graphs can further be categorized into two types: deterministic graphs and probabilistic graphs. Probabilistic graphs have a probability value associated with each edge. Deterministic graphs on the other hand have only two possibilities regarding the edge; either there will be an edge (probability of having an edge is, 1) or there will be no edge (probability of having an edge is 0). For uncertain datasets, a probabilistic graph representation is more suitable. Once a dataset is represented as a probabilistic graph an algorithm for mining the probabilistic graph can be developed.

* Corresponding author.
E-mail address: zahid.halim@giki.edu.pk (Z. Halim).

http://dx.doi.org/10.1016/j.ins.2015.04.043
0020-0255/

Probabilistic graphs (or uncertain graphs) are an important area of research for their numerous applications in an assortment of fields [17]. A few cases in point are a protein–protein interaction network [51], wireless sensor network with uncertain edges [36], and social networking. Clustering of such graphs has recently gained popularity [2]. One way to cluster probabilistic graphs is to convert it into a deterministic graph by setting a threshold value for edge weights first and then apply a clustering algorithm on the resulting deterministic graph. For instance, if 0.5 is selected as a threshold value, only the edges having probability greater than or equal to 0.5 will be considered and the rest will be ignored. The smaller the threshold, the larger the number of the edges that will remain in the graph. Obviously, there will be a loss of information in the form of some edges and/or nodes. Since connectivity is one of the important properties of a graph, approaches are needed to select a suitable threshold value that shall minimize the loss of information while transforming a probabilistic graph into a deterministic one.

The work presented here focuses on solving the problem of clustering large probabilistic graphs using an evolutionary algorithm (EA). The proposed solution is based on a multi-population genetic algorithm (GA). The input to GA starts with initializing its multiple populations, each representing a deterministic version of the same probabilistic graph. Multiple deterministic versions of the input graphs are generated by applying different thresholds to the edges of the input graph. Each chromosome of the multiple populations represents one complete clustering solution. For clustering, an EA is employed; however, the EA is guided by the pKwikCluster algorithm [28]. The GA runs for a fixed number of iterations and after examining each of the clusters produced, most appropriate threshold is selected and hence the most appropriate clusters. The proposed approach is tested using two probabilistic graph datasets and also using nine deterministic graphs by adding noise in order to make them probabilistic. It is also compared with two baseline clustering algorithms for uncertain data, FDBSCAN (Fuzzy-DBSCAN) [48] and UK-Mean (uncertain K-mean) [9], and two recent approaches [21,22,27] for clustering probabilistic graphs. The results suggest that the proposed algorithm gives better results than the baseline and other recent approaches.

1.1. Contribution

The key contributions of the work are described in this section. The given probabilistic dataset is transformed into a probabilistic graph and evolutionary algorithm (EA) is used for clustering. The graph based representation of the problem makes it convenient to extract neighborhood information of the nodes and uses the same to process probabilities on the edges between nodes. The EAs are inherently slow. The proposed solution uses a multi-population EA that converges at multiple solutions simultaneously. It incorporates both the local and global search capability. The multiple populations of the EA help the algorithm to find multiple local optimum clusters. During the decision phase, the algorithm selects the best option from the converged solutions of the multiple populations. This enables it to select the global best clusters. Since probabilistic graphs are being dealt with, a strategy that converts a given probabilistic graph into multiple deterministic graphs is devised. These graphs are later clustered using multi-population EAs. The work also contributes towards selecting the most appropriate clustering formation based on individual cluster properties from the pool of solutions that EA provides after convergence. The proposed approach produces better results when compared with two baseline methods and two recent algorithms using well-known cluster validity indices.

The rest of the paper is organized as follows: Section 2 presents the related work and Section 3 explains the probabilistic graphs and defines related key terms. Section 4 explains the proposed methodology covering chromosome structure, reproduction operators, and the fitness functions. Section 5 lists the detailed experiments and the results obtained including comparison with the baseline and recent methods. Finally, Section 6 presents conclusion along with some future directions.

2. Related work

The study of clustering methods is one of the major machine learning and data mining area and has been applied to an assortment of applications ranging from biomedical to physics experiments. Graph clustering makes it convenient to cluster data where geometric coordinates of the items to be clustered are not available. Clustering of graphs can be categorized based on the type of a particular graph. This section lists the recent work on clustering noisy graphs, deterministic graphs, probabilistic graphs, and graph clustering using EAs.

In case of probabilistic graphs there is only a marginal contribution to the problem of clustering. However, much work is present in clustering other type of graphs [10,45,16]. Chen et al. [13] introduced an algorithm for clustering sparse, unweighted graphs having noise. Noise is defined in [13] as the edge density between different clusters and convex optimization formulation is used for graph clustering. Priyadarshini et al. [42] have shown a parameter graph based clustering technique (GCEPD) to identify coherent patterns from gene expression data. Work in [42] discovered highly coherent patterns containing genes with high biological relevance. Clémençon et al. [11] presented hierarchical clustering for graph visualization. Mishra et al. [41] proposed a technique for clustering social networks and illustrated the effectiveness of the algorithm through experiments on real social networks. Satuluri and Parthasarathy [44] proposed an approach for clustering directed graphs. Satuluri and Parthasarathy [44] first converted the directed graph into undirected graph and then applied a clustering algorithm. They also introduced a degree-discounted similarity measure which was suitable for large-scale networks. Balakrishnan et al. [6] have analyzed the performance of a fault tolerant spectral algorithm for hierarchical clustering.