International Journal of Multimedia Data Engineering and Management
Volume 12 • Issue 1 • January-March 2021

A Comparative Study of Graph Kernels and Clustering Algorithms

Riju Bhattacharya, National Institute of Technology, Raipur, India
Naresh Kumar Nagwani, National Institute of Technology, Raipur, India
Sarsij Tripathi, Motilal Nehru National Institute of Technology, Allahabad, India

ABSTRACT

Graph kernels have evolved as a promising and popular method for graph clustering over the last decade. In this work, comparative study on the five standard graph kernel techniques for graph clustering has been performed. The graph kernels, namely vertex histogram kernel, shortest path kernel, graphlet kernel, k-step random walk kernel, and Weisfeiler-Lehman kernel have been compared for graph clustering. The clustering methods considered for the kernel comparison are hierarchical, k-means, model-based, fuzzy-based, and self-organizing map clustering techniques. The comparative study of kernel methods over the clustering techniques is performed on MUTAG benchmark dataset. Clustering performance is assessed with internal validation performance parameters such as connectivity, Dunn, and the silhouette index. Finally, the comparative analysis is done to facilitate researchers for selecting the appropriate kernel method for effective graph clustering. The proposed methodology elicits k-step random walk and shortest path kernel have performed best among all graph clustering approaches.

KEYWORDS

Graph Kernel, Internal Validation Index Parameters, Unsupervised Graph Clustering

1. INTRODUCTION

Clustering is an important component in the unsupervised learning process and plays a vital role in better understanding of graphs. Over the year, graph clustering has proven to be a promising method of processing graph-based data in cluster analysis (Schaeffer, 2007). The purpose of graph clustering is to form a cluster of graphs such that the graph of the same cluster is highly correlated with similar attributes but differs significantly from the other cluster members. The intensity of similarity between graph pairs, known as graph similarity, is performed through graph analysis (Zager & Verghese, 2008). In graph analysis, a graph comparison is performed to determine cluster similarity/dissimilarity among the clusters.

There are two different approaches to graph comparison. The first method is a within-graph clustering technique comparing the graph by finding a mapping between vertices in a single graph. Secondly, a between-graph clustering method divides a set of graphs into clusters based on their structural similarity property, such as degree distribution (Kriege et al., 2020). Graph comparison incorporates various standard similarity techniques for graph comparisons that are based on graph isomorphism, topological descriptor, and inexact matching algorithm (Ghosh et al., 2018). These techniques have exponential runtime complexity for larger graph-based data. Due to this problem,

DOI: 10.4018/IJMDEM.2021010103

Graph kernels have been recommended as an alternative efficient and powerful method for measuring the similarity of graph-structured data with polynomial runtime complexity (Nikolentzos et al., 2019). A graph kernel is an asymmetrical, positive semi-definite function that is demarcated on a graph $\mathcal{G}$ space (Vishwanathan et al., 2010). This can be easily used for structural objects by implementing standard extended clustering algorithms such as partitioned, hierarchical, model-based, etc (Datta & Datta, 2003), (Ghosh et al., 2018).

Previous studies have shown that various graph kernel approaches have been used to measure the graph similarity and different classification strategies have been used to classify the graphs respectively, but to the best of the author's knowledge graph kernel is not applied together with clustering techniques. This urges a need for a methodology to incorporate various kernel methods for establishing similarity measurement of multi-graph data. The aim of this work is to provide a comparative framework for determining the best kernel method for graph clustering. For this, a three-step methodology is proposed. In the first step, the similarity measurement of the graph has been performed by applying five different kernel methods. In the second step, the similarity matrix obtained by each kernel method is processed by different clustering methods to form different graph clusters. In the last step, the graph clusters obtained are examined through connectivity, Dunn, and silhouette internal validation index parameters to determine the best performing graph kernel for clustering. The results of the proposed framework help researcher and academicians in determining the best graph kernel and clustering methods for obtaining the desired results in various domains such as social networks, chemoinformatics, and bioinformatics.

The organization of the work is as follows. In section 2, a comprehensive literature survey has been presented on graph kernels and clustering algorithms. Section 3, briefly introduces the graphs and different graph kernel methods used in the current work. In section 4, a discussion about various graph clustering techniques, in brief, has been presented. The methodology of the proposed comparative study has been discussed in section 5. In Section 6, the experimental findings are reported and the discussion on the most effective graph kernel methods for graph-based clustering has been laid down. In section 7, practical implications of graph kernel methods in various graph mining problems have been presented. Finally, the conclusion has been presented in section 8.

2. RELATED WORKS

This section presents a detailed literature survey on graph kernels and clustering algorithms.

2.1 Graph Kernel

Over the past decade most research in graph analysis has emphasized the use of graph kernels. Graph kernel is one of the latest approaches to graph comparison over the traditional comparison methods. Graph kernels can be broadly classified into two major categories based on their primary driving force: model-based and syntax-based kernels (Ghosh et al., 2018).

In the first category, Model driven kernels are based on certain information about the sample space, i.e. the relationships between data. This approach primarily has two subcategories, generative models and transformative models. In computer science research, generative kernel models such as the hidden Markov model (Eddy, 1996) are frequently used. The generative model is used to apply kernel-based techniques to sequential data, while the information about the instance properties and possible transformations between instances (graphs) is the basis of transformative model kernels. The diffusion kernel is the most prominent kernel in transformative model kernels class. The fundamental principle behind diffusion kernels (Kondor & Lafferty, 2002) is that the locality of an instance can be more easily represented rather than the whole structure of the instance space represented.

In the second category, the Syntax-based kernel model comprises a wide range of kernel methods in the R-Convolution kernel framework (Haussler, 1999). Kernel methods are used to measure the similarity of two objects. In order to find the similarity, it must satisfy the two mathematical properties.