Abstract

Explainable graphlet distance-based whole-graph embedding

Enikő Zakar-Polyák

2023

Complex systems with inherent connections, which can be represented by graphs, appear in every field of life. People create extensive relationship networks, build various connections between settlements that form complex road networks, but nature also builds its own systems of connections, like food chains. The analysis and creation of such systems is essential to understanding and developing our world. In addition to traditional graph and network theoretical approaches, numerous graph-based problems could benefit from the use of machine learning algorithms as well. In order to utilise such powerful tools on graphs, a vector representation, i.e. an embedding into a Euclidean space of graphs is necessary.

Numerous graph embedding approaches have been introduced throughout the years, but many of them can be hardly interpreted, which is essential to properly understand the solution of a problem. Therefore, the aim of this work is to introduce a novel explainable embedding approach, which is based on the widely-known graphlet-based characterisation of graphs, but intends to give a more accurate representation by taking into consideration relative positions as well. The relative positions are defined with graphlet and instance centroids, and both the distance of different graphlets, and the characterisation of each graphlet based on the location of its instances are included in the vector representation of the graph. The Graphlet Distance-based embedding algorithm (GraD) is a general, task-agnostic approach, it gives graph representations that can be used for arbitrary purposes. It is also explainable since we know exactly what the components of the embeddings represent.

To see the performance of the introduced GraD embedding algorithm, we carried out a comparative analysis on a large number of graph classification benchmark tasks, with various embedding approaches. We found that GraD is always among the top-performing methods. Moreover, the fact that GraD is at least as efficient as the other considered graphlet-based approaches on each dataset, but mostly achieves better results, suggests that taking into consideration the relative positions and higher order graphlets result in a more accurate characterisation of graphs. By exploring the features that highly influence the prediction of the classifiers for some tasks, we also demonstrated the explainability of the embeddings. These analyses also showed that the relative position-related features can have significant predictive power.