Classification of Graph Structures

Andrzej Dominik
Warsaw University of Technology, Poland

Zbigniew Walczak
Warsaw University of Technology, Poland

Jacek Wojciechowski
Warsaw University of Technology, Poland

INTRODUCTION

Classification is a classical and fundamental data mining (machine learning) task in which individual items (objects) are divided into groups (classes) based on their features (attributes). Classification problems have been deeply researched as they have a large variety of applications. They appear in different fields of science and industry and may be solved using different algorithms and techniques: e.g. neural networks, rough sets, fuzzy sets, decision trees, etc. These methods operate on various data representations. The most popular one is information system/decision table (e.g. Dominik, & Walczak, 2006) denoted by a table where rows represent objects, columns represent attributes and every cell holds a value of the given attribute for a particular object. Sometimes it is either very difficult and/or impractical to model a real life object (e.g. road map) or phenomenon (e.g. protein interactions) by a row in decision table (vector of features). In such a cases more complex data representations are required e.g. graphs, networks. A graph is basically a set of nodes (vertices) connected by either directed or undirected edges (links). Graphs are used to model and solve a wide variety of problems including classification. Recently a huge interest in the area of graph mining can be observed (e.g. Cook, & Holder, 2006). This field of science concentrates on investigating and discovering relevant information from data represented by graphs.

In this chapter, we present basic concepts, problems and methods connected with graph structures classification. We evaluate performance of the most popular and effective classifiers on two kinds of classification problems from different fields of science: computational chemistry, chemical informatics (chemical compounds classification) and information science (web documents classification).

BACKGROUND

There are numerous types of pattern that can be used to build classifiers. Three of these are frequent, common and contrast patterns.

Mining patterns in graph dataset which fulfill given conditions is a much more challenging task than mining patterns in decision tables (relational databases). The most computationally complex tasks are subgraph isomorphism (determining if one smaller graph is included in other larger graph) and isomorphism (testing whether any two graphs are isomorphic (really the same). The former is proved to be NP-complete while the complexity of the latter one is still not known. All the algorithms for solving the isomorphism problem present in the literature have an exponential time complexity in the worst case, but the existence of a polynomial solution has not yet been disproved. A universal exhaustive algorithm for both of these problems was proposed by Ullman (1976). It operates on the matrix representation of graphs and tries to find a proper permutation of nodes. The search space can be greatly reduced by using nodes invariants and iterative partitioning. Moreover multiple graph isomorphism problem (for a set of graphs determine which of them are isomorphic) can be efficiently solved with canonical labelling (Fortin, 1996). Canonical label is a unique representation (code) of a graph such that two isomorphic graphs have the same canonical label.

Another important issue is generating all non-isomorphic subgraphs of a given graph. The algorithm for generating DFS (Depth First Search) code can used to enumerate all subgraphs and reduce the number of required isomorphism checking. What is more it can be improved by introducing canonical labelling.

Contrast patterns are substructures that appear in one class of objects and do not appear in other classes

whereas common patterns appear in more than one class of objects. In data mining, patterns which uniquely identify certain class of objects are called jumping emerging patterns (JEP). Patterns common for different classes are called emerging patterns (EP). Concepts of jumping emerging patterns and emerging patterns have been deeply researched as a tool for classification purposes in databases (Kotagiri, & Bailey, 2003). They are reported to provide high classification accuracy results. Frequent pattern is a pattern which appears in samples of a given dataset more frequently than specified threshold. Agarwal and Srikant proposed an efficient algorithm for mining frequent itemsets in the transaction database called Apriori.

In graph mining contrast graph is a graph that is subgraph isomorphic to at least one graph from particular class of graphs and is not subgraph isomorphic to any graph from any other class of graphs. The concept of contrast graphs was studied by Ting and Bailey (2006). They proposed an algorithm (containing backtracking tree and hypergraph traversal algorithm) for mining all disconnected contrast graphs from dataset. Common graph is subgraph isomorphic to at least one graph in at least two different classes of graphs while frequent graph is subgraph isomorphic to at least as many graphs in a particular class of graphs as specified threshold (minimal support of a graph). Kuramochi and Karypis (2001) proposed an efficient (using canonical labelling and iterative partitioning) Apriori based algorithm for mining frequent graphs.

MAIN FOCUS

One of the most popular approaches for graph classification is based on SVM (Support Vector Machines). SVMs have good generalization properties (both theoretically and experimentally) and they operate well in high-dimensional datasets. Numerous different kernels were designed for this method (Swamidass, Chen, Bruand, Phung, Ralaivola & Baldi, 2005).

Another approach is based on k-NN (k-Nearest Neighbors) method. The most popular similarity measures for this method use concept of MCS (maximum common subgraph) (Markov, Last, & Kandel, 2006).

Deshpande, Kuramochi and Karypis (2003) proposed classification algorithms that decouples graph discovery process from the classification model construction. These methods use frequent topological and geometric graphs.

Recently a new algorithm called the CCPC (Contrast Common Patterns Classifier) was proposed by Dominik, Walczak, & Wojciechowski (2007). This algorithm uses concepts that were originally developed and introduced for data mining (jumping emerging patterns - JEP and emerging patterns - EP). The CCPC approach uses minimal (with respect to size and inclusion (non-isomorphic)) contrast and common connected graphs. Classification process is performed by aggregating supports (or other measure based on support) of contrast graphs. Common graphs play marginal role in classification and are only used for breaking ties.

Applications: Chemical Compounds Classification

Chemical molecules have various representations depending on their dimensions and features. Basic representations are: 1-dimensional strings expressed in SMILES language (language that unambiguously describe the structure of chemical molecules using short ASCII strings), 2-dimensional topological graphs and 3-dimensional geometrical structures containing coordinates of atoms. We are particularly interested in 2-dimensional graph representation in which atoms correspond to vertices and bonds to edges. Nodes are labelled with molecule symbols and links with bond multiplicities. These graphs are typically quite small (in terms of number of vertices and edges) and the average number of edges per vertex is usually slightly above 2.

Sample classification problems in the area of chemical informatics and computational chemistry include: detection/prediction of mutagenicity, toxicity and anti-cancer activity of chemical compounds for a given organism. There are two major approaches to classifying chemical compounds: quantitative structure-activity relationships (QSAR) and structural approach. The former one (King, Muggleton, Srinivasan, Sternberg, 1996) requires genuine chemical knowledge and concentrates on physico-chemical properties derived from compounds while the latter one (Deshpande, Kuramochi, & Karypis, 2003; Kozak, Kozak, & Stapor, 2007; Dominik, Walczak, & Wojciechowski, 2007) searches directly structure of the compound and discover significant substructures (e.g. contrast, common, frequent