A Connectionist Approach to the Distance-Based Analysis of Relational Data

Kristina Schädler and Fritz Wysotzki

Technical University of Berlin
FR 5-8, Franklinstr. 28/29, D-10587 Berlin, Germany
Phone 0049-30-314 25491, e-mail: schaedle@cs.tu-berlin.de

Abstract. Objects with higher structural complexity often cannot be described by feature vectors without losing important structural information. Several types of structured objects can be represented adequately by labeled graphs. The similarity of such descriptions is difficult to define and to compute. However, many algorithms in machine learning, knowledge discovery, pattern recognition and classification are based on the estimation of the similarity between the analysed objects. In order to make similarity based algorithms like nearest neighbor classifiers, clustering, or generalised prototype learning accessible for the analysis of relational data, a connectionist approach for the determination of the similarity of arbitrary labeled graphs is introduced.

Using an example from organic chemistry, it is shown that classifiers based on the connectionist approach to structural similarity to be considered in this paper perform very satisfactorily in comparison with recent logical and feature vector approaches. Moreover, being able to handle relational data in a natural way without any loss of structural information, the algorithms need only a subset of the given features of the objects for classification.

1 Introduction

By now, an immense number of algorithms for the solution of different tasks in the analysis of datasets and the extraction of knowledge from collections of data is available. Many methods in intelligent data analysis require the estimation of the similarity between the entities of the data base and the query objects. Clustering algorithms divide the dataset into clusters with high intra-class and low inter-class similarity. Distance based classifiers use the similarity between classified objects and the query object for determining its class. In machine learning and knowledge discovery [26, 27], often the aim is to find common characteristic features of objects having similar properties or belonging to the same class. If the objects are represented by feature vectors, some kind of Euclidian or generalised Minkowski metric might be employed.

Objects having a complex structure often cannot be described as fixed length feature vectors without losing important structural information (see [13, 14, 15]).

In machine learning, these objects often are represented in a logical framework (see [1, 23] for an overview). Inductive Logic Programming is used to tackle the problems of learning and classification of structured objects. In many real world applications, it is more natural to describe complex objects or other structures by labeled graphs, for instance chemical structures by structural formulas or computer programs by trees or flow charts, respectively. In this paper, the estimation of the similarity between graphs is discussed and a connectionist approach for computing the similarity of two graphs is introduced. It is a neural net approach to the graph theoretic problem of subgraph isomorphism. The net is able to find an approximate solution of the NP-complete graph matching problem efficiently including domain knowledge about the similarity of objects. For more details about the algorithm and its application to case-based reasoning, see [29, 28].

Based on these results, quite a lot of similarity or distance based algorithms can be made accessible for the treatment of relational data. Two groups of classification algorithms and their performance on a specific problem using the neural approach to similarity estimation described in Section 5 are considered in detail in this paper. The first group consists of two generalised prototype classifiers, learned from examples by a new similarity-based inductive graphtheoretic learning algorithm called SIG-Learning and a prototype learning algorithm from [37] which reduce the set of given instances to a smaller set of generalised prototypes used for classification. The second group of algorithms performs a weighted nearest neighbor classification. The results of the algorithms are compared with those produced by some recent logic, graphtheoretic and feature vector based algorithms, applied to the discovery of cause-effect relationships of some organic compounds (mutagenesis data), a typical data mining problem.

2 Generalised Prototypes: Learning and Classification

2.1 Relational Learning Algorithms

Learning tasks for structured objects often aim at the discovery of certain substructures of a set of objects, for instance the part of the structure that causes a specific common property of the objects. One main approach to relational learning models objects and relations in a logical framework, for instance in the field of Inductive Logic Programming (ILP). In this paper, elementary objects that constitute a structured object and the (binary) relations between them are described by the nodes and edges of graphs representing the structured objects, respectively\(^1\). Colors (labels) of nodes denote one-place relations, i.e. properties of elementary objects, and colors of edges represent the names of the binary relations between objects. A graph is a more general description than another graph if it has less nodes or relations, i.e. it is a part of the latter. A data mining or classification learning task is to find a general description of objects showing or not showing a common property or behavior. Using a graph representation

\(^1\) For a comparison of graph theoretic methods with ILP see [21], Chapter 6. [32] and [31] discuss some more aspects of the graph vs. logic generalisation.