Integrated Data Preparation and Feature Selection in Knowledge Discovery Systems

Iryna Skrypnyk

Metaintelligence Lab., State Technical University of Radio-electronics, Kharkov, Ukraine
Currently visiting researcher at Dept. of CS&IS, Un-ty of Jyvaskyla, Finland (till December 2000)
e-mail: iryna@cs.jyu.fi

Abstract

The goal of this research is to analyze present problems with data preparation in knowledge discovery, and to develop an integrated data preparation and feature selection approach that solves these problems and works in accordance with data characteristics and data mining methods used jointly.

Keywords: feature selection, data mining, knowledge discovery.

1. Research background

Across a wide variety of fields, data is being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist human in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the newly emerging field of Knowledge Discovery in Databases (KDD). The main goal of KDD tools is to automate as much of the knowledge discovery process as possible [1].

The KDD process can be divided into two important parts: data preparation for data mining, and data mining. Input data for KDD systems is usually multidimensional, and presented by a number of features (attributes). The goal of feature selection methods, in general, is to choose (or to produce) a minimum sized subset of features that retains all information needed for further processing [2].

2. Research problems and goals

Nowadays industry professionals and academicians almost universally agree that one of the most important parts of any KDD project, and one of the most time-consuming and difficult, is data preparation. However, even understanding the importance of data preparation, not enough attention is paid to the problem both in literature and in industry applications. Industrial applications do not use one unified and well-defined approach to data preparation. There is a need for a new effective integrated approach to data preparation and feature selection. The choice of feature selection technique has to be carefully substantiated as well as the choice of an appropriate data mining technique. In order to automate this important part of the KDD process, it is necessary to find out formal dependencies and heuristics that will help put into correspondence some feature selection technique to appropriate data. Successful automation of data preparation and feature selection can help provide high accuracy and efficiency for the next step, data mining. In practice, successful data preparation is more than a half of the problem solution, and it is often the key to solving the problem. Recently have been elaborated several advanced techniques for integration of data mining methods, and a meta-level approach to the method selection management that try to automate the choice of an appropriate method for data mining.

This approach can be adopted for integration of data preparation methods as well. It is necessary to take into account also the problems of noise, redundancy and irrelevancy in the data [3]. They may

---

1 This research is partly supported by the Centre for International Mobility (CIMO).
have a bad influence on the final processing result. On the other hand, in practice we have to solve additional problems with distributed, heterogeneous, and different-typed data [4].

The goal of this research is to carefully analyze described above problems, and to develop an integrated data preparation and feature selection approach that solves these problems and works in accordance with data characteristics and data mining methods used jointly.

3. Research approach and methods

The theory creation research method is used for creating conceptual basics of the problem area. A new approach to selection of the most appropriate method or combination of methods for data preparation and, in particular, feature selection, is under development now. This approach is based on the wrapper model, and is dynamical in order to take into account possible heterogeneity of the data. A way of finding important particularities in the initial data is proposed to put into correspondence a group of data preparation methods with the data possessing some concrete characteristics. This knowledge can be used by the procedure of automated search of the optimal data preparation strategy. Besides the theoretical reasoning used in theory formation also constructive research methods to produce software prototypes needed for experimental testing the proposed theories are used. Using the software prototypes the elaborated approach is tested on different real-world and artificial datasets. In the experiments, the proposed integrated data preparation approach will be compared with several other data preparation techniques.

Three basic research approaches are used: theoretical approach, constructive approach, and experimental approach. These approaches are tightly connected and applied in an iterative way.

4. Research results

Preliminary results are already received and published in the given research area. The effectiveness of an advanced feature selection approach is experimentally evaluated [5]. This approach takes into account local information about feature relevance, reducing the time costs and improving the quality of the knowledge discovery. The integrated data preparation approach is being elaborated and tested on real-world data as well as on standard KDD benchmark datasets. As a constructive practical result, a software prototype will be created that will implement the elaborated approach. This software prototype will be applied to several real-world problems in order to check the elaborated theories, and to find their weak and strong places.

5. References