Chapter 47
A Knowledge-Oriented Recommendation System for Machine Learning Algorithm Finding and Data Processing

Man Tianxing
https://orcid.org/0000-0003-2187-1641
Itmo University, St. Petersburg, Russia

Ildar Raisovich Baimuratov
Itmo University, St. Petersburg, Russia

Natalia Alexandrovna Zhukova
St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia

ABSTRACT

With the development of the Big Data, data analysis technology has been actively developed, and now it is used in various subject fields. More and more non-computer professional researchers use machine learning algorithms in their work. Unfortunately, datasets can be messy and knowledge cannot be directly extracted, which is why they need preprocessing. Because of the diversity of the algorithms, it is difficult for researchers to find the most suitable algorithm. Most of them choose algorithms through their intuition. The result is often unsatisfactory. Therefore, this article proposes a recommendation system for data processing. This system consists of an ontology subsystem and an estimation subsystem. Ontology technology is used to represent machine learning algorithm taxonomy, and information-theoretic based criteria are used to form recommendations. This system helps users to apply data processing algorithms without specific knowledge from the data science field.

DOI: 10.4018/978-1-6684-6291-1.ch047

INTRODUCTION

Due to the popularization of Internet people's lives are increasingly dependent on Internet technology. A lot of relevant data are generating from human daily activities. The information contained in these data is very valuable for Internet of Thing (IoT) developers and data analysts in various fields. Extracting useful information from huge data and applying it to life has become a hot topic in academia. Machine learning (ML) algorithms are the most effective tools to extract knowledge from data. Experts enhance and improve ML and data processing technologies, therefore, the number of different types of algorithms for ML increases, and the algorithms itself become more and more complicate. This situation causes the confusion about how to choose, when and how to apply the appropriate algorithm or technology for data processing for researchers.

Currently, the taxonomy of ML algorithms (Ayodele, 2010) is the main basis for researchers to make choices. But these taxonomies have some limitations. First, they usually do not cover all the information about data analysis, they represent only a single "has-a" relationship. Second, building a taxonomy is a complex, long-term process, while data processing technology rapidly advances, therefore, it is hard to keep an ontology up to date. Finally, a taxonomy cannot help the user to decide, which algorithm is more appropriate in one or another specific situation.

In opposition to this, authors propose a knowledge-oriented system to help data analysts build the process of data processing. It consists of two parts: ontology subsystem and estimation subsystem. The ontology subsystem uses the existing taxonomies of ML algorithms. Authors represent them with ontology techniques and create some new ontologies to describe the processes of ML algorithms (e.g. dataset features, output model features, mathematics, process). In these ontologies' authors define new properties to represent the performance of ML algorithms and the process of algorithms in more detailed way. To form recommendations about ML algorithms and evaluate the results of its implementation, authors propose the estimation subsystem. Several information-theory based measures were adopted and incorporated into a system of comprehensive estimation of data and results of ML algorithms implementation. It is worth mentioning that authors also create an ontology of preprocessing technology, which can provide solutions for the defects in dataset neatness.

The workflow of the knowledge-oriented system consists of the following steps. First, users need dataset feature estimation and task requirements as the basis for algorithm selection. After selecting the appropriate algorithm, the interpretation system will give the process of algorithm execution, related parameter settings and measure selection. Authors consider a clustering task as an example. After evaluating, the system suggests implementing clustering and propose a number of clusters for the output model.

The main advantages of this knowledge-oriented system are as follow: 1) This system describes the entire data processing including preprocessing and data analysis. 2) It is built based on ontology technology so that it is expandable and understandable. 3) This system provides a new information measure for estimate the output model.

BACKGROUND

Provide broad definitions and discussions of the topic and incorporate views of others (literature review) into the discussion to support, refute, or demonstrate your position on the topic.[1]