Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 123 (2018) 234–239

Procedia Computer Science

www.elsevier.com/locate/procedia

8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017

Overview of different approaches to solving problems of Data Mining

Kochetov Vadim

National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Moscow, Russian Federation

Ko4etovvadim@gmail.com

Abstract

This paper is devoted to the main tasks in the analysis of large amounts of information and comparison of methods for their solution. The analysis of large volumes of information and identification of valuable knowledge provided by Data Mining tools. The concept of Data Mining is translated as data mining, data analysis, data collection. Due to of the huge variety of data types and forms of organizing information actual data may not always be analyzed by machine learning tools. For the transformation of "raw" data to the data, which can work efficiently Data Mining techniques, solve the problem of pre-processing. The methods k-nearest neighbor and decision trees solve such problems as the Data Mining classification and regression in the specified domains.

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures

Keywords: Data Mining, the nearest neighbor method, the method of k-nearest neighbor, decision trees, classification, regression, forecasting.

1 Introduction

Due to the wide variety of Data Mining techniques and many different types of information and forms of data presentation it is necessary to define the limits of the applicability and relevance of certain methods according to the provided data and the achieved objectives. It is also necessary to understand how the problem should be solved with the Data Mining south as classification, regression, clustering and so on. After the success of the data processing result every step is determined from the choice of data and ending with an explanation of the anomalies in the results. In this paper, the characteristics methods k-nearest neighbor and decision trees will be studied solutions in various subject areas.

As we study the data for the different subject areas data features and methods of work with them are change depending on the area. Data Mining is an important step in data preprocessing. The process of pre-treatment is often the most time-consuming and protractedly. Sometimes it occupies a

Peer-review under responsibility of the scientific committee of the 8th Annual International Conference on Biologically Inspired Cognitive Architectures
10.1016/j.procs.2018.01.036

large part of the entire process of Data Mining. In addition, a lot of resources and time spent on the choice of the model and its training.

2 Data Mining Tasks

There are the following categories of Data Mining problem solving: supervised learning (with teacher training) and unsupervised learning (learning without a teacher). The term comes from Machine Learning. Is Machine Learning a concept that combines the entire set of Data Mining Technologies. [2]

In the case of supervised learning analysis task is solved achieved in several stages. Firstly, with a Data Mining algorithm a model of the analyzed data (classifier) is built. Then over qualifier made education. In other words, we checked the quality of his work, and if it is not satisfactory, there is an additional training of the classifier. This continues until it reaches the desired quality level or not will be clear that the selected algorithm does not work correctly with the supplied data or the data do not have the structure that could be identified. This type of problems are the problems of classification and regression.

Unsupervised learning combines tasks, tapping descriptive models, such as the laws in the procurement of large customers wholesale supplier. Obviously, if these laws are, the model should present them and inappropriate to talk of her training. The advantage of such problems is the ability to solve them without any prior knowledge about the data being analyzed.

The problem of classification and regression.

Classification of - one of the areas of machine learning problems. Tasks that area solve the following problem. For example, there is a certain set of objects (entities), distributed in a certain way to classes. It is also known for a limited set of entities, distributed by classes in a known manner. This set is called the training sample. The distribution of the remaining entities in classes is unknown.

Regression (forecasting) - Regression - a task based on the classification, only the data consist of the values of the dependent variable (response variable) and independent variables (explanatory variables). The independent variables are the values of object attributes. Variables can take on any value on the set of real numbers.

Solution occurs in two stages.

In the first stage, based on the training sample is based model for determining the value of the dependent variable. It is often referred to as a function of the classification or regression. For the most accurate function to the training sample must meet the following basic requirements:

- the number of objects included in the sample must be sufficiently large;
- The sample should include objects representing all possible classes

  in the case of the classification of tasks or the entire range of values in the case of the regression problem;
- for each class in the classification problem and for each interval range of values in the problem of regression sample should contain a sufficient number of objects;
- distribution of a sample of objects in classes should be as uniform.

At the second stage of construction of the model is applied to the analyzed objects (Objects with an undefined value of the dependent variable). The problem of classification and regression has a geometric interpretation. Consider it an example with two independent variables, which will submit it to the two-dimensional space.