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.

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

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.
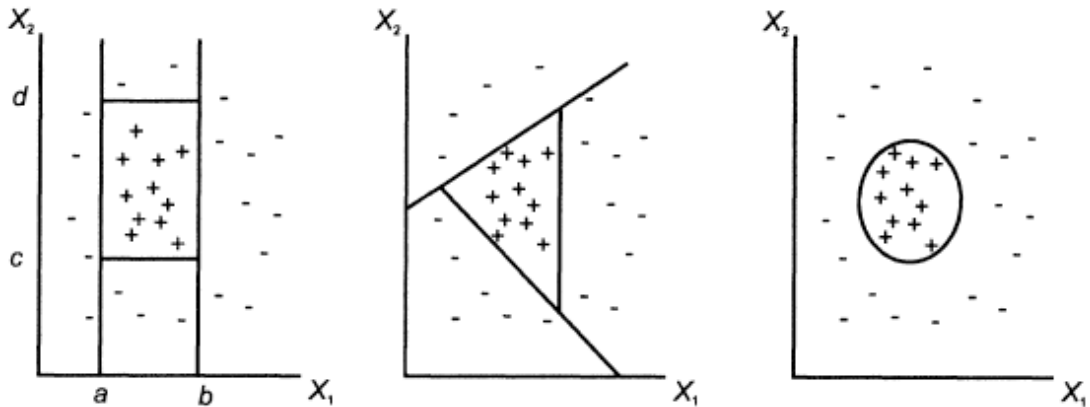
Fig. 1 Classification of two-dimensional space

The set of objects is displayed on the set of points on the plane. The "+" and "-" represent the class of the object. The classification function should be structured in such a way as to result in a surface in any way cover all the points in the middle of the field, where there is a cluster of objects of the same class. The function must be set to "+" in the field and "-" outside.

As can be seen from the figure, there are several options for building Stroke area. View function depends on the algorithm used.

The main problems faced in dealing with classification problems and regression - is the poor quality of the original data, in which there are as erroneous data and missing values; attribute types, not suitable for the analysis of Data Mining tools, the presence of the dominant classes, as well as the so-called problem of overfitting (retraining) and underfitting (weak model). The essence of the first of these is that the classification function in the construction of "too good" adapts to the data and found in them the errors and trying to interpret anomalous values as part of the internal data structure. It's obvious that in the future, this model will not work with other data where the error character will be slightly different. The term underfitting denote a situation where too large number of errors when checking the classifier on the training set. This means that specific patterns in the data was found, and either they do not exist at all, or you need to select a different method of detection. [2]

3 Classification Data Mining stages

Data Mining The process can be divided into the following stages: [3]
• The first stage - the stage of determining the relationships and regularities.
• The second stage - the application of these rules to predict other objects.
• The third stage - the analysis of variance. This process deviations and error analysis.
Thus, Data Mining process can be represented as a sequence of three steps: [4]
Defining relationships => application of these rules => variance analysis
Let us consider in detail each of the stages.
1. Defining Relationships.
At the stage of determining the relationships explored data to identify hidden patterns and relationships.
Identify the relationship attained by the following actions:
• identification of patterns of conditional logic;
• identification of patterns of associative logic;
• identify trends and variances.
The tools for these activities are:

• induction rules conditional logic (problem of classification and clustering, the description in the compact form of close or similar groups of objects);

• induction rules associative logic (Association task and sequences and retrieve them with the help of information);

• determining trends and fluctuations (initial stage of the prediction task).

At the stage of determining the relationship dependencies are checked, that is, checking their reliability on the data samples that have participated in the formation of relationships. This method of data division to train and test set used for the method of decision trees.

2. Predictive modeling.

The second stage of Data Mining - results obtained from the first stage are used to build a predictive model. Here, the observed regularities are directly used for the prediction.

This phase includes the following steps:

• determination of unknown values;

• forecasting the dynamics of the processes.

In the process of predictive modeling classification problem solved and forecasting.

In solving the problem of forecasting the results of the first stage (definition of a trend or oscillation) is used to predict the unknown (or missing the future) target variable values (variables).

Predictive modeling, on the other hand, deductively. The regularities obtained at this stage, formed from the general to the particular and the individual. Here we obtain new knowledge about some object or group of objects on the basis of:

• Class of knowledge, to which belong to the objects under study;

• Knowledge of general rules in force within a given class of objects.

3. Analysis of the exceptions (forensic analysis) In the third stage, Data Mining analyzes the exceptions or anomalies identified in the found regularities.

The action performed at this stage - the identification of deviations. To identify deviations is necessary to determine the rate, which is calculated on the free search stage. The reason for deviation can be a logical deviation, which can be formulated in the form of rules. Another reason may be errors of initial data. In this case, exclusion analysis step can be used as a data cleaning. [5]

4 Classification Data Mining methods

Consider the most widely used classification of the Data Mining techniques.

All Data Mining techniques are divided into two large groups according to the principle of working with the original data. The upper level of this classification is determined based on whether the data Data Mining after they are stored or distilled for later use.

1. The direct use of the data, or save the data. In this case, the original data is stored in a detailed and explicit directly used for predictive modeling steps and / or analysis exceptions. The problem with this group of methods is that their use can be difficult analyzing very large databases. Methods nearest neighbor method and k-nearest neighbor belongs to this group.

2. Identify and use of formal laws or distillation templates. When distillation technology templates one sample (template) information is extracted from the raw data is converted into a certain formal structure whose form depends on the Data Mining method. This process is performed on the stage of the free search. Note that the first group of methods, this step basically absent.

At stages predictive modeling and analysis of variance used definition phase relations means that they are much more compact databases themselves. Decision trees and symbolic rules relate to this group.

Methods of this group is perhaps the most interpreted - in most cases they make out patterns found in a sufficiently transparent manner from the point of view of the user.

Predictive methods use the values of some variables to predict unknown (missing) or future values of other (target) variables. The methods used in this study (decision trees, nearest neighbor) are methods aimed at obtaining predictive results.

5 Conclusions

When solving specific problems that can be higher than those of the tested samples, since for a certain model of parameters it is necessary to be selected in a special way. However, in order to trace the trend and the behavior of the characteristics of Data Mining enough universal models without special configuration.

Analyzing the results, for each descriptor of the characteristic.

Trees of solutions.

Reaction to increase the volume of data:

• the average accuracy does not change;

• the coefficient of determination for regression trees is almost unchanged, but there is little growth;

• The operating time increases slightly.

Trees of solutions built for given subject areas, personal effectiveness. However, with an increase in the number of informative signs, the method falls sharply. This is due to the curse dimension, when the linear increase in the number of objects of experimental data increases exponentially. The method is well interpreted both in graphical form and in the form of symbolic rules. Has a high working speed.

Let us consider the results for the method of k-nearest neighbors.

Reaction to increase the volume of data:

• the average accuracy does not change;

• time to increase increases;

• the coefficient of determination on average has a weak growth.

Change in quantitative coefficients of satisfactory quality. Models built for given subject areas by this method of efficiency. The method is well interpreted for problems of small dimensions, for example in the form of graphs, on a plane or in space. The sharp increase in the operating time on increasing linear volumes of data is explained by the most complicated technique. This method stores all known information about objects in memory, unlike decision trees that store only the results of the constructed data relationships. Has a high working speed.

So, both methods are suitable for use on specified subject areas, high accuracy and fastness. The decision tree method is faster than the method of k-nearest neighbors. The k-nearest neighbor method has a relatively low scalability, while the decision tree method scales without increasing the learning time. Precision simulates solutions at a more complex level. Trees of decision making are satisfactory better interpretability.

References

[1] *Data mining*, Chubukova I. A., Binom. 2008 384 p.

[2] *Analysis of data and processes: Textbook*. allowance Barseghyan A. A., Kupriyanov M.S., Kholod I.I., Tess M.D., SI Elizarov. - 3rd ed., Pererab. and additional. - St. Petersburg: BHV-Petersburg, 2009 512 p.

[3] *Characteristics of technologies and processes for data mining Data log*. Parsaye K A 1998.

[4] *Down with the dirt!* Aragon L. Week of PC / RE. 1998. No. 6. C. 53-54

[5] *Methods of analytical data processing to support DBMS*, Shchavelev L.V., decision-making. 1998. № 4-5

[6] *Electronic textbook "Statistica"*. Access mode: http://www.statsoft.ru/home/textbook.htm

[7] *Department of Transportation Statistics USA* Access mode: http://www.transtats.bts.gov

[8] UCI. *Machine learning repository*. Center for machine learning and intelligent systems. Access mode: https://archive.ics.uci.edu/ml/datasets/Wholesale+customers#

[9] *Yahoo Finance* Access mode: http://finance.yahoo.com/

[10] *Self-learning systems, machine learning (course of lectures)* Access mode: http://logic.pdmi.ras.ru/~sergey/index.php?page=ml - Заклл. From the screen

[11] *Pandas* - Access mode: http://pandas.pydata.org