Code 004.891

APPLIED USE OF OLAP-TECHNOLOGIES IN MACHINE LEARNING

Ulezla Dzmitry

Yanka Kupala Grodno State University

Abstract

Metric classification methods and their features, OLAP-cubes advantages, methods of their application in practice are studied. This work discusses the methods of OLAP-technologies applied use in machine learning, their synthesis with metric classification objects methods.

Аннотация

Изучаются методы метрической классификации и их особенности, преимущества OLAP-кубов данных, способы применения их на практике. В данной работе рассматриваются способы прикладного использования OLAP-технологий в машинном обучении, их синтез с метрическими методами классификации объектов.

Introduction

Identification of patterns, similar object's properties searching is important process in every applied science. They allow you to put forward new hypotheses and conjectures, confirm suggestions in practice. Modeling of complex systems, objects and processes allows you to make it.

In complex system modeling there arises the problem of classifying some objects with a set of certain features. In some cases, replacement of specialists by automated expert system can save significant amount of money, staff hours and resources. Such problems can be solved by using of machine learning techniques and methods that enables to classify with high probability some objects, according to certain parameters.

Object classification

There is a need to consider the algorithmic and mathematical aspects of the problem, identify the strengths and weaknesses of these approaches.

Metric classifier - a classification algorithm based on the ratings of similarity between objects calculation. The simplest metric classifier is the nearest-neighbor method, in which the classified object belongs to the class which owns the most similar to their features objects. A classic example of the problem of metric classification is the classification of iris flowers (Fisher, 1936.).

To formalize concept of similarity we will enter term function of the distance between objects. As a rule, the strict requirement that this function was metric - is not shown; in particular, the triangle inequality quite can be violated.

"Compactness hypothesis" suggests that similar properties are more often in the same class than in different. This means that the boundary between the classes usually has rather simple shape and compact classes form localized areas in the object space. Metric algorithms are based on this hypothesis [1].

If objects are described by numerical vectors, Euclidean metric is usually taken. If thera are too much features, and the distance is calculated as the sum of the deviations for individual characteristics, than the curse of dimensionality can appear. The sum of a large number of deviations with high probability will have very similar values (according to the law of large numbers). It turns out that in the space of high-dimensional objects all objects are equally distant from each other.

The problem is solved by the selection of a relatively small number of informative features. In the estimation algorithms we construct a variety of different feature sets - support

sets, for each set we construct its proximity function. We may also consider, that different features have different weights.

Advantages: Easy to implement and understand, easy interpretability - automatic expert system can explain its decision.

Disadvantages: instability to noise and emissions - random erroneous objects used during the training phase, the need to store whole entire sample set.

Noise problem is solved by a comprehensive analysis of prior learning sample, allowing us to filter the sample to get rid of noise.

Various methods may use the same data for classification in different ways, depending on the goal and the characteristic you want to predict. Also methods of noise emissions control and filtering can be varied. For easy storage, filtering and access to training samples, the size of which can reach gigabytes and terabytes, you can use OLAP-technology.

OLAP-technologies

OLAP (online analytical processing, real time analytical processing) — data processing technology, which consists of the preparation of the total (aggregated) information based on large amounts of data, structured according to the multi-dimensional principle. OLAP technology implementations are components of Business Intelligence software solutions class.

OLAP enables you to handle requests with great speed. Relational databases store entities in separate tables, which are usually well-normalized. This structure is suitable for operating with the database, but complex multi-table queries works usually relatively slow [2].

OLAP-cube - OLAP-structure created from the operational data. OLAP-cube contains the basic data and information about the dimensions (units). Cube potentially contains all the information that may be required to answer any requests. When data has a huge number of units, full calculation is often made only for some measurements. And remaining measurements are used only when request is connected with them[3].

Queries to OLAP-cubes containing training samples are written considering the necessary filtering on MDX language.

Conclusion

As you can see, the possibilities afforded us by OLAP-technologies can be used in machine learning for the object classification. Range of solved tasks and problems is extensive. But this approach requires a preliminary analysis, the selection of individual metrics and filter implementation using MDX-scripts for each specific task on the same training set.

References:

1. Hasie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning. Data Mining, Inference and Prediction, 2-nd edition – Springer, 2009 – 764p.

2. Ulezla, D.S. Some problems of processing Big Data / D.S. Ulezla, A.M. Kadan // Информационные компьютерные технологии: проектирование, разработка, применение : сб. науч. ст. / ГрГУ им. Я. Купалы ; редкол.: А. М. Кадан (гл. ред.) [и др.]. – Гродно : ГрГУ, 2013. – С.86-89.

3. Кадан, А.М. Информационно-технологические решения на основе Бизнес Интеллекта в сфере управления университетом / А.М. Кадан, Е.Н. Ливак. - Вестник ГрГУ, Серия 2, Математика. Физика. Информатика, вычислительная техника и управление. Биология.– Гродно: ГрГУ, 2010. - №2(96). – С. 123-131.