анализ), необычные записи (обнаружение аномалий) и зависимости (извлечение правил ассоциации). Это обычно включает использование методов базы данных, таких как пространственные индексы. Эти шаблоны затем можно рассматривать как своего рода сводку входных данных, и их можно использовать в дальнейшем анализе или, например, в машинном обучении и прогнозной аналитике. Например, этап интеллектуального анализа данных может идентифицировать несколько групп в данных, которые затем можно использовать для получения более точных результатов прогнозирования системой поддержки принятия решений. Ни сбор данных, ни подготовка данных, ни интерпретация результатов и создание отчетов не являются частью этапа анализа данных, связанные термины «драгирование данных», «промысел данных» и «отслеживание данных» относятся к использованию методов интеллектуального анализа данных для выборки частей большого набора данных о населении, которые (или могут быть) слишком малы, чтобы можно было сделать надежные статистические выводы о достоверности любых данных. Шаблоны обнаружены. Эти методы, однако, могут быть использованы при создании новых гипотез для проверки больших массивов данных.

Заключение

И в конце хотелось бы отметить, что область использования интеллектуального анализа данных ничем не ограничена - он везде, где есть большое количество информации.

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining. БХВ-Петербург, 2004
2. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. Питер, 2013
3. Бергер А. Microsoft SQL Server 2005 Analysis Services. OLAP и многомерный анализ данных. БХВ-Петербург, 2007
4. Макленнен Д., Танг Ч., Криват Б. Microsoft SQL Server 2008: Data Mining — интеллектуальный анализ данных. БХВ-Петербург, 2009
5. Чубукова И. А. Data Mining. Интернет-Университет Информационных Технологий (ИНТУИТ), 2016
6. Мастицкий С.Э., Шитиков В.К., Классификация, регрессия и другие алгоритмы Data Mining с использованием R. Тольятти, 2017
7. Сегаран Т. Программируем коллективный разум. Символ-Плюс, 2008
8. Mitra S., Acharya T., Data Mining. Multimedia, Soft Computing, and Bioinformatics. – John Wiley & Sons, Inc. – Hoboken, New Jersey. – 2003.
9. Konar A., Artificial intelligence and soft computing: behavioral and cognitive modeling of the human brain. – CRC Press LLC. – Boca Raton, Florida/ – 2000.
10. Han J., Kamber M., Data mining: Concepts and Techniques. – Morgan Kaufmann Publishers. – 2001
11. [Electronic resource] URL: https://habr.com/ru/hub/data_mining/ (accessed: 15.09.2019)

Komlev G.V., Mitrofanova A.S.
Data mining models and methods

Reshetnev Siberian State University of Science and Technology
(Russia, Krasnoyarsk)

doi: 10.18411/lj-10-2019-27

idsp: ljjournal-10-2019-27

Аннотация

В этой статье описаны модели, методы интеллектуального анализа данных и возможные варианты их использования: базовая архитектура моделей и структур, свойства моделей интеллектуального анализа данных.

Ключевые слова: интеллектуальный анализ, модели, методы.

Abstract

This article describes the models, data mining methods and possible options for their use: the basic architecture of models and structures, the properties of data mining models.

Keywords: intellectual analysis, models, methods.

Data mining models are created by applying an algorithm to data. But this is more than an algorithm or a metadata container: it is a set of data, statistics and templates that can be applied to new data to form forecasts and derive relationships.

The data mining model obtains data from the data mining structure and analyzes it using the data mining algorithm. The data mining structure and data mining model are separate objects. The data mining structure stores information that identifies the data source. The data mining model contains information obtained from the results of statistical data processing, for example, patterns discovered as a result of analysis.

The process of data analysis, search and building a model is often iterative, as you need to find and identify various information that can be extracted. You must also understand how to link, transform and combine them with other data to obtain a result. After the discovery of new elements and aspects of data, the approach to identifying data sources and formats with subsequent matching of this information with a given result may change.

There are many different methods of data mining, query modeling, processing and collection of information.

In fact, data mining is the processing of information and identifying patterns and trends in it that help make decisions. The principles of data mining have been known for many years, but with the advent of big data they have become even more widespread.

Big data has led to an explosive increase in the popularity of wider data mining methods, partly because there is much more information, and by its very nature and content it is becoming more diverse and extensive [1].

This article describes the models, data mining methods and possible options for their use: the basic architecture of models and structures, the properties of data mining models.

Predictive models.

Predictive models are built on the basis of a data set with known results. They are used to predict results based on other data sets. At the same time, of course, it is required that the model work as accurately as possible, be statistically significant and justifiable, and so on.

These include:

- classification models - describe the rules or a set of rules according to which the description of any new object can be attributed to one of the classes. Such rules are based on information about existing objects by breaking them into classes.

Classification can be used to get an idea of the type of customer, product, or item, by describing several attributes to identify a particular class. For example, cars can be easily classified by type (sedan, SUV, convertible) by defining various attributes (number of seats, body shape, drive wheels). Studying a new car, you can attribute it to a certain class by comparing attributes with a well-known definition. The same principles can be applied to customers, for example, classifying them by age and social group.

- sequence models - describe functions that allow predicting a change in continuous numerical parameters. They are based on data on a change in a parameter over the past period of time.

Descriptive Models

Descriptive models pay attention to the essence of dependencies in a data set, the mutual influence of various factors, that is, the construction of empirical models of various systems. The key point in such models is lightness and transparency for human perception. It is possible that the discovered patterns will be a specific feature of specific research data and will not be found anywhere else, but it can still be useful and therefore should be known.

These include:

- regression models - describe the functional relationships between dependent and independent indicators and variables in a human-readable form. It should be noted that such models functionally describe the relationship not only between continuous numerical parameters, but also between categorical ones.

- clustering models - describe groups (clusters) into which you can divide objects whose data are analyzed. Objects (observations, events) are grouped based on data (properties) that describe the essence of objects. Objects within a cluster must be “similar” to each other and different from objects included in other clusters. The more similar the objects within the cluster and the more differences between the clusters, the more accurate the clustering.

By examining one or more attributes or classes, you can group individual data elements together to get a structured conclusion. At a simple level, clustering uses one or more attributes as the basis for determining a cluster of similar results. Clustering is useful in defining various information because it correlates with other examples, so you can see where the similarities and ranges are consistent.

- exception models - describe exceptional situations in records (for example, individual patients), which differ sharply in anything from the main set of records (group of patients). Knowledge of exceptions can be used in two ways. It is possible that these records are an accidental failure, for example, errors of operators who entered data into a computer. A typical case: if the operator, mistakenly, puts the decimal point in the wrong place, then such an error immediately gives a sharp “splash” by an order of magnitude. It makes sense to discard such a “noise”, random component, and exclude it from further studies, since most of the methods that will be discussed in this chapter are very sensitive to the presence of “outliers” - sharply different points, rare, atypical cases. On the other hand, separate, exceptional records may be of independent interest for research, since they may indicate some rare but important abnormal diseases. Even the identification of these records, not to mention their subsequent analysis and detailed consideration, can be very useful for understanding the essence of the studied objects or phenomena.

- final models - identification of restrictions on the data of the analyzed array. For example, when studying a sample of data on patients no older than 30 years who have had myocardial infarction, it is found that all patients described in this sample either smoke more than 5 packs of cigarettes per day or have a weight of at least 95 kg. Such limitations are important for understanding array data; in fact, it is new knowledge extracted from analysis. Thus, Data Summarization is the finding of some facts that are true for all or almost all records in the studied data sample, but which would rarely be found in the entire imaginable variety of records of the same format and, for example, would be characterized by the same distribution of values fields. If we take information on all patients for comparison, the percentage of either heavily smokers or overly obese people will be very small. It can be said that the seemingly implicit classification problem is being solved, although in fact only one class is presented, represented by the available data.

- association models - identifying patterns between related events. An example of such a pattern is a rule indicating that event Y follows from event X. Such rules are called associative.

An association (or attitude) is probably the most famous, familiar, and simplest data mining model. To identify patterns, a simple comparison of two or more elements is made, often of the same type. For example, by tracking your buying habits, you may notice that cream is usually bought with strawberries.

To build the considered models, various methods and algorithms of intellectual analysis are used. Due to the fact that the technology of mining analysis has developed and is developing at the intersection of such disciplines as statistics, information theory, machine learning, database theory, it is quite natural that most algorithms and methods of mining were developed on the basis of various technologies and concepts. Consider the technologies most often implemented by the methods of mining.

Several basic methods that are used for data mining describe the type of analysis and the data recovery operation. Unfortunately, different companies and solutions do not always use the same terms, which can exacerbate the confusion and apparent complexity.

Let's look at some key methods and examples of how to use certain tools for data mining.

Basic methods

It is customary to attribute to the basic methods of mining analysis, first of all, algorithms based on enumeration [2].

A simple enumeration of all the studied objects requires 2^N operations, where N is the number of objects. Therefore, with an increase in the amount of data, the amount of computation grows exponentially, which with a large amount makes the solution of any problem using this method practically impossible.

To reduce the computational complexity in such algorithms, as a rule, different types of heuristics are used, leading to a reduction in enumeration. The optimization of such algorithms is reduced to reducing the dependence of the number of operations on the number of data under investigation to a linear function. At the same time, the dependence on the number of attributes, as a rule, remains exponential. Provided that there are few of them (in the vast majority of cases there are significantly fewer than data), this dependence is acceptable.

The main advantage of these algorithms is their simplicity, both in terms of understanding and implementation. The disadvantages include the lack of a formal theory on the basis of which such algorithms are built, and therefore the difficulties associated with their study and development.

The basic methods of mining can also include approaches that use elements of the theory of statistics. Due to the fact that mining is the development of statistics, there are a lot of such methods. Their main idea is reduced to correlation, regression, and other types of statistical analysis. The main disadvantage is the averaging of values, which leads to a loss of information content.

The main way to study the problems of data analysis is to map them to a formalized language and then analyze the resulting model. The uncertainty in the amount of missing information from a system analyst can be divided into three large groups:

- suspense;
- incompleteness (insufficiency, inadequacy);
- unreliability.

Inaccuracy is physical (its source is the external environment) and linguistic (arises as a result of verbal generalization and is caused by the need to describe an infinite number of situations by a limited number of words for a limited time).

There are two types of physical uncertainty:

- inaccuracy (inaccuracy of measurements of values of a certain value performed by physical devices);
- randomness (or the presence in the external environment of several possibilities, each of which may randomly become a reality; knowledge of the corresponding law of probability distribution is assumed).

There are two types of linguistic uncertainty:

- the uncertainty of the meanings of words (ambiguity, vagueness, ambiguity, fuzziness). It arises if the objects of the control problem displayed by the same word are different;
- the ambiguity of the meaning of phrases (distinguish syntactic and semantic).

Probability theory methods and classical set theory are successfully used to handle physical uncertainties. However, with the development of systems that use the methods of the theory of artificial intelligence, in which it is necessary to process the concepts and relations of a natural language, it became necessary to expand many formal methods to take into account the linguistic uncertainty of tasks.

Genetic Algorithms

Genetic algorithms are among the universal optimization methods that allow solving problems of various types (combinatorial, general problems with and without restrictions) and of varying degrees of complexity. Moreover, genetic algorithms are characterized by the possibility of both single-criteria and multi-criteria search in a large space, the landscape of which is nonsmooth [3].

In recent years, the number of works, especially foreign scientists, devoted to the development of the theory of genetic algorithms and questions of their practical use, has increased dramatically. The results of these studies show, in particular, that genetic algorithms can be more widely used when integrated with other methods and technologies. There are works that prove the effectiveness of the integration of genetic algorithms and methods of the theory of fuzziness, as well as neural computing and systems.

Neural networks

Neural networks are a class of models based on biological analogy with the human brain and designed after passing through the stage of the so-called training on the available data to solve various problems of data analysis. When applying these methods, first of all, the question arises of choosing a specific network architecture (the number of “layers” and the number of “neurons” in each of them).

Data mining is not only used database tools or software. Data mining can be performed with relatively modest database systems and simple tools, including creating your own, or using off-the-shelf software packages. Sophisticated data mining relies on past experience and algorithms defined using existing software and packages, and different specialized tools are associated with various methods.

1. Gavrilova, T.A. Intelligent systems knowledge base / T.A. Gavrilova, V.F. Horoshevskij. – SPb. : PITER, 2001. – 382 p.
2. Davydenko, V. Data Mining / V. Davydenko // Software products and systems. – 2007. – № 3(79). – P. 20 – 31.
3. Korneev, V.V. Database. Intelligent information processing / V.V. Korneev, A.F. Gareev, S.V. Vasyutin. – M. : NOLIDZH, 2001. – 351 p.

Komlev G.V., Mitrofanova A.S.

Expert systems

*Reshetnev Siberian State University of Science and Technology
(Russia, Krasnoyarsk)*

doi: 10.18411/lj-10-2019-28

idsp: ljjournal-10-2019-28

Аннотация

В статье рассматриваются определение экспертных систем, их структура, классификация, и также технология работы с ними.

Ключевые слова: экспертная система, искусственный интеллект, анализ, синтез.

Abstract

The article discusses the definition of expert systems, their structure, classification, and also the technology of working with them.

Keywords: expert system, artificial intelligence, analysis, synthesis.

Artificial intelligence, as a scientific discipline, consists of several large currents. One of them is expert systems.