CLASSIFICATION OF BASIC METHODS FOR DATA ANALYSIS IN INDUSTRY

Nikolina Ivanova

University of Mining and Geology “St. Ivan Rilski”, 1700 Sofia; E-mail: nikolina.ivanova@mgu.bg

ABSTRACT. In today's technological world, data is being generated in vast quantities by machines, people, and 'things' – everything we do in the digital world is turning into big data sets. Their sheer volume, variety and velocity categorises them as Big Data. Coming from various sources, they often come in different sizes and formats, necessitating the use of new methods and technologies to use and turn them into useful information.
Effectively used data leads to a better understanding of the past performance of business and helps to make informed decisions about its future activities. Different industries are implementing new technologies and advanced data analytics models based on IoT and ML to improve and optimise their processes and predict future consequences or risks. Below are presented some interesting uses of data analytics in industry.
This paper examines a data analysis model that sets out the steps that are followed in the process of preparing, cleaning, and analysing data sets. In order to determine what analysis can be performed on the data, it is extremely important for data analysts to identify the characteristics of the given data and the types of the underlying variables. For this reason, a classification of the main types of variables is presented. Also, the types of basic data analytics are classified, with an emphasis on big data analytics approaches.

Key words: big data, data analysis, data analysis lifecycle.

КЛАСИФИКАЦИЯ НА ОСНОВНИТЕ МЕТОДИ ЗА АНАЛИЗИ НА ДАННИ В ИНДУСТРИЯТА

Николина Иванова

[1]Минно-геоложки университет „Св. Иван Рилски“, 1700 София

РЕЗЮМЕ. В днешния технологичен свят, данни биват генерирани в огромни количества от машини, хора и „неща“ – всичко, което правим в дигиталния свят, се превръща в големи набори от данни. Техният огромен обем, разнообразие и скорост ги категоризира като големи данни. Идващи от разнообразни източници, те често биват в различни размери и формати, което налага използването на нови методи и технологии за използването и превръщането им в полезна информация.
Ефективното използване на данните води до по-добро разбиране на досегашното представяне на бизнеса и спомага за вземането на информирани решения за бъдещите му дейности. Различните отрасли на индустрията внедряват нови технологии и усъвършенствани модели за анализи на данни, базирани на IoT и ML, за да подобрят и оптимизират своите процеси, както и да предвидят бъдещи последствия или рискове. По-долу биват представени някои интересни употреби на анализите на данни в индустрията.
В статията се разглежда модел за анализ на данни, задаващ стъпките, които се следват в процеса на подготвяне, изчистване и анализиране на набори от данни. За да могат да определят какви анализи може да се извършват върху данните, за специалистите по извършване на анализи на данни е изключително важно да идентифицират характеристиките на данните, които им се предоставят, както и от какъв тип са основните променливи. По тази причина е представена класификация на основните типове променливи. Също така са класифицирани основните видове анализи на данни, с акцент върху подходите за анализи на големи данни.

Ключови думи: големи данни, анализи на данни, жизнен цикъл на анализите на данни.

Introduction

One of the significant consequences of the digital world is the creation of vast volumes of raw data. Data generated from a variety of sources, including sensors, logs, and social media, can be used both on its own and as a supplement to the existing transactional data that many organisations already have.

Companies can analyse customer preferences and promising trends by using data analytics. That can improve their product and services, reduce costs, and lead to new, more successful ways of doing business.

According to research by the International Data Corporation (IDC, 2021), global spending on big data and business analytics (BDA) solutions in 2021 reached $215.7 billion, which is 10.1% more than in 2020. They also present a global compound annual growth rate (CAGR) forecast for BDA spending of 12.8% in the period 2021-2025.

As the amount of data stored increases, especially in technologically advanced companies, the issue of managing raw "big data" becomes much more important.

Big data analytics is a process that discovers patterns, trends, and relationships in large volumes of data that cannot be discovered with traditional data processing techniques and tools. Typically, big data analysis occurs in real time, at the time the data is generated.

Quantity and quality data analysis

When performing an analysis, it is crucial to define the key characteristics to be observed or measured. These characteristics are called variables. Recording the values, patterns, and events for a set of variables is called observation. The collection of observations constitutes the data set for analysis. When looking for meaningful patterns in data, it is often investigated whether correlational relationships exist between variables. There are two main groups of studies being conducted, according to the type of variables in the data set (Fig. 1).

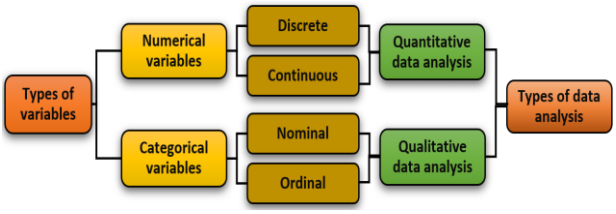

Fig. 1. Classification of types of variables and analyses

Quantitative data analysis

Quantitative data analysis involves working with numerical variables and focuses on numbers and mathematical calculations. Quantitative data is defined as statistical data and is mostly structured. Techniques for quantitative analysis of all data include working with algorithms, statistical observations, mathematical analysis tools, and software for manipulating data and uncovering insights. Quantitative analysis is often used to quantitatively explain certain occurrences or to make predictions. Numerical variables are divided into two groups:

- continuous – can be measured in a range of values, e.g. measuring temperature, body weight, etc.
- discrete – they are continuous but have a specific value within a finite set of values, e.g. number of visits, number of sales, etc.

Qualitative analysis

Qualitative data describes information that is usually not numerical. Qualitative data is non-statistical and usually unstructured or semi-structured. This data is not measured using numbers. Instead, it is categorised based on properties, attributes, tags, and other identifiers. Qualitative data can be used to ask the question "Why?". The qualitative analyses are used for theorising, interpretations, hypothesis development, and initial understandings. Qualitative variables are divided into the following groups:

- ordinal – consist of two or more categories in which order matters, e.g. ranking in a competition, level of education, etc.
- nominal - consist of two or more categories whose value is assigned based on the identity of the object, e.g. gender, car brand, etc.

Data analysis lifecycle

The data analysis lifecycle represents a process in which raw data is analysed to identify patterns, discover correlations and hidden relationships that cannot be found with traditional data processing techniques and tools. Any different type of data can be subjected to analysis to provide useful insights. They can help to optimise processes and increase overall business or system performance.

The steps followed in performing data analysis are called a data analysis lifecycle. They are shown in Fig. 2, as well as discussed in detail after it:

Fig. 2. Data analysis lifecycle steps

Defining the goal

First of all, it should be specified why the particular data analysis is being conducted, what the purpose of the study is, and what data will be used. Companies specify their requirements, and it is up to the data analytics specialists to decide which type of analysis is most appropriate in the particular case.

Data collection

Data collection is the extraction of the data that will be needed for the analysis. With Big Data, data almost always comes from more than one source and in different formats. Nevertheless, they can be structured, making them easy to understand and process. Unstructured data, on the other hand, requires significant processing to make sense of and process it. Significant value can be derived from combining structured and unstructured data for analysis. Another important aspect is the timing – they are often transmitted in near real time.

Data cleaning

The data making up the dataset for the analysis can be very heterogeneous (coming from files, the Internet, sensors, databases). When combining them, it may turn out that their formats and types are incompatible with each other, which requires a good cleaning and organisation before proceeding with analysis. Cleaning may include:

- removal/standardisation of missing or inappropriate data;
- removal of duplicate records;
- data aggregation;
- sorting and arrangement of data;
- application of validation rules;
- conversion of data types – e.g. numeric data represented as text must be converted to the appropriate numeric type;
- unification of data formats from different sources – time and date format is essential when processing time series.

Data analysis

Once the data has been thoroughly cleaned and processed, it can be analysed. The first thing in this step is to select the tools that will be used (Python, R, Excel, MATLAB, Rapid Miner, etc.). The choice of tools depends on the type of analysis to be performed. Some of the tools are specialised in processing and visualising large volumes of data, while others use complex mathematical modelling and simulations to predict future events and outcomes.

Data visualisation

Visualisation techniques range from simple line graphs or histograms, to complex charts such as related graphs. Graphical presentation of the results helps companies to better understand the results, to more easily extract valuable insights from the data, and to more clearly display relationships and trends within them. Graphics can even be interactive or animated for even easier perception.

Types of data analysis

There are four main categories of data analysis (Fig. 3) that are used in all branches of industry. They are connected to each other, with each subsequent category upgrades on the previous one. The level of complexity and required resources increases as deeper analyses are conducted, but this leads to better and more accurate business insights.

Fig. 3. Main types of data analysis

Descriptive analytics

The first step in data analysis is descriptive analysis. It is the collection, organisation, analysis and summarisation of large data sets that are historical or observational. These sets may come from different sources, be heterogeneous, or presented in different formats. Through sets of procedures, key features are identified within them, with the idea that the data may reveal interesting patterns. Descriptive analysis allows for the discovery of hidden relationships in data, issues with data storage, and more questions that have yet to be answered. Its purpose is to eliminate complexity in understanding data and make it more representative. This is most often achieved through numerical or graphical summarisation, using graphics, tables or charts, which allows the business to more easily extract valuable information from them. The main questions that descriptive statistics answers are “What happened?” and “How much?”. Some of the approaches used to answer these questions are:

- How widely dispersed is the data?
- Are there values that repeat more often than others?
- What are the largest and smallest values?
- Are there any particular trends in the data?

However, the answers to these questions will simply alert that something has happened or is wrong, but do not provide an explanation as to why it happened. Therefore, the analysis should be more in-depth in order to find the cause of the phenomenon.

Diagnostic analytics

Diagnostic analysis builds on descriptive analysis. It works with the results obtained from descriptive analysis to answer the question "Why did the event happen?". It enables companies to gain insight into the causes of patterns that have been observed in the data. Its purpose is to explain why anomalies and deviations occur by formulating various hypotheses and testing them.

The critical aspect of performing a diagnostic analysis is to collect detailed data from the company. Sometimes it is also necessary to use additional sources to help draw a complete picture, after which the conclusions reached are tested against the original database.

Performing diagnostic analysis can include multiple approaches and techniques, such as regression analysis, clustering analysis, time-series analysis, probability theory, data drilling and data mining.

Predictive analytics

Predictive analytics is used to identify future trends in data based on historical data. It uses the findings of descriptive and diagnostic analysis to identify patterns and predict "what is likely to happen” with a certain degree of confidence. It belongs to the advanced types of analysis. The accuracy and correctness of the predictive analytics result is highly dependent on the quality and completeness of the data.

To perform a predictive analysis, methods and algorithms such as regression analysis, machine learning, and neural networks are used. Software products like SAS Enterprise Miner that have these methods built into them make it much easier to understand and use. (Watson and Hugh, 2014).

Prescriptive analytics

Prescriptive analytics builds on predictive analytics by not only predicting what will happen in the future, but making suggestions for optimal solutions for future phenomena in the data.

It works with the results achieved so far from previous analyses, prescribing what action is best taken to mitigate or even avoid future risks, or for the organisation to make the most of a promising trend. Prescriptive analysis recommends actions or decisions based on a complex set of objectives, constraints, and choices. Prescriptive analytics implementations may require a feedback system to track the outcome of actions taken.

Prescriptive analytics uses advanced tools and technologies – ML, AI and sophisticated algorithms to achieve the best results, however, making it complex to implement and manage. AI systems can be designed to continuously learn new data from the organisation and use that information to help the business make the best decisions.

Approaches for Big Data Analytics

According to a statistical study by Taylor (2022), an average of 328.77 million terabytes of data will be generated daily in 2023.

Most of this data could be analysed to extract additional value from it. They represent a huge volume of various types of data, including those that require real-time processing. Approaches to analysing big data need to be re-examined, as the data is often too complex, multidimensional, unstructured or incomplete to be handled by traditional approaches. This is where advanced analytics techniques applied to big data sets come into play. As much as these methods have evolved, however, the larger the volume of data, the more difficult it is to organise and manage it. Which methods will be used depends on the type of analysis to be done, the specific problem to be solved, and the type of data to be analysed (Ma et Al., 2014)

Data mining

Clifton (2023) provides a definition of Data mining, presenting the method as the process of discovering interesting and useful patterns and relationships in volumes of large data. Tools from the fields of statistics and artificial intelligence, such as neural networks or machine learning, are used. Combining them with database management and analysing large sets of digital data enables the analysis and the discovery of patterns in the data.

Machine Learning

Bangert (2021) separates the machine learning into three main parts. First, it consists of many prototypical models that could be applied to the data at hand. These are known by names such as neural networks, decision trees, or k-means clustering. Second, each of these comes with several prescriptions, so called algorithms that tell us how to calculate the model coefficients from a data set. This calculation is also called training the model. After training, the initial prototype has been turned into a model for the specific data set that we provided. Third, the finished model must be deployed so that it can be used. It is generally far easier and quicker to evaluate a model than to train a model. In fact, this is one of the primary features of machine learning that make it so attractive: Once trained, the model can be used in real-time. However, it needs to be embedded in the right infrastructure to unfold its potential.

There are two main types of machine learning - supervised learning, where the algorithm is trained using already classified and marked data, and unsupervised learning, which tries to find the classification indicators and mark them by itself.

Classification

Classification is essentially supervised machine learning. In order to perform a classification analysis, it is necessary to pre-determine the classification groups (classes) which will serve as labels to mark the data. The attributes of objects that define the essential characteristics for classification are called predictors. The purpose of this type of analysis is to create a generalised model of the data, whereby the change of the dependent attributes (their belonging to a certain class) is determined, which is related to the behavior of the predictor attributes (Ivanov, 2016).

Association

Association is unsupervised machine learning. This algorithm measures the strength of co-occurrence of variables in large databases. Its goal is to find hidden patterns and correlations that can be turned into easy-to-understand and recognisable rules (Kotu and Deshpande, 2019).

Cluster analysis

Cluster analysis is an unsupervised machine learning-based algorithm that operates on unlabelled data. It divides large databases into smaller groups based on their similar characteristics. These groups are called clusters. Before applying the analysis, it is not known how many clusters exist in the data. It is usually used when there are no assumptions about likely relationships in the data. It presents information about where associations and patterns exist in the data, but not what they might be and what they mean.

Regression Analysis

Regression analysis is a statistical technique of measuring the relationship between variables to gain a deeper understanding of particular trends or patterns. It is done by choosing one dependent variable and specifying others (independent variables) on which it should depend. It measures the extent to which the change in the value of the dependent variable depends on the change in the values of the independent variables. The purpose of this type of analysis is to filter out only the significant variables and exclude those that have no particular value.

Cases of BDA in Industry

The mining industry is characterised by dynamics of the processes and a variety of objects. Its impact on the environment is also significant (Kolev et al., 2019). Making informed and efficient decisions in such a situation requires high quality data. Anastasova and Yanev (2021) describe modern approaches to data storage and processing, as well as technologies to achieve the data quality required for specific purposes in the mining industry.

Arsova and Hristov (2021) propose an approach using Deep neural networks for the study of a geological indicator, comparing the results obtained with those achieved with Shallow Neural Networks and demonstrating how the number of hidden layers and the number of neurons in them affect the quality of data analysis from the exploration of ores.

The question of choosing software tools for data processing is also important. Drankov (2020) proposed a resource-constrained approach based on Python and XML configuration.

Fakete (2015) highlighted the critical factors of successfully utilising big data, IoT, predictive maintenance technologies to enhance production, decrease delay time, and boost business, in general, while he considers the current challenges of the industry with its low profitability and high expenditures. He has constructed a generic model that mining companies can use to incorporate these methods into their business.

Tyuleneva (2020) considers the benefits of implementing digital technologies in the mining industry. They are related to the optimisation of costs for materials, equipment, real-time monitoring of the implementation of production plans, prediction of future production risks and failures.

Tylečková and Noskievičová (2020) describe the challenges of providing and monitoring the data quality. Low-quality data does not lead to accurate information, resulting in unexpected outcomes that may damage trust in the information system. In contrast, the acquisition, storage, and utilisation of Big Data is expected to be a critical factor in staying competitive, promoting business growth, and fostering innovation. As a result, companies that regularly incorporate Big Data into their decision-making and strategic planning will have an advantage over their competitors.

Hyder et al. (2019) conducted and analysed a number of interviews with representatives of the mining industry regarding the possible uses, benefits, and challenges of implementing AI and ML in the mining industry. They also describe what they believe are the necessary steps for the implementation of such technologies in the sector, as well as their possible future uses.

Bag et al. (2022) build and apply a theoretical model that uses partial least squares structured equating modeling. They explore the direct effect of big data predictive analytics on supply chain and resource resilience in the South African mining industry under extreme weather events. For this purpose, they conducted interviews with 229 persons involved in the mining industry.

Yin et al. (2021) use local indicators of spatial association (LISA), principal component analysis (PCA), and deep autoencoder network (DAN) procedures to build a data-driven geochemical model that helps discover knowledge and associations in geochemical and mineral exploration in the Daqiao district, Gansu Province, China.

Conclusion

The field of information technology is developing at an incredible pace. IoT and ML are helping to automate many processes in industry that were previously done by humans, sometimes with a high risk of accidents. Data analytics systems and algorithms are becoming increasingly sophisticated. With their help, organisations can not only analyse their behavior, but also improve their efficiency, reduce costs, predict future risks or optimise their work.

Many of the processes are monitored by IoT devices and sensors which sometimes record huge amounts of data in real time. Their processing and analysis is impossible with traditional techniques and tools for data manipulation. This is where BDAs, which are specifically designed to handle large volumes of heterogeneous data as well as real-time analytics, come to the rescue.

Successfully selected BDA leads to positive business results, but there are also examples of unsuccessful implementations of such systems. Due to the wide variety of BDA systems, methods used by them and software tools for their implementation, choosing a suitable BDA is a complex and responsible process.

References

Anastasova Y., Yanev N. 2021. Models and data quality in information systems applicable in the mining industry, *E3S Web Conf. 280* 08012 DOI: 10.1051/e3sconf/202128008012

Arsova-Borisova, K., Hristov V. 2021. Deep neural networks applications in the study of a geological indicator – *Sustainable Extraction and Processing of Raw Materials (https://seprm.com)* Volume 2, November Issue, https://doi.org/10.58903/b15171913

Bag S., Rahman M., Srivastava G., Chan HL, Bryde D. 2022. The role of big data and predictive analytics in developing a resilient supply chain network in the South African mining industry against extreme weather events, International *Journal of Production Economics,* Volume 251, 108541, ISSN 0925-5273

Bangert P. 2021. Chapter 3 - Machine Learning, *Machine Learning and Data Science in the Oil and Gas Industry, Gulf Professional Publishing*. Pages 37-67. ISBN 9780128207147

Clifton, C. 2023. Data mining. *Encyclopedia Britannica.* Available on: https://www.britannica.com/technology/data-mining. Accessed 10 April 2023

Drankov, I., Gorbounov Y. 2020. Advanced Formatting of Delimited Big Data with Python, *16-th Annual Conference Computer Science and Education in Computer Science (CSECS)*. ISSN 1313-8624. At: Metropolitan College, Boston University, USA

Fekete J. 2015. “Big data in mining operations”. Master’s Thesis Copenhagen Business School

Hyder, Z., Siau, K., & Nah, F. 2019. Artificial Intelligence, Machine Learning, and Autonomous Technologies in Mining Industry. *Journal of Database Management (JDM),* 30(2), 67-79. http://doi.org/10.4018/JDM.2019040104

IDC Tech. Rep. 2021. Worldwide Big Data and Analytics Spending Guide. Available on: https://www.idc.com/getdoc.jsp?containerId=prUS48165721

Ivanov M. 2016. Syvremenni metodi na inteligenten analiz na danni. *Научен електронен архив на НБУ* (in Bulgarian)

Kolev, P. Savov, M. Vatzkitcheva, K. Velichkova, D. Dimitrov, B. Vladkova, Toncheva. 2019. The impact of outdoor mining activities on atmospheric air quality in nearby settlements, *JOURNAL of Mining and Geological Sciences – Sofia: Univ. of mining of geology "St. Ivan Rilski"*, Vol. 62, part 2, p. 45-49, ISSN 2682-9525; ISSN 2683-0027.

Kotu V., Deshpande B. 2019. Chapter 6 - Association Analysis. *Data Science (Second Edition) - Concepts and Practice*, Pages 199-220. ISBN 9780128147610.

Taylor P. 2022. Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025, https://www.statista.com/ statistics/ 871513/worldwide-data-created/

Tylečková E., Noskievičová D. 2020. The role of big data in industry 4.0 in mining industry in Serbia. *CzOTO, Volume 2*, Issue 1, pp.166-173

Tyuleneva T. 2020. Problems and Prospects of Regional Mining Industry Digitalisation. *E3S Web of Conferences 174, Vth International Innovative Mining Symposium.* https://doi.org/10.1051/e3sconf/202017404019

Watson, Hugh J. 2014. Tutorial: Big Data Analytics: Concepts, Technologies, and Applications. *Communications of the Association for Information Systems: Vol. 34*, Article 65.

Yin B., Zuo R., Xiong Y., Li Y., Yang W. 2021. Knowledge discovery of geochemical patterns from a data-driven perspective. *Journal of Geochemical Exploration, Volume 231*, 106872, ISSN 0375-6742.