TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

The art of Data Analysis: Review on Essential, Techniques and Methodologies

Dr. Shalini Lamba[a], Harsha Sahni [b] and Aanya Sharma [c]

[a,b,c] National Post Graduate College, Lucknow,

India

drshalinilamba@gmail.com, harshasahni2505@email,
aanyasharma.1103@email

KEYWORD

Data Analytics; Knowledge Discovery; Data Mining; Preprocessing; Transformation; Interpretation; Big Data; Computational Complexity; Information Security

ABSTRACT

This paper provides a comprehensive overview of the key processes and methodologies in data analytics, with a focus on knowledge discovery in databases (KDD). It explores the evolution of data analytics and its importance in today's data-driven world. The paper discusses the significance of each KDD operator, including selection, preprocessing, transformation, data mining, and interpretation/evaluation, in shaping the outcomes of data analytics processes. Through a detailed examination of these operations, the paper highlights the essential role of data analytics in gathering, analyzing, and presenting insights derived from data. Furthermore, the paper explores the challenges and advancements in big data analytics, emphasizing the importance of understanding computational complexities, information security, and computational methods for effective data analysis.

1. Introduction

In today's data-driven world, the ability to extract meaningful insights from vast and complex datasets is paramount. Data analytics, the systematic exploration and analysis of data, has emerged as a crucial tool in unlocking the value hidden within these datasets. By employing various processes and methodologies, data analytics enables organizations and researchers to uncover patterns, trends, and correlations that can drive informed decision-making and strategic insights.[1]

The evolution of data analytics has been marked by significant advancements in technology, particularly in the areas of data storage, processing power, and machine learning algorithms. These advancements have enabled the analysis of larger and more diverse datasets, leading to groundbreaking discoveries and innovations across industries.

At the core of data analytics lies the process of knowledge discovery in databases (KDD), which encompasses several key operations, including selection, preprocessing, transformation, data mining, and interpretation/evaluation. Each of these operations plays a critical role in shaping the outcomes of data analytics processes, from gathering and cleaning data to deriving actionable insights.

Despite its many benefits, data analytics also presents challenges, particularly in the areas of computational complexity, data security, and scalability. As datasets continue to grow in size and complexity, there is a need for

Corresponding Author: Dr. Shalini Lambaa, National Post Graduate College, Lucknow, India
Email: drshalinilamba@gmail.com

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

more efficient and scalable data analytics solutions to handle these challenges effectively.

2. Foundations of Data Analytics: Key Processes and Methodologies in Knowledge Discovery

In the field of data analytics, the process of knowledge discovery in databases (KDD) consists of several key operations, as outlined by Fayyad and colleagues. These operations include selection, preprocessing, transformation, data mining, and interpretation/evaluation. Together, these operations form the foundation of a

robust data analytics system, enabling the collection, analysis, and presentation of insights derived from data. While data mining often takes the spotlight in research and reports, it's essential to recognize the importance of all KDD operators. Each operator, from gathering and selecting data to preprocessing, transformation, and beyond, plays a crucial role in shaping the outcomes of KDD processes.[2]

The initial stages of data analysis involve input operations such as gathering, selection, preprocessing, and transformation. Selection focuses on identifying relevant data for analysis, while preprocessing detects and refines data to ensure its utility. Transformation then standardizes data into formats suitable for data mining. These preparatory steps, including data extraction, cleaning, integration, and reduction, are essential for refining raw data into actionable insights for subsequent analysis.[3]

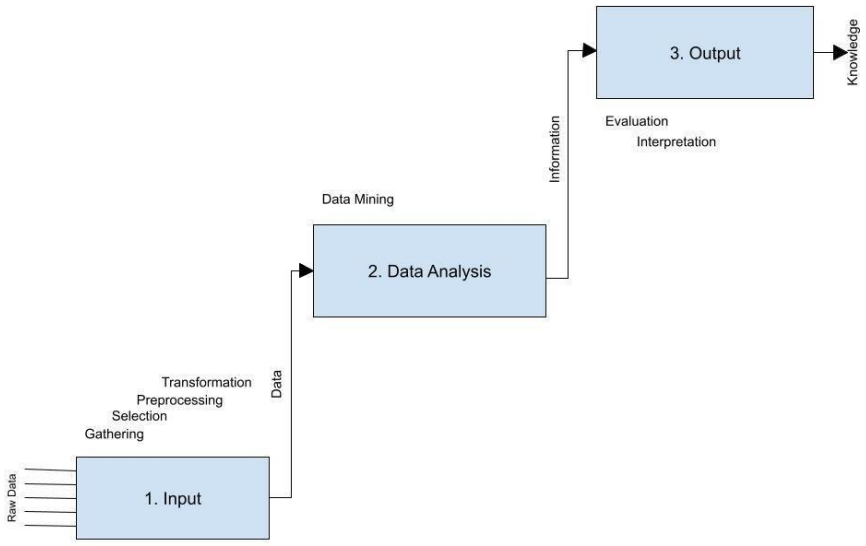

Fig1: The process of knowledge discovery in databases

At the core of data analysis is data mining, where hidden patterns, rules, and information are discovered from the data. While data mining includes various methodologies, such as statistical and machine learning approaches, its goal remains consistent: to extract valuable knowledge from raw data. Techniques like clustering, classification, association rules, and sequential patterns are used to uncover meaningful insights, often involving iterative processes like initialization, data scanning, rule construction, and update.

1. Input data D
2. Initialise candidate solutions r
3. **While** the termination criterion is not met
4. *d=Scan(D)*
5. *v=Construct (d, r,o)*
6. *r=Update(v)*
7. **End**
8. Output rules r

Outputting the results involves evaluation and interpretation, crucial for assessing the effectiveness of data mining outcomes. Evaluation metrics like precision, recall, F-measure, and computational costs offer quantitative insights into the quality and efficiency of data analysis. Interpretation focuses on translating data findings into

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

actionable information for decision-making. Graphical user interfaces, summarization techniques, and effective data visualization play pivotal roles in presenting complex data analysis results in a clear and understandable manner.

Incorporating these fundamental concepts and methodologies into data analytics frameworks ensures a systematic approach to extracting value from data, enabling informed decision-making and strategic insights.

3. Data Analysis and Data Preparation

Data analysis refers to the process of converting raw data into valuable insights that can drive decision-making. [4]It utilizes a variety of techniques and tools to identify patterns, trends, and relationships within the data, which are crucial for researchers and organizations seeking informed decisions and predictions.[5]

To begin data analysis, the data must first undergo preparation to ensure its quality and relevance. Data preparation involves several steps to clean, transform, and format the data into a usable format. These steps include:[6]

3.1. Data Cleaning: This step involves identifying and rectifying errors or inconsistencies in the data. Tasks typically include removing duplicate entries, correcting typos, and handling missing values.[7]

3.2. Data Transformation: Data often requires transformation to be suitable for analysis. This can involve converting data types (e.g., from text to numerical), normalizing data (e.g., scaling data to a standard range), or creating new variables based on existing ones.

3.3. Data Integration: Often, data is sourced from multiple locations and needs to be integrated into a single dataset. This process involves identifying common variables and merging data from different sources.

3.4. Data Reduction: In cases where datasets are excessively large or complex, data reduction techniques such as aggregation or sampling can be used to reduce the dataset's size while preserving its essential characteristics.[8]

3.5. Data Visualization: Visual representations of data, such as graphs and charts, can help researchers comprehend complex patterns and trends. Data visualization makes it easier to convey findings to others.

In summary, data analysis and preparation are critical components of the research process. They ensure that the data is accurate, reliable, and suitable for analysis, ultimately leading to more robust and meaningful results.

4. Data Analytics and its Methods

Data analysis encompasses a range of techniques for examining large datasets, including structured architecture, data mining, and analytical tools. A key aspect of analyzing big data involves exploring significant values, offering recommendations, and utilizing decision support tools like executive information systems and online analytical processing. To handle the complexity of data analysis, algorithms are used to process real-time data, yielding highly accurate results. Data analysis can reveal valuable insights for business growth, such as predictive analytics for future projections. The field of data analytics is vast and dynamic, given the diverse types of data and its rapid growth. The analysis's purpose varies depending on the application, aiming to address three general questions: past events, current situations, and future expectations.[9]

Extracting and processing information from large databases is time-consuming and resource-intensive. The interdisciplinary nature of data analytics makes it challenging for businesses to identify the specialized skills required for comprehensive data analysis. Effective research provides essential features for completing this task and overcoming the analytical challenges.[10]

Data analytics can be described as a data science that breaks down data into individual components for examination and integrates them to generate knowledge. Oracle and Cloudera have proposed a seven-step approach for extracting value using data analytics, including identifying objectives, business levers, data collection, data cleaning, data modeling, team building, and optimization.[11]

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

5. Types of Data Analysis

Data analysis can be categorized into six main methods:[12]

5.1. Descriptive Analysis: Descriptive analysis summarizes data for easy presentation. It includes two main categories: Univariate and Bivariate analyses.

Univariate Analysis: This set of statistical tools focuses on understanding the characteristics and general properties of a single variable. Common techniques include:[13]

- Frequency: Determines all possible values for a specific variable and the frequency of each value in the dataset, helping to understand the distribution of variables.
- Central Tendency: Identifies the most represented value in a dataset using methods like mean, mode, and median. Mean is the average of the values, median is the middle value, and mode is the most frequent value.
- Dispersion: Indicates how the tools like range, standard deviation and variance can spread out the values around the central tendency. Range is the difference between the highest and lowest values, variance displays the concentration of values around the average value, and standard deviation is square root of the variance which helps to understand the variability of data.

Bivariate Analysis: This involves analyzing the relationship between two variables. Common techniques include:

- Correlation: Measures the strength and direction of the relationship that exists between the two variables, using a particular formula based on sample mean values and standard deviations, and can be extended to analyse more than two variables.

Software like SPSS simplifies the computation of these measures, especially for complex formulas, enabling researchers to efficiently analyze and interpret data, leading to meaningful insights.

5.2. Exploratory Analysis: Exploratory analysis focuses on identifying influences and answering research questions related to relations, connections, and patterns between variables. The major techniques used are Dependence and Interdependence methods.

Dependence Techniques: These analyse the impact of predictor (variables) on an outcome variable. Common tools include:

- Analysis of Variance: Compares outcomes of multiple groups (predictor variables) and can be applied in inferential analysis.
- Multiple Analysis of Variance (MANOVA): Compares outcomes across two or more variables, often used in experimental research.
- Structural Equation Modeling (SEM): Analyses relationships between interrelated and predictor variables, defining structural relationships between latent and measured constructs, fruitful for testing hypotheses and estimating theoretical networks of relations.

- Logistic Regression: Similar to multiple regression but with dichotomous outcome variables and metric predictor variables.
- Multiple Discriminant Analysis: Handles multiple predictor variables and a single outcome variable, both dichotomous and multichotomous categories.

Interdependence Techniques: These analyze relationships between variables without assuming influence direction. Common techniques include:

- Factor Analysis: Reduces a large number of variables to a new variate or a smaller set of factors to discover patterns and relationships.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

- Cluster Analysis: Classifies objects or individuals into mutual groups based on homogeneity and heterogeneity between clusters.
- Multidimensional Scaling: Identifies key dimensions based on individuals' judgments and perceptions, transforming judgments and perceptions using distances represented in multidimensional space.

These techniques are valuable for analyzing complex relationships between variables and uncovering patterns that can inform decision-making and research conclusions.

5.3. Inferential Analysis: Inferential analysis bridges the gap between sample and population data, involving widely used statistical procedures, including:

- T-Test: Also known as Student's t-test, it compares means or averages between groups using a single dichotomous independent variable along with a continuous dependent variable. The t-test can be non-directional (two-tailed) or directional (one-tailed), used for hypothesis testing.
- Analysis of Variance (ANOVA): Compares means across different groups, similar to using multiple t-tests, more efficient and reduces experiment-wise error, using differences between mean values for comparison, handling more than two groups.
- Chi-Square ($\chi 2$): Tests relationships between categories of two categorical variables from the same population, used for nominal and ordinal data to compare observed and expected frequencies.
- Regression: Uses one or more independent variables for predicting a value on a dependent variable, similar to correlation but focusing on predicting the dependent variable, simple (single independent variable) or multiple (several independent variables).
- Time Series Analysis: Analyzes variables changing continuously over time, often used in longitudinal research designs, aiming to summarize data, fit low-dimensional models, and make predictions based on regular interval observations.

These inferential analysis techniques are valuable for understanding relationships, predicting outcomes, and making decisions based on sample data that can be generalized to a larger population.

5.4. Predictive Analysis: Predictive analysis uses historical data to make predictions about future events or trends, building predictive models using statistical algorithms and machine learning techniques to forecast future outcomes based on past data.

5.5. Explanatory or Causal Analysis: Explanatory analysis aims to understand the relationships between variables and determine if one variable causes changes in another, done through experiments or observational studies to establish causal relationships.

5.6. Mechanistic Analysis: Mechanistic analysis focuses on understanding the underlying mechanisms or processes driving observed phenomena, involving developing theoretical models based on existing knowledge and testing these models using data.

These methods of data analysis serve different purposes and are used in various contexts to extract meaningful insights from data, researchers choosing the most appropriate approach based on their research questions and data characteristics.

6. Advancements and Challenges in Big Data Analytics

In recent years, the accumulation of big data has become widespread across various fields, including healthcare, public administration, retail, biochemistry, and interdisciplinary scientific research. This surge in data is also prominent in web-based applications, such as social computing, internet text and document.[1] I, and internet search indexing. Social computing includes aspects like social network analysis, online communities, recommender systems, reputation systems, and prediction markets. Internet search indexing involves platforms like ISI, IEEE Xplorer, Scopus, and Thomson Reuters. Big data offers new opportunities for knowledge processing tasks among researchers, but it also comes with significant challenges.[14]

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

To address these challenges, understanding computational complexities, information security, and computational methods for big data analysis is necessary. For example, statistical methods that perform well with small data might be problematic to scale with large datasets. Similarly, many computational techniques encounter obstacles when dealing with large-scale data analysis. In the health sector, researchers have identified four primary challenges: data storage and data analysis, knowledge discovery and computational complexities, scalability and visualization of data, and information security.

6.1. Data Storage and Analysis: The exponential growth in data size, driven by mobile devices, aerial sensory technologies, remote sensing, and RFID readers, poses challenges in data storage mediums and input/output speeds. Storing immense data can be costly, leading to data being ignored or deleted due to space constraints. This underscores the need for efficient storage mediums that prioritize data accessibility for analysis. While technologies like solid-state drives (SSD) and phase-change memory (PCM) have been introduced to address these challenges, existing solutions still struggle to meet the performance demands for processing big data.

The diversity of data presents another challenge, requiring increased data mining tasks and the need for data reduction, selection, and feature selection, especially with large datasets. Automating these processes and developing new machine learning algorithms are essential for consistency and efficiency in data analysis. Clustering large datasets using technologies like Hadoop and MapReduce has become feasible, facilitating the collection and analysis of semi-structured and unstructured data within reasonable timeframes. However, there is still a need for a standard process for effectively analyzing and extracting knowledge from such data.[15]

6.2. Knowledge Discovery and Computational Complexities: Knowledge discovery and representation are important for big data analysis, covering sub-fields like authentication, management, preservation, information retrieval, archiving, and representation. Several tools and techniques, such as fuzzy set, near set, formal concept analysis, rough set, soft set, and principal component analysis, have been developed, but their effectiveness with large datasets is a concern. As data sizes continue to grow, existing tools might not effectively process these data to derive meaningful insights. Data warehouses and data marts are commonly used for large dataset management, with data warehouses storing data from operational systems and data marts facilitating analysis.[16]

Analyzing large datasets involves dealing with computational complexities, uncertainties, and knowledge exploration systems. Systematically modeling computational complexity is important, although establishing a comprehensive mathematical system applicable to big data remains challenging. Domain-specific data analytics, leveraging machine learning techniques with minimal memory requirements, have emerged to minimize computational costs and complexities. However, current tools often struggle to handle these complexities effectively, necessitating the development of more efficient and scalable machine learning algorithms for data analysis.

6.3. Bio-inspired Computing for Big Data Analytics: Bio-inspired computing offers promise for intelligent data analysis and application to big data. Techniques like fuzzy set, rough set, soft set, near set, formal concept analysis, and principal component analysis have been beneficial in processing large datasets. Hybridized techniques and domain-specific data analytics are being developed to optimize big data analysis and address computational complexities and uncertainties.

6.4. Quantum Computing for Big Data Analysis: Quantum computing presents a revolutionary approach to big data analysis, using quantum bits (qubits) and quantum mechanics to process information exponentially faster than traditional computers. Quantum computing has the potential to address complex big data problems efficiently, offering a pathway to analyse large datasets and derive meaningful insights.

In conclusion, advancements in bio-inspired computing and quantum computing present new possibilities for addressing the challenges and complexities in big data analytics. These advancements pave the way for enhanced data processing, knowledge discovery, and decision-making capabilities across various domains.

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

7. Future Scope

In our further study we aim to examine the eco-friendliness of various data analytic techniques. Data, though useful, is affecting the planet on a great scale. Using better and efficient techniques for analysing data and reducing it to the smallest possible chunk of useful information is a task which must be handled efficiently. The complexities of data analysis techniques must be considered to get an overall better way of handling huge data. Thus, different data analysis techniques should be examined thoroughly to get proper insights of their impact on our planet.

8. References

[1] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *Int J InfManage*, vol. 35, no. 2, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[2] K. Kelley, "What is Data Analysis? Methods, Process and Types Explained," *Simplilearn - Online Certification Training Course Provider*, 2023.

[3] D. Jhonshon, "What is Data Analysis? Research | Types | Methods | Techniques," *Guru99.Com*. 2021.

[4] V. Çetin and O. Yıldız, "A comprehensive review on data preprocessing techniques in data analysis," *Pamukkale University Journal of Engineering Sciences*, vol. 28, no. 2, 2022, doi: 10.5505/pajes.2021.62687.

[5] C. W. Tsai, C. F. Lai, H. C. Chao, and A. V. Vasilakos, "Big data analytics: a survey," *J Big Data*, vol. 2, no. 1, 2015, doi: 10.1186/s40537-015-0030-3.

[6] D. Fife, "The Eight Steps of Data Analysis: A Graphical Framework to Promote Sound Statistical Analysis," *Perspectives on Psychological Science*, vol. 15, no. 4, 2020, doi: 10.1177/1745691620917333.

[7] S. R. Durugkar, R. Raja, K. K. Nagwanshi, and S. Kumar, "Introduction to data mining," *Data Mining and Machine Learning Applications*. 2022. doi: 10.1002/9781119792529.ch1.

[8] H. Masood, S. Qadri, R. Sabah Scholar, G. University, P. Masood Hassan, and S. Salman Qadri PhDScholar, "Research Process and Steps Involved in Data Analysis," *Journal of Xidian UniversityVOLUME*, vol. 16, no. 3, 2022.

[9] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," *Journal of King Saud University - Computer and Information Sciences*, vol. 30, no. 4. 2018. doi: 10.1016/j.jksuci.2017.06.001.

[10] H. Taherdoost, "Different Types of Data Analysis Data Analysis Methods and Techniques in Research Projects," *International Journal of Academic Research in Management (IJARM)*, vol. 9, no. 1, 2020.

[11] Prateek Bihani and S. T. Patil, "A Comparative Study of Data Analysis Techniques," *International Journal of Emerging Trends and Technology in Computer Science*, vol. 3, no. 2, 2014.

[12] M. Islam, "Data Analysis: Types, Process, Methods, Techniques and Tools," *International Journal on Data Science and Technology*, vol. 6, no. 1, 2020, doi: 10.11648/j.ijdst.20200601.12.

[13] J. C. O'Neill, "Statistical Techniques for Data Analysis," *Technometrics*, vol. 47, no. 3, 2005, doi:

TEJAS Journal of Technologies and Humanitarian Science
ISSN-2583-5599
Vol.03, I.02 (2024)
https://www.tejasjournals.com/

10.1198/tech.2005.s301.

[14] R. H. Hariri, E. M. Fredericks, and K. M. Bowers, "Uncertainty in big data analytics: survey, opportunities, and challenges," *J Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0206-3.

[15] J. Fan, F. Han, and H. Liu, "Challenges of Big Data analysis," *National Science Review*, vol. 1, no. 2. 2014. doi: 10.1093/nsr/nwt032.

[16] F. Amalina *et al.*, "Blending Big Data Analytics: Review on Challenges and a Recent Study," *IEEE Access*, vol. 8. 2020. doi: 10.1109/ACCESS.2019.2923270.