Methods and Algorithms for Data Analysis and Knowledge Discovery from Data

Violeta Todorova

***Abstract* – The rapid development of information technology has led to the huge amount of information generated by large or complex systems. Applications in the field of information technology, telecommunications, business, robotics, economics, medicine, and many other fields generate information volumes that challenge professional analysts. Data mining analysis finds application in areas where statistical and analytical methods and the models built through them are not enough. Data mining analysis is suitable for areas where heterogeneous, large data predominates.**

***Index Terms* – data analysis, data mining, knowledge discovery, medical data sources.**

I. Introduction

Nowadays huge amounts of data were generated as result of computer simulations. Medicine is fundamental fields highly dependent on big data technologies. This stimulates the progress of data processing technologies and methods.

Big data technology allows large groups of biological specimens to be collected and the data can be stored, managed and analyzed [1].

Machine learning algorithms can generate additional output data that may differ from the original input data, thus creating knowledge from big data.

Rapid Learning Healthcare (RLHC) models using artificial intelligence can detect data of varying quality that must be compared to validated data sets to be truly meaningful. The extracted information can then be processed into decision support systems (DSS) to put knowledge-based healthcare into practice.

A. Sources of Medical Data

Sources of big data in healthcare include EHRs, smart devices, genetic databases, government, and more. (Fig. 1).

1) Internet of Things (IoT):

- Wearable devices
- Smartphone applications
- Medical devices and sensors

2) Electronic Medical Records/Electronic Health Records (EMR/EHR).

3) Other clinical data.

Violeta Todorova is with the Theoretical Electrotechnics Department, Faculty of Automation, Technical University of Sofia, 1000 Sofia, Bulgaria, (e-mail: violetatodorova@yahoo.com).

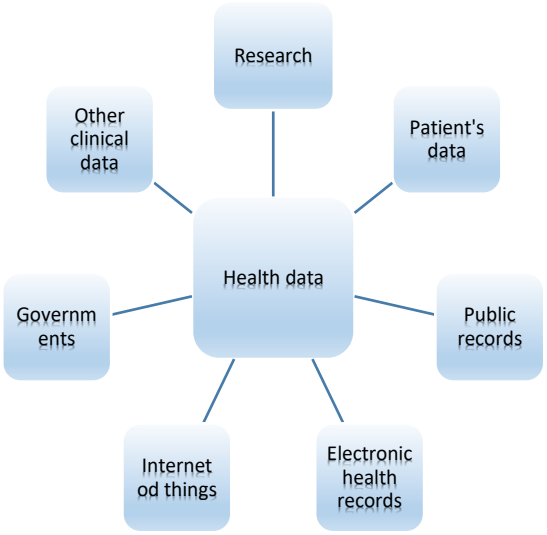

Fig. 1. Sources data for healthcare

Next generation of data analysis methods must manage huge amounts of data from different sources types with differentiated characteristics, confidence levels and frequency of actualization [2]. Data analysis aims to gain knowledge in an efficient manner.

Knowledge gained must meet the following requirements: to be accurate, understandable, and useful. Data collection tasks include classification, dependency modeling, grouping, function retrieval, and associative rule discovery [2].

Knowledge discovery from data is focused on methods for extracting useful knowledge from data. The growing volume of data and the widespread use of databases lead to challenges to innovative methodologies for discovering data for knowledge. Data collection skills are based of research in the field of statistics, databases, modeling, machine learning, data visualization to achieve automated complex and intelligent solutions [2].

B. Big Data Analytics in Medicine Areas

Image processing. Medical images are an important source of data, often used for diagnosis, therapy evaluation and planning [3].

Medical imaging data can range from a few megabytes per study to hundreds of megabytes per study. Such data requires a large storage capacity. Also, fast and accurate algorithms are required if automated processing is to be performed to make decisions using the data.

https://doi.org/10.47978/TUS.2022.72.03.002

Signal processing. Medical signals also present volume and speed hurdles, especially during continuous high-resolution acquisition and storage from multiple monitors connected to each patient.

Genomics. The cost of sequencing the human genome is rapidly decreasing with the development of high-throughput sequencing technology [3]. Analyzing genome-scale data for timely recommendation development is a significant challenge in the field of computational biology.

C. Big Data Use Cases and Data Analytics in Healthcare

The insights gained from big data analytics provide healthcare professionals with insights not previously available [4, 5]. Big data in healthcare is applied at every step of the healthcare cycle: from medical research to patient experience and outcome. Real-world applications that demonstrate how an analytics approach can improve processes, improve patient care, and ultimately save lives.

1) Diagnostics
2) Modeling and forecasts
3) Real-time monitoring of the patient's vital signs
4) Treatment of serious diseases
5) Population health
6) Preventive care
7) Electronic Health Records (EHRs)
8) Telemedicine
9) Real-time alerts
10) Data integration with medical images

II. KNOWLEDGE DISCOVERY BASED ON DATA ANALYSIS

The objectives of knowledge discovery based on data analysis are predictive and descriptive as follows [2].

- Prediction involves the use of certain variables or fields in the database to predict unknown or future values of other interesting variables.
- The description focuses on finding models that describe the data.

Prediction and description purposes are achieved by using the following data retrieval tasks [2]:

- Classification is a function that classifies a data element into one of several predefined classes.
- Regression is a function that classifies a data element of a variable with a real value.
- Grouping is a general descriptive task that identifies a limited set of categories or clusters to describe data.
- Relationship modeling is finding a model that describes significant relationships between variables.
- The change and the detection of deviations are aimed at detecting the most significant changes in the data from previous measured or normative values.

The process of knowledge discovery and data extraction (Fig. 2) consists of six main stages [5-7]:

Fig. 2. Knowledge discovery and results on the basis of data analysis and results interpretation

1. Understanding the problem area - this is the initial stage, which focuses on defining research objectives and relevant requirements from the user's point of view. Once this stage is completed, the knowledge should be translated into definitions of data retrieval tasks and a preliminary plan for how these goals can be achieved.
2. Data comprehension - begins with the initial data collection and continues with activities aimed at deepening the knowledge regarding the nature of the data. At this stage, it is necessary to identify problems related to data quality, to find the appropriate subsets of data in order to form initial hypotheses about hidden knowledge.
3. Data preparation - covers all activities for creating raw data from the final set of raw data. The data preparation stage often has to be repeated many times at different stages of the computational workflow. Data preparation tasks include data selection, defining attributes, examining individual records, and transforming and clearing data.
4. Modeling - this stage consists in selecting and applying various modeling techniques to extract data dependencies. The parameters of the model are adapted to their optimal values. As some models have their own specific data format requirements, it is often necessary to return to the data preparation stage.
5. Model evaluation - consists of a careful review of all steps taken in building the model to ensure that they achieve the specific objectives. At the end of this stage, a decision is made to use the results.
6. Use of the model - related to the applied strategy for monitoring and operation. At this stage, it must be determined whether and when to resume the data retrieval procedure and under what conditions.

As in many cases the data are imperfect, containing inconsistencies and abbreviations, they cannot be directly applicable to start the data retrieval process. The rapid increase in the rate of generation must also be taken into account of data and their size in various academic and scientific applications.

Most of the data collected require more complex analysis mechanisms. Data preprocessing aims to adapt the data to the requirements set by each data retrieval

algorithm, which allows data to be processed that would otherwise be inappropriate.

The data analysis conceptual model is shown on Fig. 3.

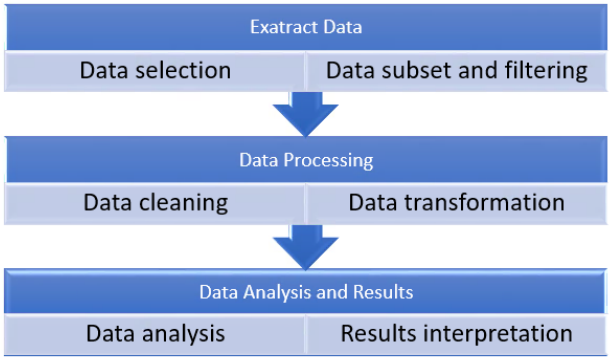

Fig. 3. Conceptual model for data analysis

The overall process of extracting and interpreting data models involves repetition of the following steps [8-11]:

- Defining goal of knowledge discovery process - defining task and corresponding prior knowledge and its application.
- Defining scope, appropriate end-user knowledge and goals.
- Creating a target dataset: choosing a dataset or selecting subset of variables or data instances.
- Filtering and preparation: removal of redundant or negative values; collecting the necessary modeling information; missing data fields processsing.
- Data set simplifying by deleting unwanted variables: finding suitable data presentation futures in relation to the task purpose; applying measurement or conversion methods to reduce the effective number of variables considered.
- Combining the objectives of the data discovery process with data extraction methods – determining whether the purpose of the knowledge-based process is classification, regression, etc.
- Selecting an algorithm for data extraction. This process includes appropriated models and parameters for overall process: selecting the method for searching a model in the data; determining appropriated models and parameters; compliance of a method for data extraction with the general criteria.
- Data extraction - searching interesting models as classification rules or trees, regression, clustering, etc.
- Interpretation of basic knowledge of derived models.
- Using knowledge and integrating it into another system for further action.

III. Workflow for Knowledge Discovery from Medical Data

The research techniques that follow the method of the discovery of knowledge from a collection of data includes the following:

- Data preparation, cleaning and selection;
- Knowledge discovery and decision making;
- Visualization and interpretation of results.

Data preprocessing covers integration of data from different sources, data clearing, selection of important functions. The data discovery workflow is shown on Fig. 4.

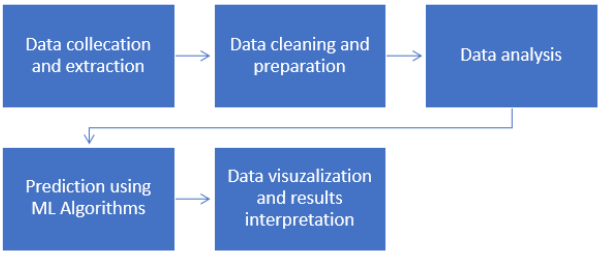

Fig. 4. Workflow for knowledge discovery from data

Data mining is the analysis of data from different perspectives and their summarization into useful information [12-14]. Data mining is the process of discovering hidden connections, correlations or patterns between dozens of records in large databases. Data mining tools use artificial intelligence, statistics, databases and machine learning systems to find the connection between data.

Important features of data mining are:

- Uses automated pattern detection.
- Predicts the expected results.
- Focuses on large arrays and databases
- Creates useful information.

Challenges related to big data analysis:

- Need for synchronization in different data sources
- Obtaining quality data through big data analysis
- Consolidation of a large amount of data in one platform
- Uncertainty in the field of data management
- Data storage and quality
- Data security and confidentiality
- The amount of data that is collected - automatic collection and organization
- Real-time data collection
- Visual presentation of the data
- Data from multiple sources
- Unavailable data
- Poor quality data
- Data scaling

A. Data preparation

An important step in reviewing data is to identify data quality issues. This can be done by checking for missing values, inconsistent data records.

Good data preparation is crucial for the creation of valid and reliable models that have high accuracy and efficiency. The accuracy of each analytical model strongly depends on the quality of the data submitted in it. Excellent data leads to more useful information that improves organizational decision-making and improves overall operational efficiency.

The data preparation process includes data cleaning, data integration, data selection and data transformation. The second phase involves extracting information from data, evaluating models and presenting knowledge.

B. Model tracking

Model tracking is a basic technique for data retrieval. It involves identifying and monitoring data trends or patterns to draw intelligent conclusions.

C. Classification

Classification is a function that assigns elements from a collection to target categories or classes. The purpose of the classification is to predict exactly the target class for each case in the data. For example, a classification model can be used to identify loan applicants as low, medium, or high credit

The classification task begins with a set of data in which the class assignments are known. Classification models are tested by comparing predicted values with known target values in a set of test data. Historical data for a classification project is usually divided into two data sets: one for model construction; another to test the model.

Data classification is a form of analysis that builds a model that describes important class variables. Classification methods are used in machine learning and pattern recognition.

Process for building a classification model:

- First step (training): A classification model is built based on the training data.
- In the second step (classification) the accuracy of the model is checked and then the model is used to classify new data.

D. Association

Association is a function that detects the likelihood of elements in a collection appearing together. The links between the accompanying elements are expressed as rules of association.

E. Detection of large differences in values

The detection of large differences in values determines all anomalies in the data sets. Once data deviations are detected, it becomes easier to understand why these anomalies occur.

F. Clustering

Analysis technique that relies on visual approaches - uses graphs to distribute data across different types of indicators. Determines the inherent grouping between the available data.

G. Regression

This is a technique that clearly reveals how the variables are related. Regression techniques are used in aspects of forecasting and data modeling. Regression and classification are techniques for extracting knowledge from data used to solve such problems. Both are used in predictive analysis, but regression is used to predict a numerical or continuous value, while classification divides the data into discrete categories.

More advanced techniques, such as multiple regression, predict a relationship between multiple variables. Adding more variables significantly increases the complexity of the forecast.

H. Forecast analysis

Forecasting analysis uses models found in current or historical data to extend them in the future. In this way it gives an idea of what trends will occur further in the data. Approaches include aspects of machine learning and artificial intelligence.

I. Sequential models

This technique for extracting knowledge from data focuses on revealing a series of events that take place sequentially. Especially useful for retrieving transaction data.

J. Decision trees

The decision tree is used to build classification and regression models. Used to create data models built from the training database submitted to the system (controlled training). With the help of a decision tree, solutions can be visualized.

K. Statistical techniques

Statistical techniques are the basis of most analyzes involved in the process of extracting knowledge from data. The various analytical models are based on statistical concepts that derive numerical values applicable to specific business purposes. Models for some statistical techniques are static, while others involving machine learning improve over time. Statistics include planning, design, data collection, analysis, preparation and interpretation and reporting of research results.

IV. CONCLUSION

Models and algorithms from machine learning, data extraction, statistical visualization, computational statistics, and other computer-intensive statistical methods are designed to learn from these complex volumes of information. These tools are often used to increase the efficiency and productivity of large and complex systems. This naturally makes these models, methods, and algorithms increasingly popular in all areas of economics and life.

Statistics is a major pillar of machine learning. Without it, a deep understanding and application of machine learning cannot be developed. Modern machine learning methods use the power of statistics to build efficient models, make reliable predictions and look for optimal solutions.

REFERENCES

[1] Mallappallil M, Sabu J, Gruessner A, Salifu M. A review of big data and medical research. SAGE Open Med. 2020 Jun 25;8:2050312120934839. doi: 10.1177/2050312120934839. PMID: 32637104; PMCID: PMC7323266.

[2] P. Borovska, V. Gancheva and I. Georgiev, "Platform for adaptive knowledge discovery and decision making based on big genomics data analytics," In: Rojas I., Valenzuela O., Rojas F., Ortuño F. (eds) Bioinformatics and Biomedical Engineering. IWBBIO 2019, Lecture Notes in Computer Science, vol. 11466. Springer, Cham, pp. 297–308, https://doi.org/10.1007/978-3-030-17935-9_27.

[3] Belle A, Thiagarajan R, Soroushmehr SM, Navidi F, Beard DA, Najarian K. Big Data Analytics in Healthcare. Biomed Res Int. 2015;2015:370194. doi: 10.1155/2015/370194. Epub 2015 Jul 2. PMID: 26229957; PMCID: PMC4503556.

[4] Esfandiari, Nura, Mohammad Reza Babavalian, Amir-Masoud Eftekhari Moghadam, and Vahid Kashani Tabar. "Knowledge discovery in medicine: Current issue and future trend." Expert Systems with Applications 41, no. 9 (2014): 4434-4463.

[5] Parihar, Astha, and Shweta Sharma. "Knowledge Discovery and Data Mining Healthcare." (2020).

[6] Cios, Krzysztof J., Witold Pedrycz, and Roman W. Swiniarski. Data mining methods for knowledge discovery. Vol. 458. Springer Science & Business Media, 2012.

[7] Mariscal, Gonzalo, Oscar Marban, and Covadonga Fernandez. "A survey of data mining and knowledge discovery process models and methodologies." The Knowledge Engineering Review 25.2 (2010): 137-166.

[8] Janiesch, Christian, Patrick Zschech, and Kai Heinrich. "Machine learning and deep learning." Electronic Markets 31.3 (2021): 685-695.

[9] Sun, Shiliang, et al. "A survey of optimization methods from a machine learning perspective." IEEE transactions on cybernetics 50.8 (2019): 3668-3681.

[10] Sarker, Iqbal H. "Machine learning: Algorithms, real-world applications and research directions." SN Computer Science 2.3 (2021): 1-21.

[11] Mahesh, Batta. "Machine learning algorithms - a review." International Journal of Science and Research (IJSR).[Internet] 9 (2020): 381-386.

[12] Hlaing, Khin Sein, and Y. M. K. K. Thaw. "Applications, Techniques and Trends of Data Mining and Knowledge Discovery Database." Int. J. Trend Sci. Res. Dev 3.5 (2019): 1604-1606.

[13] Chand, Satish. "Knowledge Discovery and Data Mining for Intelligent Business Solutions." Advances in Data and Information Sciences. Springer, Singapore, 2022. 205-214.

[14] Pareek, Mayank, and Purushottam Bhari. "A review report on knowledge discovery in databases and various techniques of data mining." Open Access International Journal of Science and Engineering 5.12 (2020): 79-82.