International Journal of New Findings in Health and Educational Sciences (IJHES)
ISSN 2980-0641
DOI : 10.63053/ijhes.6

Data Mining (Concepts, Algorithms) and Its Application to predict and Control Covid-19 Epidemic

Elham Khosravipour[1], Bahman Khosravipour[2]

1.M.A in Human Resource Management, Tehran, Iran
2. Professor in Agricultural Extension And Education, Tehran,Iran

ARTICLE INFO

Keywords: *Data Mining, Data Mining Algorithms, Health field, Covid-19 Disease*

ABSTRACT

In recent years, the issue of data mining has been considered by researchers due to the widespread access to large amount of data, the urgent need for information, proper, fast and accurate identification and diagnosis of various topics. Data mining discovers valid, fresh and useful patterns of available data which can provide valuable analytics for very large data sets. Today, the scope of its use has expanded and various data mining algorithms are used to identify and predict important issues in the field of health. The health sector is in the greatest need for data mining. The move from traditional medicine to evidence-based medicine is one of the things that can confirm this. In today's world, Covid-19 quickly hit other countries in addition to China like a hurricane, killing many people around the world. Due to the large amount of data in the field of medicine, the data mining process has played a very effective role in the management of various diseases such as prognosis, diagnosis and treatment. The main purpose of using data mining algorithms in medical sciences is to make better use of databases and discover hidden knowledge in order to help physicians make better decisions. The present article, which is prepared in a review manner using internet and library resources, tries to examine the concepts, data mining algorithms and their application in the field of health, especially to help diagnosing Covid-19. Finally, practical suggestions are provided.

1. Introduction

During the Third Industrial Revolution in the 1960s, humanity underwent a social transformation that led to the emergence of the information society. It was a new form of social existence whose primary function was to collect, store, analyze, and share network information. The development of technology has made it possible to transform from a service-oriented society to a human-centred technology, and with the use of the Internet of Things and big data, various industries and the human social environment have entered the information process. Information has created the Cyber-physical environment (CPE) and big data, enabling the information community to connect intangible assets as information networks. Since the amount of data collected has long exceeded the human ability to discover useful information, new techniques and tools must be developed to discover new knowledge from this data (Petrovik, 2021). Data is like a hidden treasure in any business and many businesses do not appreciate this information, while having this data alone may not be useful for a business and only It is the correct use of this data that helps a lot in business growth [2].

1. Data mining concept:

Perhaps the most important aspect in introducing data mining is the issue of knowledge discovery from databases (KDD) so that in many cases DM and KDD are used synonymously. The concept of data mining was first introduced in a workshop on KDD by Shaper. Subsequently, from 1991 to 1994, KDD introduced new concepts in this branch of science so that many sciences and concepts were related to it (Hamrahmansoft, 2017).

Data mining is a computational process that is used in many fields. Its purpose is to gain useful and hidden predictive knowledge. Data mining techniques are used to build a model in which new information can be identified with unknown information. One of the most popular features of all DM techniques is automatic learning, which detects its patterns in the observed data set. (Awadh et al, 2021). Over the past decade, vast amount of data has been collected in databases that most of them are obtained from business software, financial applications, resources and business management and business relations. The result of this large data collection can be that organizations and business units, services, etc. have rich information but poor knowledge in this regard and the main purpose of data mining is to extract patterns in this data set and turn them into "cognition". Data mining or knowledge discovery in databases is a powerful and technical tool that is used to extract potentially hidden knowledge and useful previous information from a set of data. This process automatically discovers the relationships and patterns in the raw data and executes the results. According to the definitions proposed from different perspectives, two basic components can be identified in data mining, the first is the discovery of hidden patterns in the data and the second is the use of these patterns to predict future results (Moradi and Ghasemi, 2012).

2. Data mining development process:

Table 1- Data mining development

Era	Business question	Technology	Manufacturers	Features
Data collection(1970)	What is the average profit of	Computers, tapes, disks	CDC-IBM	Delivery of static and expired data

	the company in the last 5 years?			
Data access(1980)	How much did the New England unit sales amount last March?	Relationship databases like SQL	Oracle-sybase-IBM-Microsoft	Delivery of outdated dynamic data
Data management(1990)	How much did the New England unit sales amount last March, and what are the details of Boston?	OLAP- Data warehouse, multidimensional database	Arbor –IRI-Pilot-Redbrick-Evolutionarytech	Delivery of outdated dynamic information on several levels
Data mining (2000)	How much will Boston City sales amount next month? Why?	Advanced algorithms, huge databases	Lockheed-IBM-SGL-Nascent industry	Delivery of prospective and forward-looking information

Source: (Tarokh and Sharifian, 2007)

3. Discover knowledge in data mining and basic data mining algorithms:

Data mining is closely related to artificial intelligence and machine learning. In data mining, it can be said that database theories combine artificial intelligence, machine learning, and statistics to provide a practical context. It should be noted that the term data mining is used when dealing with large amount of data in the mega or terabyte range. This has been emphasized in all data mining sources. The larger the data amount and the more complex the relationships between them, the more difficult it becomes to access the information hidden in the data, and the clearer the role of data mining as a method of knowledge discovery. (hamrahansoft, 2017).

The term knowledge discovery in a database is different from data mining. Knowledge discovery refers to the general process of discovering useful knowledge from data, dealing with the possible and existing interpretation of patterns to decide which conditions are known. But data mining is the application of algorithms to patterns extracted from data without the additional steps of the knowledge discovery process. Data mining algorithms on databases are useful for modeling or for discovering patterns in data. When these patterns are useful, new and understandable, we say that this is the discovery of knowledge. Data mining is in fact a step in a broader field of knowledge discovery process (Khosravipour, 2009).

Data mining algorithm is based on big data, through the establishment of mathematical model to analyze, classify and predict the data. The data processed by the algorithm can provide reference for the decision-making of organization management. The main functions of data mining algorithms include concept description, association rule analysis, classification and prediction, clustering analysis, anomaly analysis and trend analysis, evolution analysis, etc. Among them, **classification and prediction** are to classify similar data and predict some trends according to the characteristics of different types of data. Cluster analysis and classification have similarities, but they are different in essence. Cluster analysis is to aggregate data of the same category, and this category is usually known. Anomaly analysis is mainly used in intrusion detection, abnormal data discovery and so on. (Hai, 2021).

4. Applications of data mining in the field of health:

Data mining has many applications in the field of health, including the diagnosis of diseases,

classification of patients in disease management, finding patterns for faster diagnosis of patients and prevention of complications in them. (Baskabadi and Dostparast, 2020). The health sector is in the greatest need for data mining. The move from traditional medicine to evidence-based medicine is one of the things that can confirm this. (moghadasi et al, 2012).

Nowadays, effective use and extraction of hidden data from mass data in health management is considered by managers as a major goal to improve the current situation and the use of data mining techniques can be useful as an operational tool in identifying infected areas as well as evaluation and monitoring for scientific decisions of planners and managers (Hashemi Foumani and Matieian, 2020). One of the fields of application of data mining is its use in hospitals and pharmaceutical factories to discover unknown patterns and models of the effect of drugs on different diseases as well as patients of different age groups. (Moradi and Ghasemi, 2012). The modern world of medicine is rich in information but poor in knowledge. Therefore, striving to new pandemic and possible future pandemics has become one of the notable concerns of scientists. In the last decades, some valuable studies have been published regarding pandemics and data mining techniques. Such studies were conducted with the aim of better understanding, controlling, and manage pandemics using various data mining methods. Due to the importance to fight the COVID-19 pandemic, conducting a survey on the most popular and efficient data mining methods could have a significant impact on selecting the most effective techniques in pandemic studies. Thus, it can help us to reveal the unknown character of the new pandemic and the next possible pandemics. (safdari et al, 2021).

Extracting information from large amount of information using the data mining process helps us identify and predict disease. In data mining, various statistical methods, machine learning, artificial intelligence, etc. are used to discover hidden relationships between data (Mirzakhani et al., 2018).

5. Using data mining algorithms to predict and control Covid-19:

In late December 2019, the city of Wuhan in China reported an outbreak of a virus of unknown cause, known as Covid-19, which was approved by the World Health Organization. According to the World Health Organization, Covid-19 quickly hit other countries in addition to China like a hurricane, killing many people around the world. Prolonged latency and severity of the disease in the early stages led to a rapid increase in the number of patients. Infection of this severe respiratory disease is mild to moderate. The mechanism of this virus is unknown and so far there is no specific drug for this virus; Therefore, control of the source of infection is very important at present. Interruption of transmission and use of drugs and equipment is very important to control the progression of the disease (Ghasemi Tabagh et al., 2020).

In an epidemic, a country's health care system tolerates tremendous pressure due to the increase in the use of healthcare services and surge in hospitalizations. However, most COVID-19 patients usually are asymptomatic or have minor symptoms and can be advised to self-quarantine and get better under ambulatory or virtual care services. For severe or advanced ill patients who progressed to fast deterioration, instant hospitalization is of great significance to receive early interventions and supportive treatments that may increase the patient's survival chance. Furthermore, this pandemic has led to the shortage of hospital resources and the overtiredness of healthcare workers, which demands accurate forecast models to successfully triage hospitalized patients with poor prognoses and make the best use of restricted resources. Thus, using an Artificial Intelligence (AI)-based risk assessment tool is valuable to mitigate the burden of health systems from unnecessary hospital visits, charges, and mental and physical pressure of the health workers especially in countries with intensive medical resources shortages. Machine Learning (ML) is a sub-form of AI that provides new insight or knowledge via extract functional patterns and applicable models from large amount of the raw dataset. ML is a valued solution that is ever more deployed in clinical researches to conduct deep analyses and make

known new contributing factors of a specific target outcome. ML algorithms consist of supervised and unsupervised methods, which we considered supervised methods. In the supervised approach, we use part of our data as training data set to develop our model, and then we test the model with a section of data that is new to the algorithm. Accordingly, applying ML-based prediction models may aid decision-making by generating rapid and reliable predictions to determine the mortality risk of COVID-19 patients and effectively triage them. It can be beneficial to reduce the overwhelmed burden on healthcare systems by helping to predict the risk of deterioration and possible deaths. (Shanbehzadeh et al, 2021).

During the fierce outbreaks, not only clinical specialists have been trying to invent novel treatments and vaccines, but also scientists in the field of data science and technology are trying to discover the infectious and help control it by applying information-based methods. (Safdari et al, 2021). Due to the large amount of data in the field of medicine, the data mining process has played a very effective role in the management of various diseases such as prognosis, diagnosis and treatment. The results of a review study conducted by Albahri et al. on the role of data mining and artificial intelligence in the detection and diagnosis of Covid-19 disease showed that the use of this technology to provide diagnostic models and identify the most optimal and effective data mining algorithm can significantly help in timely, effective and economical diagnosis of the disease (Nopour et al., 2021).

The data mining algorithms used in the diagnosis and prediction of this disease are as follows:

- **NB[1] Classification algorithm**: NB algorithm is one of the most popular machine learning methods, specifically data analysis and classification. This method relies on the statistical concept of Bayes' theorem, which calculates the probability of a specific result by using available information. This classifier is called naive because it relies on the principle of independence assumptions, that is, the relationship among all attributes and features is considered independent of one another. The NB classifier model is characterized by the ease of construction and development and the ability to process large data, outperforming a number of sophisticated algorithms. The model is trained with the data and its available properties in databases, determines the type of new records and then classifies them on the basis of data and statistics previously available. It is used in many systems, such as for identifying harmful messages, classifying documents, such as in news sites, to anticipate the type of document (e.g. politics, sports and technology), recognizing the views and feelings in the text content (negative, positive or optimistic) and in face recognition in pictures.
- **Multilayer Perceptron Neural Network Algorithm:** MLP algorithm is one of the most popular neural networks algorithms. It consists of a perceptron, an input layer for receiving the signal, an output layer for making a decision or prediction about the data and an infinite number of hidden layers, which are MLP's true computational engine between the two layers.
- **Decision tree classification algorithm, j48:** This classifier is one of the most widely used machine learning language processing domains. The main advantages of this algorithm are easy construction of graphical classification and low-cost formal generation. However, this algorithm does not generate multiple redundant attributes and modules and it is quite susceptible to noise in the data. (Awadh et al, 2021). Also in the research conducted by

Naive Bayes[1]

Nopour et al. (2021), the results of comparing the performance of four famous data mining algorithms using different evaluation criteria showed that the use of selected data mining methods and in particular the decision tree classification algorithm j48 is highly capable in the timely and effective diagnosis of Covid-19 disease in the form of clinical decision support systems (Nopour et al., 2021).

The COVID-19 datasets provided by Johns Hopkins University, contain information on COVID-19 cases in different geographic regions since January 22, 2020 and are updated daily. In the research that have done by Ahouz & Golabpour (2021), Data from 252 infected regions were analyzed as of March 29, 2020, with 17,136 records and 4 variables, namely latitude, longitude, date, and records. In order to design the incidence pattern for each geographic region, the information was utilized on the region and its neighboring areas gathered 2 weeks prior to the designing. Then, a model was developed to predict the incidence rate for the coming 2 weeks via a Least-Square Boosting Classification algorithm.

As another example, researchers at the University of Southampton predicted how the disease spreads during the 40-day New Year celebrations, which are the most frequented, based on data from the International Air Transport Association and Bayloo location data using data mining algorithms. (Ara Research Center, 2019).

6. Conclusions and suggestions:

In today's world, Covid-19 quickly hit other countries in addition to China like a hurricane, killing many people around the world. In an epidemic, a country’s health care system tolerates tremendous pressure due to the increase in the use of healthcare services and surge in hospitalizations. Data mining is a computational process that is used in many fields and the main purpose of data mining is to extract patterns in this data set and turn them into "**cognition**". Data mining is the application of algorithms to patterns extracted from data without the additional steps of the knowledge discovery process. Also it has many applications in the field of health, including the diagnosis of diseases, classification of patients in disease management, finding patterns for faster diagnosis of patients and prevention of complications in them. Due to the large amount of data in the field of medicine, the data mining process has played a very effective role in the management of various diseases such as prognosis, diagnosis and treatment. Many algorithms, including ***the Simple Naive Bayes, multilayer, and j48 decision tree***, help researchers to diagnose Covid-19 disease. Also, the results of many research studies show that the use of this technology to provide diagnostic models and identify the most optimal and effective data mining algorithm can significantly help in timely, effective and economical identification of the disease. Therefore, according to the stated contents, **it is suggested:**

- Attempts in the using data mining methods and algorithms in the diagnosis and management of Covid-19 disease
- Employing specialized forces and proficient in data mining and artificial intelligence and medical sciences to improve the current situation
- Holding online or face-to-face workshops in higher education to familiarize students and researchers in this field with the benefits of using data mining, discovering knowledge from a medical database, and diagnosing and predicting effective medicine
- Carry out future research on other methods of data mining in the diagnosis of this disease

References

1. Ahouz, F & Golabpour,A.(2021). *Predicting the incidence of COVID-19 using data mining.* BMC Public Health, 21:1087, https://doi.org/10.1186/s12889-021-11058-3.
2. *Application of data mining in marketing*. (2020) Available on: https://ihemmat.com/%DA%A9%D8%A7%D8%B1%D8%A8%D8%B1%D8%AF-%D8%AF%D8%A7%D8%AF%D9%87-%DA%A9%D8%A7%D9%88%DB%8C-%D8%AF%D8%B1-%D8%A8%D8%A7%D8%B2%D8%A7%D8%B1%DB%8C%D8%A7%D8%A8%DB%8C/
3. Ara Research Center (1398). *Artificial intelligence and its application in controlling the corona virus.* Available on: https://iranthinktanks.com/artificial-intelligence-and-its-application-in-corona-virus-control
4. Awadh, W et al. (2021). *Predictions of COVID-19 Spread by Using Supervised Data Mining Techniques.* Journal of Physics: Conference Series. doi:10.1088/1742-6596/1879/2/022081.
5. Baskabadi, M. and Dostparast, M. (2020). *Modeling and data mining of global data of Covid-19 virus patients.* Iranian Journal of Emergency Medicine, Volume 7, Number 1, Article 40.
6. Ghasemi Tabaq, F et al. (1399). *Psychological Impairments of New Coronavirus Disease. A Review Study*. Rahyaft, No. 79.
7. Hai, L.(2021). *Data Mining of Enterprise Financial Management Based on AHP*. The 2nd International Conference on Computing and Data Science (CONF-CDS 2021). doi:10.1088/1742-6596/1881/4/042077.
8. Hashemi Foumani, M. and Motieian, H. (2020). *Modeling the prevalence of super-acute Avian influenza in Guilan province with data mining models and spatial information system in 2016: An ecological study*. Journal of Rafsanjan University of Medical Sciences, Volume 19.
9. Hamrahan soft (1396). *Data mining and knowledge exploration*. Available on: https://www.hamrahansoft.ir/1396/09/20/891/
10. Jahani, A. (2020). *The difference between machine learning and data mining.* Available on: https://onlinebme.com/clearly-explained-how-machine-learning-is-different-from-data-mining
11. Khosravipour, A. (2009). *Security applications of data mining and its techniques.* Bachelor Thesis, Islamic Azad University, Central Tehran Branch.
12. Mirzakhani, F. et al. (2018). *Comparison of Artificial Neural Network and Decision Tree Model to Identify and Predict Factors Related to Type 2 Diabetes.* Mashhad Journal of Paramedical Sciences and Rehabilitation - Volume 7- Number 4.
13. Moghaddasi, H. et al. (2012). *Data mining and its application in health.* Health Information Management / Volume 9 / Number 2 .
14. Moradi, G and Ghasemi, V. (2012). *Data mining technique and its application in social studies.* Journal of Social Sciences, Faculty of Literature and Human resource, Ferdowsi University of Mashhad, Spring and Summer 2012, pp. 157-178.
15. Nopour et al. (2021). *Proposing an effective technological solution for early detection of Covid-19 disease: A study based on data-driven machine learning.* Journal of Modern Medical Information, Seventh Volume, First Issue.

16. Petrovic, n.(2021). *Decision Support Based on Data Mining for Post COVID-19 Tourism Industry*. XV International SAUM Conference.
17. Safdari, R.(2021). *Using data mining techniques to fight and control epidemics: A scoping review.* Health and Technology (2021) 11:759–771. https://doi.org/10.1007/s12553-021-00553-7.
18. Shanbehzadeh, M et al.(2021). *Comparing of Data Mining Techniques for Predicting In-Hospital Mortality Among Patients with COVID-19.* Journal of Biostatics and Epidemiology. 2021;7(2): 154-173
19. Tarokh, M. and Sharifian, K. (2007). *Application of data mining in improving customer relationship management.* Journal of Industrial Management Studies, Year 6, No. 17.