Clinical Cancer Investigation Journal-ISSN: 2278-0513

Studying data mining techniques and their applications

Abstract

Data mining is a scientific field and a process that must be put into practice as a project rather than being available for purchase. The information hidden in the data may be utilized, however, the data are often large and cannot be used alone. As a result, it is becoming more and more important to use the power of data mining to find patterns and models as well as to connect various database elements to find the knowledge concealed in the data and, ultimately, to transform the data into information. Finding practical patterns in the data is sometimes referred to as data mining. An effective model is innovative, valid, easy to grasp, and explains the connection between a subset of facts. In this study, we investigated the literature on data mining methods and applications.

Keywords: data mining, classification, deep learning

Paniz Arshadi Kalvanagh*
Department of Computer Engineering, West Tehran Branch, Islamic Azad University, Tehran, Iran
Ali Mansoor
Department of Computer Engineering, West Tehran Branch, Islamic Azad University, Tehran, Iran
** Corresponding*
Email: panizarshadi2011@gmail.com

Introduction

As soon as the computer began to be used for data analysis and storage, in 1950, the quantity of data stored in this doubled every couple of decades. Meanwhile, as IT progressed, the amount of data in databases increased on a biennial basis. The amount of work has increased, and data storage capacity is increasing faster than in the past. Since the World Wide Web has existed, integrated information systems, integrated banking systems, electronic commerce, etc., have added to the volume of data in databases continuously and have led to the creation of enormous data warehouses, making the need for rapid and precise knowledge discovery and extraction from these databases more obvious.

The importance of rapid information access or timeliness has doubled due to the degree of competition in areas of science, social, economic, political, and military fields. Because of this, systems must be created that can rapidly find information of interest to users while focusing on minimizing human involvement on the one hand and using analytic techniques suited for the amount of massive data on the other. Now, data mining is the most crucial technique for the efficient, precise, and quick production of huge data, and its significance is growing. Between statistics, computer science, artificial intelligence, pattern theory, and machine learning, data mining serves as a link. Data mining is a challenging process in which finds accurate, new, and potentially useful models and models can be found in a large amount of data in a form that works for people.

The subject of data mining

Statisticians, database and management information system researchers, and business groups all use the phrase "data mining." Data mining is a key phase in this process, which is known as knowledge discovery in databases. It relates to the general process of learning meaningful information from data. Data preparation is one of the supplementary processes in the knowledge discovery process in databases. We can draw out valuable information from the data with the help of data selection, data cleansing, and a thorough grasp of the data mining process. Traditional data analysis and statistical methods are where data mining gets its start, therefore it covers analytical methods from different areas:

Numerical analysis

Adaptive patterns and levels of artificial intelligence, such as machine learning

Neural networks and genetic algorithms

Definitions of data mining

We can better grasp the phrase "data mining" if we take a look at its literal meaning. The Latin term "mine" refers to the exploitation of precious and hidden ground resources. The combination of this term with data denotes a thorough investigation of the data set to unearth previously undiscovered information that is helpful. Depends on the individual. Every author, researcher, and the user is influenced by the backgrounds and viewpoints of the individuals. Data mining has been defined differently by each author, researcher, and user following their point of view and attitude.

Data mining is the extraction of conceptual, unknown, and potentially useful information from the database.

Data mining is the science of extracting useful information from databases or datasets.

Data mining is the semi-automatic extraction of models, changes, dependencies, abnormalities, and other statistically meaningful structures from big databases.

Data mining is the process of extracting valid, previously unknown, understandable, and reliable information from big databases and using it in making decisions in important business activities.

The term data mining refers to the semi-automated process of analyzing big databases to find useful patterns.

Data mining means searching a database to find patterns among the data

Information storage and access management:

Information data is known as one of the vital resources of the organization and many organizations treat their organizational information and knowledge like their other valuable assets.

The precise translation of data mining will help us better comprehend this phrase. The Latin term Mine refers to the extraction of the earth's hidden and precious riches. The combination of this term with Data signifies data on a thorough examination of accessible data with a huge volume to uncover previously concealed beneficial information. It varies on the individual. Each author, researcher, and user relies on the histories and perspectives of the individuals. Each author, researcher, and the user has provided a unique definition of data mining based on their perspective and attitude.

Note: Informational data refers to raw information of the organization and information refers to processed data. Furthermore, after classification and analysis, the processed data becomes the organization's knowledge.

Imagine how difficult it would be to retrieve information in a situation where the data were improperly kept or where there were no authorized means to access them. Data must be organized logically, categorized, and stored to be more simply utilized, more quickly evaluated, and subject to better management. This is necessary to create an appropriate information system.

Structure of the organization's database:

Organizational data is kept in a variety of information banks with a variety of architectures. One of the services offered by information technology units is designing and arranging these structures, utilizing and transferring them to sophisticated information banks, and optimizing them.

Data mining and its applications

Data mining is a technique that looks for an organization's database for hidden patterns in the data, the relationships between them, and their trends and patterns. To anticipate the link between two sets of data and the likelihood that a result will occur in the future, data mining employs sophisticated mathematical functions and algorithms.

Some of the applications of data mining in real environments are:

•Banking:

Predicting fraud patterns via credit card recognition of fixed customers

Identifying the amount of using credit cards based on social groups

•Insurance:

Analysis of claims

Prediction of the number of purchases of new insurance policies by customers

•Medical:

Identifying the type of behavior with patients and predicting the success rate of surgery

Identifying the success rate of treatment methods in dealing with difficult diseases

Data mining techniques like decision trees and neural networks have attracted a lot of interest as a result of the increase in access to databases of medical facilities and medical data [1]. However, additional research is still required in the area of choosing the right algorithm type.

The kind of algorithms and their verification are the two most crucial considerations among them. The input and verification data should be based on data gathered from inside hospitals since the accuracy of the algorithms may also rely on the areas.

Data mining and its application in diabetes:

Much has been done to apply data mining to the domains of health and treatment, which is known as clinical data mining. Through the development and enhancement of information and communication technology tools, clinical data mining can be a new and useful method for discovering patterns in health data. Information containing substantial knowledge is abundant in healthcare systems; thus, approaches and tools are required to extract meaningful information from this massive data collection.

The research potential to find hidden patterns in the data set in the area of healthcare, which can be used to both diagnose and cure illnesses, is the first benefit of adopting this approach. On the other hand, several aspects are taken into account while making medical choices about this condition owing to the impact of numerous elements on it.

Therefore, the implementation of a method that can determine whether or not a diabetes diagnosis is accurate can be a crucial step in the prevention and control of this disease, especially in its early stages. These cases, along with the existence of numerous information banks that provide information about patients with diabetes registered and kept in different periods, can also help in preventing and controlling this disease.

Data mining may be a useful tool for academics studying diabetes to uncover patterns and knowledge buried in the voluminous information available on the condition and aid doctors in making a diagnosis.

Data mining steps

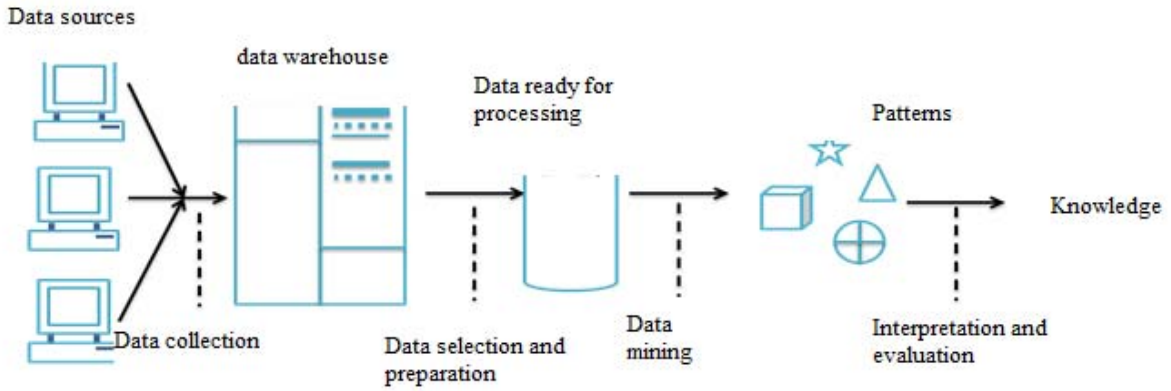

Figure 1: Steps in the knowledge extraction process

•Data collection
• Data selection
• Data Preparation
•Data mining
• Interpretation and evaluation
• Data mining algorithms

We discuss a few of the models and algorithms used for data exploration in this section. To accomplish this purpose, the majority of products use various sorts of algorithms described in computer science or statistical papers together with their unique implementation. For instance, numerous suppliers provide CART decision tree variants that may operate in parallel on computers. Even though some suppliers' algorithms may not have any extra requirements or features, they could nevertheless function properly.

The fact that no model or algorithm can and should not be utilized alone is maybe the most crucial factor. The models and techniques you use for any particular issue will depend on the kind of data you utilize. In this sector, there is no such thing as the best model or algorithm. To discover the optimum model, you will thus need a variety of tools and technologies.

- Neural networks

For classification, clustering, mining features, prediction, and pattern recognition, the neural network approach is utilized. Commonly referred to as "neural networks," artificial neural networks (ANN) are a kind of mathematical model inspired by biological processes. Neural systems are algorithms for optimization and unrestricted learning that are based on ideas drawn from studies into the workings of the brain.

- Decision tree

Artificial intelligence uses trees to illustrate a variety of ideas, including phrase structure, equation structure, game modes, etc. A decision tree is a tree in which samples are categorized using a method that advances from the root to the leaf nodes at the end. The decision tree's designation derives from the fact that it illustrates the process of choosing a category for an input sample.

- Simple Bayes classifier

Simply described, the Bayes method is a way to categorize things according to their likelihood of occurring or not. The basic Bayes classifier will provide accurate results after initial training based on the intrinsic properties of probability, particularly the probability distribution. The Naive Bayes method uses supervised learning as its learning strategy.

Data mining software

During the past years, there was a rapid flow of interest in data mining in software markets. Most of the users of data mining software, thinking of commercial use of this software, want to apply it. Data mining software commonly utilizes three various methods to apply data mining:

1) Discovery
2) Using predictive models
3) Use of controversy analysis

Without any preconceived notions, discovery is the act of exploring the data to find hidden patterns. The future is predicted using the patterns found in a data bank by data mining software based on predictive models. Predictive models let the user use unknowable data, and the program finds these unknowable values.

The patterns found from the data are utilized to identify abnormal values in competing models. The normal values must be known to derive the abnormal values, which may then be determined based on them. Data mining software is now less popular than other intelligent software. However, data categorization, an estimate of unknown values, and prediction of unknown values are three broad categories that may be used to categorize this software's commercial activity. Data clustering, an approximate grouping of data, and descriptions of links between data.

Today's companies, organizations, colleges, and institutes of higher learning are drowning in data and information, most of which is utilized just to carry out immediate tasks and not yet for strategic decision-making. Strategic decision-making may benefit from the use of data mining, whose use is growing every day. This information is accessible at higher education institutions and centers. Utilized information resources Data mining makes use of a variety of information sources, such as:

- Data warehouses
- Files
- Web
- Object-oriented databases
- Multimedia
- Data warehouse

A single data set known as a data warehouse is used by many firms to collect and store their data from heterogeneous and homogeneous data sources. Data from the past and present that will be utilized for planning and forecasting in decision support systems may be found in the data warehouse.

Data mining techniques

- Category

This method, which is the most popular, uses several specified samples to build a model that can categorize different recorded situations. This approach combines learning with classification, and it often employs decision tree or network classification methods.

To estimate the rules properly, educational data is assessed using a classification algorithm; if the accuracy is adequate, it may be applied to new situations. The technique for categorizing training items employs pre-classified samples to identify a set of parameters necessary for accurate classification. The program then transforms these parameters into a model referred to as classification.

- Clustering

Even the greatest data mining algorithms may not be able to uncover significant patterns from complicated data structures. Using clustering, it is possible to identify intricate data structures and disentangle conflicting competing signals. The process of dividing a diverse population into several subsets or homogeneous clusters is known as clustering.

The difference between clustering and classification:

Every piece of data is categorized according to a model in the chosen monitoring section. These categories were established by earlier studies. The clustering approach, however, groups the data based on similarity rather than a preset category, and the user chooses the names for each group.

- Regression

Regression analysis may be used to make predictions. One or more independent and dependent variables may be modeled using regression analysis.

The independent factors in information extraction are the features that are previously known, whereas the dependent variables are connected to the outcome that we want to predict. Unfortunately, many real-world issues are difficult to forecast, necessitating the use of more intricate decompositions and a variety of models for regression and classification. For instance, classification trees and regression trees may be created using the decision tree classification algorithm (CART). Models for classification and regression may be produced using neural networks.

Genetic algorithm

The development of the genetic algorithm, a heuristic search method that mimics the course of natural evolution, has been consistently used to provide practical answers to optimization and search issues. The creation of optimum solutions to problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover, is what genetic algorithms are. This is how nature selects the stronger (more worthy) for life.

Performed Research

We may use the study done in [1] as an example of internal research done in the area of employing data mining in the prediction and prevention of gestational diabetes. In [1], studies have been conducted, and two approaches of neural network and decision tree in data mining to analyze and experimental data analysis and forecasting have been used. This survey is about data mining based on neural networks and decision tree methods in early diagnosis of the risk of gestational diabetes. The extracted data were prepared in Matlab software before being normalized and examined. Are the two data mining techniques of neural networks and decision trees reliable in quickly and correctly diagnosing the risk of gestational diabetes? is the issue that this study tries to find an answer to. And are they capable of accurate diagnosis? The research's findings indicate that data-oriented methods are effective in increasing forecasting accuracy and correctness, that they perform satisfactorily in identifying tacit knowledge and hidden relationships between data, and that the decision-making error in both methods is acceptable and fairly comparable.

The authors recognize in [2] that the goal of this reference is to identify methods for precisely diagnosing the illness and to examine data patterns utilizing classification analysis and straightforward Bayes and decision tree classification algorithms. Researchers had hoped to develop a quicker and more effective way of illness diagnosis, which would enable patients to get prompt care. Automatic diabetes diagnosis has been identified as a significant medical issue by Eyre and colleagues [2]. The secret to treating diabetes is identifying it early on. The usefulness of these models is shown by their experimental findings. These algorithms have been shown to

perform well in studies assessing the effectiveness of approaches for diabetes diagnosis [2].

It has been acknowledged in [3] that gestational diabetes is the diagnosis when a woman has high blood sugar (glucose) during pregnancy but does not have pre-pregnancy diabetes. It was recently reported that roughly 18% of pregnant women had gestational diabetes, according to the criteria for diabetes. Using data mining methods, this reference aims to enhance the diagnosis of gestational diabetes. Additionally, the effectiveness of supervised learning methods, a straightforward Bayes classifier, and a random tree have been examined in this reference. The random tree is one of the most effective and precise algorithms in this area, according to experimental findings. Numerous studies have been undertaken in the area of diabetes prediction, according to Meri and colleagues [4]. Some more examples of similar techniques include J48 and CRT.

The mathematical expectation-maximization (EM) algorithm, genetic algorithm, and H-means clustering have all been employed by the authors of [4] to identify diabetes patients. When all of the signs are identical to the clusters, these techniques perform better than other algorithms. A fuzzy model technique and a fuzzy expert system have been developed by Mirsharif and colleagues [5]. When a patient doesn't have access to a doctor, a fuzzy inference expert system has been utilized to make the diagnosis of gestational diabetes. A model-based technique called a fuzzy expert system necessitates the gathering of expert information. In the fuzzy expert system model, which may be utilized as a decision aid, rules are extracted and gathered in the knowledge base under the supervision of experts, experts, and medical reference books. The fuzzy inference system has a mean square error level of 0.2%.

In [6], to classify diabetic and non-diabetic people with diabetes, the traditional classification method (Binomial logistic regression and Fisher's Linear Discriminant Analysis) and machine learning classifiers (neural networks, support vector machines, fuzzy c-mean clustering, and random decision forests) were used. 6500 instances were included in the database utilized for this investigation. To assess the prevalence of the main non-communicable disease risk factors in Tehran hospitals, a cluster sample of the Iranian population was carried out in 2005 and 2007. To assess the performance of six instances in terms of sensitivity, specificity, full accuracy, and the area under the receiver operating factor curve criteria, 10 risk variables that are often linked to diabetes were chosen. A comparison in the area of approaches is shown in [6] and is based on actual data. Iran's diabetes has been predicted using data mining. Support vector machine models' superior discriminating performance over other techniques was noted in this source. As a result, it may be used to accurately diagnose diabetes using simple clinical procedures.

It is acknowledged in [7] that information collection can be done due to the following three steps:

1. Collecting information due to a questionnaire about health history and behavioral information
2. Applying standard physical measurements to collect physical and physiological information
3. Blood sampling for biochemical investigation and laboratory tests of lipids and glucose status, which is done by trained personnel.

According to [7], risk indicators including body mass index and family history of diabetes may be used to identify those who are at risk of developing diabetes. Another method for doing this is by logistic regression. According to the authors, prediabetes models' prediction abilities may be increased by learning theory and data mining techniques, which is crucial for classification without distributional presumptions. Especially when the dependent variable is bilateral, traditional methods like logistic regression and Fisher's linear discriminant analysis have been extensively employed to categorize various issues. It seems that more research is still needed to fully understand the beneficial performance of data mining methods using dividers such as neural networks, support vector machines, fuzzy C-Mean, and random forests.

In [8], the use of data mining for the prediction of diabetes has been investigated, and the methodologies of logistic regression, neural networks, Gaussian mixture models (GMM), support vector machines (SVM), and others have all been examined. The usage of artificial networks, as recognized by the authors, has improved diabetes prediction results.

The decision tree, which is straightforward to use and may provide interpretable results, has been described in [9] as one of the popular and frequent classifications for classification and prediction. In addition to providing its predictions in the form of laws that are adequate in terms of statistical parameters, decision trees are capable of producing human-understandable descriptions of the connections in a data set. This approach of learning is used to incorrect data and discrete functions and aids in knowledge discovery.

Medical centers often deal with vast, intricate, and diverse amounts of data. It should be remembered that these data come in a variety of forms, and that missing values should also be remembered. Three data mining algorithms, self-organizing maps (SOM), C45, and random forest, were examined in [9] using adult population data from the Saudi Arabian Ministry of Health. According to the findings of the study in [9], the random forest performed well when compared to other dividing cases.

Conclusion

Data mining is the study of patterns, logical relationships, and other novel and important information that may be found in data. Finding meaningful patterns in data is conveyed by a variety of titles in various civilizations (such as data mining). For instance, terms like data pattern processing, knowledge extraction, and information discovery might be stated.

By examining many studies, it is evident that each of these studies tended to pick out and analyze just a small number of data mining algorithms, and then, using the results, they provided their recommendation for the best algorithm. Additionally, there haven't been many studies in our nation on gestational diabetes, and it's still unclear which methodology would be best for the Iranian sample. A proper statistical population may be extremely helpful since previous studies often had a small statistical population. Given the uniqueness of this particular sector, a more thorough investigation is required in every field.

Acknowledgments

None.

Conflict of interest

None.

Financial support

None.

Ethics statement

None

References

1. Mirsharif and Rouhani, data mining based on neural network and decision tree methods in early diagnosis of the risk of gestational diabetes, Journal of Informatics Center for Health Research and Biomedical Informatics, 139

2. Iyer et al, diagnosis of diabetes using classification mining techniques, International Journal of Data Mining

3. Nagarajan et al, Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes, ijsr 2014

4. C. M. Velu and K. R. Kashwan, “Visual Data Mining Techniques for Classification of Diabetic Patients”, 3rd IEEE International Advance Computing Conference (IACC), 2013.

5. Mirsharif M, Alborzi M. A fuzzy expert system & neuro-fuzzy system using soft computing for gestational diabetes mellitus diagnosis. International Journal of Information, Security and Systems Management 2014;3(1):249-52.

6. Tapak et al, Real-Data Comparison of Data Mining Methods in Prediction of Diabetes in Iran, HIR, 2013.

7. Poorolajal J, Zamani R, Mir-Moeini R, Amiri B, Majzoobi M, Erfani H, et al. Five-year evaluation of chronic diseases in Hamadan, Iran: 2005-2009. Iran J Public Health 2012;41(3):71-81

8. Kim S, Kim W, Park RW. A comparison of intensive care unit mortality prediction models through the use of data mining techniques. Healthc Inform Res 2011;17(4):232-43.

9. Sapra RL, Mehrotra S, Nundy S. Artificial neural networks: prediction of mortality/survival in gastroenterology. Current Medicine Research and Practice 2015;5(3):119-29.