Advanced Production and Industrial Engineering
R.M. Singari and P.K. Kankar (Eds.)

doi:10.3233/ATDE220732

A Future Awareness for Healthcare Applications Using Data Mining

[1]K. T. Sree, [2*]R Anil Kumar and [3]G. Rama Naidu

[1,2]Aditya College of Engineering & Technology, Surampalem, India
[3]Aditya Engineering College, Surampalem, India
anilkumar.relangi@acet.ac.in

Abstract. This research report showcases various data mining (DM) techniques such as Classification, Regression, and Clustering in brief and also discusses the aptest framework method for the healthcare industry, CRISP-DM. This report also explores the various data mining applications in the healthcare industry. DM is utilized to extract the data from a lot of information. DM includes two models, predictive and descriptive. Classifying data is to form classes either with the final objective of learning new antiques or searching new areas. This is why specialists have for many years tried to locate the enshrouded examples in the knowledge that can be classified and contrasted as well as other concepts which are the result of common principles.

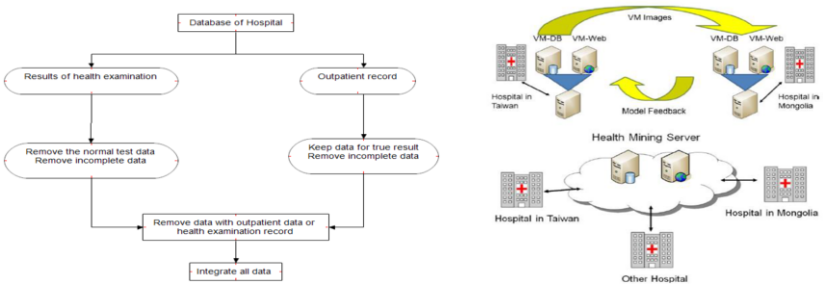
Keywords: Data mining, health, algorithm

1. Introduction

Data mining (DM) is the technique of finding patterns and relationships in huge data sets to solve problems via data analysis. DM techniques are unique in identifying unknown links within massive data sets. Healthcare has been revolutionized through the use of data mining. DM has numerous health care sector uses, such as effective hospital resource management, fraud detection, and prioritizing patients based on the severity of their disease. Data mining concepts have spread widely across several industries throughout the last two decades [1-3], [4]. Many data mining techniques are classification, association, regression, and clustering, all of which are more detailed and it is found that more efficient in discovering undiscovered relationships between data sets than processing statistics one by one [5], [6], [7], [8]. Algorithms have now been specifically designed towards improving hospital norms and creating better policies. Data mining is the analytic process of Knowledge Discovery in Databases (KDD). This consists of five processes that are namely Selection, Pre -Processing, Transformation, DM, and Evaluation. In light of these challenges, a process model for carrying out data mining projects that is independent of both the industry sector and the technology employed was established [9]. The CRISP-DM process model is meant to reduce the total cost of large data mining projects, while simultaneously increasing its reliability, reproducibility, scalability, and speed [10-14].

DM techniques face challenges as well due to the lack of quality data in the health care industry. Data generated from hospitals is heterogeneous as patients are either not interested in disclosing personal information or might simply give incorrect information as shown in Figs. 1 and 2. The basic building block for knowledge discovery is the

correct information to derive unknown relationships and thus is the biggest challenge there is to overcome in this field [15-19].



(a) **(b)**

Fig. 1: (a) Data integration Flow chart (b) transnational medicine applications [1]

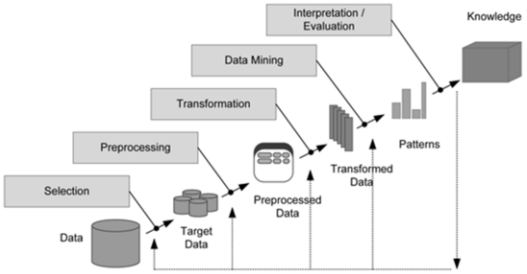


Fig. 2: Data mining technique [2]

AI is a way to replicate human intellect, much like the Wright Brothers did while designing the first successful flying machine.

2. Related Works

Data Mining techniques were developed in the early 1990s due to an ever-increasing amount of data collected by various public and private sectors and the need to make an automated framework that would work upon these large data sets generated and find valuable unknown relationships and Knowledge Discovery (KDD) [8]. To some extent, the market is under the impression that Data Mining is a simple technology that can be implemented quickly. But this is not true because data mining is a complex process needing a variety of tools and techniques and professionals with the appropriate skills and knowledge [12]. Constraints such as professional skills and various tools hinder the authenticity of the work is done; we need to set a standard process that applies to all industries. This standard process will set Data Mining as an established engineering practice [14]. The Cross-Industry Standard Process for Data Mining aims to make large data mining projects less costly, more reliable, faster, and easier to implement. This model will also set guidelines for novices in the industry helping them out by learning basic cross-industry frameworks. CRISP-DM consists of six sub-processes. The first and second sub-processes are about collecting and analyzing data generated. Data pre-processing and modeling are examined in 3rd and 4th sub-processes respectively. Evaluation and Deployment are the fifth and last sub-processes. Coming to the health

care industry, the CRISP-DM process has been helping hospitals make the right administrative as well as medical decisions [15-19].

3. Methodology analysis for DM applications

3.1. Classification

Classification is a data mining technique that groups data samples into target classes. Classification predicts the target classes for each data point. There are two methods of the classification technique i.e., binary classification and multilevel classification. Binary classification has two possible sub-classes namely high level and low level. Multilevel classification has more than two sub-classes, for example, low level, medium level, and high level. The data set is partitioned into two sets - training and testing. The training data set is used to train the classifier while the testing data set is used to test the correctness of the classifier. The method that achieved better accuracy than others was the ensemble classification method [6-9].

3.2. Decision Tree (DT)

The decision tree is similar to a flowchart where each non-leaf node denotes a test on a particular attribute and each branch indicates a result of that test and each leaf node has a class label. As illustrated in Fig. 3, the root node is the highest-level node. Decision trees are classifications that make use of a tree-shaped graph as a basis. The decision tree is most commonly used in operations research analysis to calculate conditional probabilities, which is the most popular application. When using a decision tree, decision-makers can select the most advantageous alternative and the root-to-leaf path, suggesting a distinct class separation based on the largest amount of information gained in the process. The decision tree has a wide range of applications in the field of health [8-13].

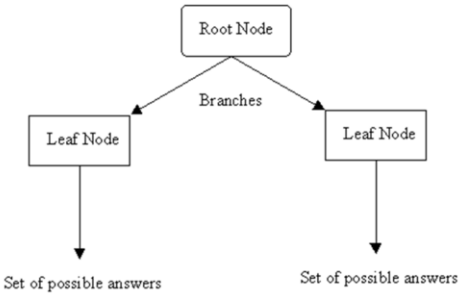


Fig. 3: Decision Tree [16]

3.3. K - Nearest Neighbor (K-NN)

The classifier that makes use of formerly recognized data points (DP) and classed DP to find previously unknown data points, known as "nearest neighbors," is one of the simplest classifiers that may be used to locate the unidentified data points. As demonstrated in Figure 4, K-NN is utilized in various industries such as healthcare,

cluster analysis, and search engine optimization (SEO). To help determine the causes of the individuals who have heart illness, it applied a K-NN classifier. This dataset was derived from the UCI and used to conduct an experiment, which found the K-NN model was more accurate for diagnosing heart problems. This categorization algorithm, based on the biological nervous system known as the nervous system, is made up of several interconnected processing elements known as neurons, and this system works in unity to find a solution to a given problem. As you can see in Figure 4, rules are derived from a trained neural network to assist with network interoperability. NN uses neutral network, which are ordered processing elements, to solve a certain problem. As a NN is adaptive, it can alter its size and geometry to decrease the mistake. The multilayer, generalized regression, probabilistic neural networks model was built utilizing the Artificial Neutral Network (ANN) for the investigation of chest disorders. It is used to find out if a person has a lung condition. These experiments evaluated the Chest Computed Tomography (CT) and discovered several lung tissues features that helped to decrease the CT data size. Then, those data features were passed on to a neural network, which was used to detect numerous disorders related to the lungs [6-13], [19-22].

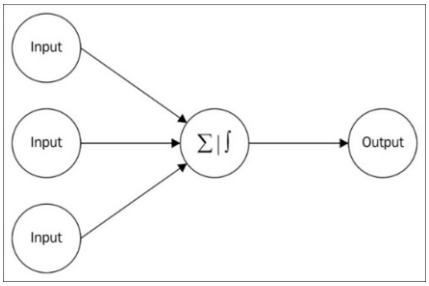


Fig. 4: Single layer Neural Network

3.4. Clustering

Regression is learning a function that maps a data item to a real-valued prediction variable. The model can have more than one independent variable but there can only be one dependent variable. Clustering is a data mining technique that is very different from the classification technique because, unlike classification techniques, clustering has no predefined classes. Clustering is an unsupervised learning method. Data points that are in the same vicinity or are closer to each other have more similarities than data points from another cluster. Over the past few decades, the clustering technique has been used and various other clustering techniques have been established. One of the advantages of clustering is that it needs less to no information for analyzing the data [10], [22].

3.5. Partitioned Clustering

Data points are first selected at random and then assigned to 'k' centroids, depending on some similarity metric. The mean for the cluster is supplied to the cluster for each iteration (the distance between the data points). This is done to take into account every recently added data point. The technique is designed to generate compact clusters of comparable data points with the disparate fare. The centroid of the cluster is often referred to as the cluster mean. Initiates sophisticated clustering strategy using K means,

as it is a self-organized approach. K-Means employed means, but K-medoids used medoids instead. Medoid is centrally positioned data. To start, choose the medoids arbitrarily for each cluster, and then data points are placed with their most similar medoid. Public domain applications and it is applied the k-means method in the public health service to find the occurrence of breast cancer. A different study paper uses Data Mining approaches in healthcare. Groundwater contamination can be studied using the technique of clustering. K-means clustering helped find risk factors linked to fluoride concentration in water [8], [18-21].

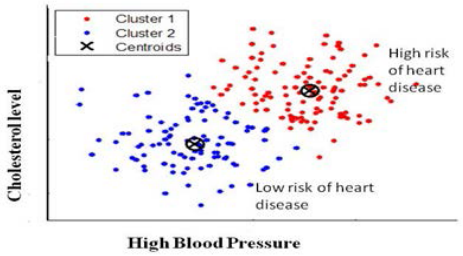


Fig. 5: heart disease patients clustering [8]

4. Different applications

Data mining techniques have proved their worth in the health care industry as the amount of data generated is immense, which is the basic building block of a data mining process.

(i) Hospital grading

The various hospital details are evaluated using a variety of data mining methodologies to determine their rankings. Locating hospitals on the high-risk patient scale is done based on their ability to deal with challenging patients. The higher-ranked hospital deals with the more dangerous patients on its highest priority, whereas the lower-ranked hospital does not take any special precautions to tackle the potential threat.

(ii) Insurance Frauds

Insurance developers use data mining techniques to resolve this issue. Working upon the data frauds can be caught by analyzing the data which gives patterns of the individual committing the fraud.

(iii) Medicine and prevention errors

Medical institutions can gain new, useful, and perhaps life-saving knowledge by applying data mining to their existing data. Issues that affect patient safety can be discovered by mining records from the hospital. These findings could then be reported to hospital administration and government regulators.

(iv) Recognize High-Risk Patients

High-risk patients need to be attended before anyone with a minor injury. First cum first serve is not a viable option since there is always the risk of losing the patient's life

who is more critical. Through data mining, predictive models can be made that identify the higher-risk patients. For example, a diabetic patient needs to be attended before a patient who needs to be treated for cough and cold.

5. Conclusion

This research report of data mining applications in the field of healthcare provides a basic overview of the current practices and present, available techniques. Healthcare organizations can tap into these techniques in more depth and extract knowledge to tackle disease outbreaks in a better way. Various techniques have been explained to understand the working and the differences they have from one another. But a crucial point that should be taken into account is the security of patient information and records. Every organization needs policies in place regarding patient privacy and security. Reasons such as fast-moving pandemics and noninvasive, painless ways to detect the onset of the disease result in increasing demand for health organizations to integrate data and use data mining to mine the data sets.

References

[1] Yoo, Illhoi, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, and Lei Hua. "Data mining in healthcare and biomedicine: a survey of the literature." *Journal of medical systems* 36, no. 4 (2012): 2431-2448.

[2] Wang, Hai, and Shouhong Wang. "A knowledge management approach to data mining process for business intelligence." *Industrial Management & Data Systems* (2008).

[3] Kona, Ashok Kumar, R. Anil Kumar, and Sanjeev Kumar. "Wireless Powered Uplink of NOMA Using Poisson Cluster Process with Two Orthogonal Signal Sets." In *ICCCE 2020*, pp. 1105-1113. Springer, Singapore, 2021.

[4] Sudeep, Sista Venkata Naga Veerabhadra Sai, S. Venkata Kiran, Durgesh Nandan, and Sanjeev Kumar. "An Overview of Biometrics and Face Spoofing Detection." *ICCCE 2020* (2021): 871-881.

[5] R. Kandwal, P. K. Garg and R. D. Garg, "Health GIS and HIV/AIDS studies: Perspective and retrospective", Journal of Biomedical Informatics, vol. 42, (2009), pp. 748-755.

[6] Sambangi, Jagadeeswari, Parvateesam Kunda, Durgesh Nandan, and Sanjeev Kumar. "An Overview of Fog Computing." *ICCCE 2020* (2021): 843-852.

[7] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "The KDD process of extracting useful knowledge from volumes of data.commun.", ACM, vol. 39, no. 11, (1996), pp. 27-34.

[8] Gundabathula, Geetanjali, Parvateesam Kunda, Durgesh Nandan, and Sanjeev Kumar. "Implementation of Cloud Based Traffic Control and Vehicle Accident Prevention System." In *ICCCE 2020*, pp. 1125-1134. Springer, Singapore, 2021.

[9] Nimmakayala, Satish, Bhargav Mummidi, Parvateesam Kunda, and Sanjeev Kumar. "Modern Health Monitoring System Using IoT." In *ICCCE 2020*, pp. 1135-1144. Springer, Singapore, 2021.

[10] Agarwal, Shivam. "Data mining: Data mining concepts and techniques." In *2013 International Conference on Machine Intelligence and Research Advancement*, pp. 203-207. IEEE, 2013.

[11] Kaur, Harleen, and Siri Krishan Wasan. "Empirical study on applications of data mining techniques in healthcare." *Journal of Computer science* 2, no. 2 (2006): 194-200.

[12] Yi, Wen, Albert PC Chan, Xiangyu Wang, and Jun Wang. "Development of an early-warning system for site work in hot and humid environments: A case study." *Automation in Construction* 62 (2016): 101-113.

[13] Khamis, Hassan Shee, Kipruto W. Cheruiyot, and Stephen Kimani. "Application of k-nearest neighbour classification in medical data mining." *International Journal of Information and Communication Technology Research* 4, no. 4 (2014).

[14] Levashenko, Vitaly, and Elena Zaitseva. "Fuzzy decision trees in medical decision-making support system." In *2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 213-219. IEEE, 2012.

[15] Geller, James. "Data mining: practical machine learning tools and techniques with Java implementations." *SIGMOD Record* 31, no. 1 (2002): 77.
[16] Filimon, Delia-Maria, and Adriana Albu. "Skin diseases diagnosis using artificial neural networks." In *2014 IEEE 9th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI)*, pp. 189-194. IEEE, 2014.
[17] Rawlings, John O., Sastry G. Pantula, and David A. Dickey. *Applied regression analysis: a research tool*. Springer Science & Business Media, 2001.
[18] Fernández-Navarro, Francisco, César Hervás-Martínez, Roberto Ruiz, and Jose C. Riquelme. "Evolutionary generalized radial basis function neural networks for improving prediction accuracy in gene classification using feature selection." *Applied Soft Computing* 12, no. 6 (2012): 1787-1800.
[19] Hruschka, Eduardo R., and Nelson FF Ebecken. "A genetic algorithm for cluster analysis." *Intelligent Data Analysis* 7, no. 1 (2003): 15-25.
[20] Coates, Adam, and Andrew Y. Ng. "Learning feature representations with k-means." In *Neural networks: Tricks of the trade*, pp. 561-580. Springer, Berlin, Heidelberg, 2012.
[21] Tomar, Divya, and Sonali Agarwal. "A survey on Data Mining approaches for Healthcare." *International Journal of Bio-Science and Bio-Technology* 5, no. 5 (2013): 241-266.
[22] Wang, Xiao-Ying, and Jonathan M. Garibaldi. "Simulated annealing fuzzy clustering in cancer diagnosis." *Informatica* 29, no. 1 (2005).