IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Issue-4, April 2019 E-ISSN: 2347-2693

Data Mining Techniques in Biological Research

Dipti N. Punjani[1*], Kishor H. Atkotiya[2]

[1]National Computer College, Jamnagar, India
[2]Department of Statistics, Saurashtra Unviersity, Rajkot, India

DOI: https://doi.org/10.26438/ijcse/v7i4.339343 | Available online at: www.ijcseonline.org

Accepted: 10/Apr/2019, Published: 30/Apr/2019

Abstract- In current era, the trend of application of data mining is widely used because of health sector is rich in information and data mining has become its necessity. In the healthcare organizations many data and information is generated on daily basis. Use of data mining and knowledge that help bring some interesting patterns which means eliminate manual tasks and easy data extraction from any electronic records, through that will secure medical records, save patient's lives and also reduce the cost of medical services as well as early detection of any infectious disease on the basis of historical and advanced data collection. Data mining can enable healthcare organizations to predict trends in the patient's medical condition and behavior proved by analysis of different prospects and by making connections between totally unrelated data and information. Generally the raw data from the healthcare organizations are tremendous and heterogeneous. These all data can be gathered from various sources or different components. Data mining has great importance for area of healthcare and also it represents comprehensive process that demands through understanding of requirement of the healthcare organization. Knowledge gained with the use of techniques of data mining can be used to make successful decisions that will improve success of healthcare organizations and also health of the patients. Data mining once started, represents continuous cycle of knowledge discovery. In this paper, I wish to discuss that how data mining is used in infectious disease like cervical cancer.

Keywords- Data Mining, Knowledge Discovery Database, Cervical cancer, Classification, Clustering

I. INTRODUCTION

Data mining is the process of analyzing messy and noisy data in an effort to find the patterns, correlation and also forthcoming. In another term data mining is also Data Discovery. With an enormous of data stored in databases and data warehouses is used in the significant business value by improving the effectiveness of managerial decision making. Recent developments in information technology have enabled collection and also different types of processing of messy or noisy personal information, such as shopping details or habits, banking credit and debit card details and also in medical history. Data mining can enable health organizations to predict tendency in the patient's medical condition and also behavior proved by analysis of prediction, by making connections between seemingly unrelated information. The raw data from healthcare or medical are voluminous and also heterogeneous. When there is a large volume of data that must be processed by the people, making decisions is generally in poor quality (1). It should be collected and also stored in organized form and also their integration allows the formation connect with medical information system. Data mining in healthcare offers unlimited possibilities for analyze different data models a lesser amount of visible or hidden to common analysis techniques.

Data mining has been used increasingly in many filed such as retail industry, finance, medical and health care, biomedical and also in DNA research (2, 3). The examination of health data is used to improve the healthcare by attractive performance of patient management tasks. The product of data mining are provide a different number of benefits of healthcare organizations like grouping of patients, having a similar type of diseases or health obstacles so that organization provides them special and also effective treatments. This technique is also used for predicting the number of days to stay of patients in hospitals, for specific type of medical diagnosis and making plan for effective information management. Data mining techniques are also useful to find out the various factors, which are responsible for disease for example, different types of foods, low education or awareness level, different working environment, living situation, availability of healthcare services etc (4). In this paper, we wish to discuss that how data mining is used in various infectious disease like cervical cancer.

II. DATA MINING AND KNOWLEDGE DISCOVERY DATABASE

Data mining originated as interdisciplinary field of statistics and machine learning and then advanced from these

beginnings to include artificial intelligence, pattern recognition, data base technology etc (5). Data mining is also known as data extraction or knowledge extraction, data dredging and also information obtain etc (6). Researchers consider that data mining is one of the steps of KDD process.

The Knowledge Discovery Databases (KDD) model is an iterative and interactive model. It has different step, it refers (7, 8) to finding knowledge in data and highlight the high level of specific data mining method.

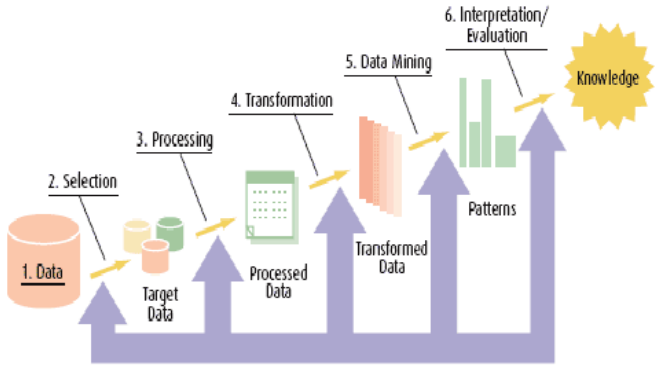

Fig: 1 KDD Process Model

The Knowledge Discovery Databases is the process of eliminating the hidden information or knowledge from any databases or data warehouse. There are different stages of KDD given detail in below;

Selection stage consists on generate a target dataset, or focusing on a subset of variables or attributes, on which discovery is to be performed.

Pre-processing consists on that targeted data cleaning and also preprocessing in order to achieve reliable or consistent data.

Transformation is focus on the transformation of the data using different transformation methods.

Data Mining consists on the searching appropriate patterns of interest in a particular presentation format, depending on the different objectives of data mining like prediction.

Interpretation/Evaluation consists on the evaluated of the mined patterns.

III. CERVICAL CANCER AS AN INFECTIOUS DISEASE

Cancer starts when cells in the body begin to grow in unstoppable manner. Cells in nearly any part of the body can become cancer, and also can spread to another part of the body. According to Dan L. Longo (2015), Cancer is an exception to the harmonize interaction among cells and organs. In general, the cells of a multi-cellular organism are programmed for collaboration. Many diseases occur because the specialized cells fail to perform their assigned task. Cancer takes their malfunction one step further. Not only there is a failure of the cancer cell to maintain its specialized function, but it also strikes out on its own; the cancer cell competes to survive using natural mutability and natural selection to seek advantage over normal cells in a recapitulation of evolution. One consequence of the traitorous behavior of cancer cells is that the patient feels betrayed by his or her body. The cancer patient feels that he or she, and not just a body part, is diseased. (pp. 467)

Cervical cancer starts in the cells line the cervix – the lower part of the uterus or womb. This is also sometimes called the uterine cervix. The cervix connects the body of the uterus to the vagina- birth canal. The part of the cervix closest to the body of the uterus is called endocervix. The next part ot the vagina is the ectocervix pr exocervix. The 2 main types of cells covering the cervix are squmous cells – on the ecovervix and glandular cells- on the endocervix. The place these cell types meet is called the transformation zone. Most cervical cancers start in the cells in the transformation zone (10).

Cervical cancer is the most common cancer cause of death among women in developing countries (11). Mortality due to cervical cancer is also an indicator of health discrimination, as 80% of all deaths (12) due to cervical cancer are in developing, low income and also in middle income countries (13).

Every year in India, 1,22,844 women are diagnosed with the infectious disease cervical cancer and 67,477 die from the disease. It is the second most common cancer in women aged 15-44 years. India also has the highest age standardized incidence of cervical cancer in South Asia at 22, compared to 19.2 in Bangladesh and 13 in Sri Lanka (14).

According to PrashantNaresh (15), lung cancer risk prediction system is very helpful in detection of any person's tendency for lung cancer. The early detection is pivot role in disease diagnosis process and for an effective preventive strategy.

Ravi Kumar et al. (16), in this paper, discuss that breast cancer is one of main causes of death in women, compared to all other cancers. So, early detection of breast cancer is important in reducing life wounded. In this paper, different methods for breast cancer detection are explored and also their accuracies are compared. With all these results, the SVM is more suitable in the classification problem of breast cancer prediction.

K. Srinivas et al (17)., illustrates the potential use of classification based data mining techniques such as Bayes theorem, Decision Tree and also Artificial Neural Network to noisy healthcare data. Using different attributes like an age, sex, blood sugar and also blood pressure, it can predict that patients getting a heart disease.

According to A. Sudha et al (18)., has emphasized an idea about major life affecting diseases and their diagnosis using data mining with minimum attributes and also awareness about diseases which direct to death.

Neha Sharma (19), illustrates an ED&P framework which is an information system to develop a Data mining model for early detection and also prevention of malignancy of Oral Cavity.

Usage of Cardiovascular disease (20) datasets from the University of California Irvine Data, the authors divided heart disease into five classes through their methods can be applied to any serious Disease (blake et al., 2001)
According to Sushmita et al., For effective pattern (21) classification and rule generation, Decision support System for cervical cancer management and also for staging was designed with the help of soft computing tools like rough set theory, genetic algorithm and also neural network is used to build up efficient decision making system.

IV. DATA MINING TECHNIQUES IN HEALTHCARE

A. Classification:

This is one of most popular technique, which is very useful in Health sector. Through the help of classification, data can be divides into the target classes. This technique is used to predict the target class for every data samples. This technique is also useful for a risk factor can be associated to patients by analyzing their patterns of diseases. This technique is also called supervised learning approach. In this technique, every dataset is divided into two categories as training and testing data set. This method is also called as supervised learning.
Some of other various classification algorithms are also used in health care sectors;

1) K-Nearest Neighbor (K-NN):

This classifier is one of the simplest classifier technique to discovers the unidentified data point using the previously data points is known as nearest neighbor and classified dataset according to such that system (22). This method has a number of functions in different areas such as health data sets, pattern recognition, cluster analysis etc.

2) Support Vector Machine (SVM):

This technique is very useful because there are two parts of implementing SVM. The first part employs mathematical programming and second is involves kernel functions. According to Vapnik, among all the accessible algorithms it provides exact and accurate results. This technique is ahead attractiveness because this can be easily extended to any problems related to multiclass though it was developed mainly for problems related to binary type classification (24).

3) Decision Tree (DT):

This technique is widely used by many researchers in health care sector. This is one of the most popular approaches for representing classification of any dataset. Using this technique, decision makers can choose best alternative and traversal from root to leaf indicates class separation based on available data set or information gain (25).

4) Neural Network (NN):

This technique is essentially based on neurons or nodes. These all neurons are interrelated within the network; they worked together in parallel format and produce the output. This technique is most widely used classification algorithms in various biomedicine and also in health sector (26, 27, 28). Neural Network is also useful to diagnosis of diseases including various types of cancers (29, 30)and predicts their outcomes(31, 32). The main use of neural network is to perform the task of classification and pattern recognition.

5) Bayesian algorithm:

Bayes theorem of statistics is plays a very important role in classification algorithm. When any probabilistic learning method implemented on data, at that time Bayesian algorithm is used (22). While in health care characteristic such as patient health condition and their symptoms are interrelated with each other but Naïve Bayesian Classifier considers that all this functions are independent with each other.

B. Regression:

Regression is basically a mathematical tool. This is also very important method of data mining. With the use of regression, we can easily find out those useful functions, which are demonstrates the correlation among different variables. When we have a training dataset, then we can easily construct it. There are two different variables, one of them is dependent and another one is independent (33).

C. Clustering:

Clustering is an unsupervised learning because it observes only independent variables; it has not any predefined classes. Using this method, large database divide into small clusters, depend upon the similarity measure (22). Clustering groups' data occurrence into small subset in such a way that similar instances are grouped together and different case belongs to different cluster as well as groups.

1) Partitional clustering:

This method is used to directly relocate objects to the number of clusters. Using this method, categorized cluster to how they relocate objects, and how they select a centroid cluster or representative cluster among within na incomplete cluster, and also find out that how they measure similarities between objects and centroid cluster.

2) Hierarchical Clustering:

This method of clustering is divided into two parts: Agglomerative and Divisive. The agglomerative method working is to merge data points into a single group, where Divisive method initially select this single group and iteratively partitioned into the smaller group until and unless every single data point relates to only one cluster (34).

3) Density based Clustering:

This method is most important in biomedical research because it is capable to handle any uninformed shape cluster. Additionally density based clustering algorithm, partitional and hierarchical clustering are handled only the spherical shape cluster not arbitrary shape cluster.

D. Association Rule:

This method did not receive more attention because of its scalability issues and also its inefficiency. But after this problem, R.Agarwal and their colleagues at IBM Research Center initiate a new association rule called Apriori algorithm (35, 36). This algorithm can be applied to any read databases, to find out the frequent patterns, interesting relationships between a set of data. It is most useful in the health care sector to find out the similarity between diseases, symptoms and also in current health situation.

1) Apriori Algorithm:

This method is invent by R. Agarwal et al (35, 36), in 1994. In this method, we have two inputs: support and confidence. Supports considered the set of transaction, which are frequently occurred in a database, where confidence considers the accuracy of it.

V. CONCLUSION

Data mining in health care management is not similar to the other fields due to the reason that the existing data in heath care sector are heterogeneous. Generally heath care sector is an emerging sector. It is more important to focus and it deals with collection, organization and storage of health related data. With the rising number of patients of any diseases and health care facilities requirements, an automated system will required to organizing, retrieving and also classifying that useful data. Data mining in medical sector has huge potential in developing diagnosis and prognosis warning systems which will help in initial proper treatment of life threatening diseases. With the help of different techniques of data mining, it is possible to improve the treatment quality of hospitals and also increase in the survival rate of patients.

REFERENCES

[1]. Eapen, A. G. (2004). "*Application of Data mining in Medical Applications*".Ontario, Canada, 2004: University of Waterloo.

[2]. Chen M. S., Han J., and Yu P. S., (1996), '*Data Mining: An Overview from Database Perspective'*, IEEE Transactions on Knowledge and Data Engineering, Vol.8, No.6, pp.866-883.

[3]. Cios K. J. (ed.), (2000), '*Medical Data Mining and Knowledge Discovery"*, Physica-Verlag (Springer).

[4]. Divya Tomar, Sonali Agarwal (2013), *"A survey on Data Mining approaches for Healthcare"* International Journal of Bio-Science and Bio-Technology, Vol-5, No-5, pp.241-266 (ISSN 2233-7849)

[5]. Illhoi Yoo, Patricia Alafaireet, Miroslav Marinov, Keila Pena-Hernandez, Rajitha Gopidi, Jia-Fu Chang, Lei Hua (2011) "*Data Mining in Heath Care and Biomedicine: A survey of the literature*" J Med Syst DOI 10.1007/s 10916-011-9710-5, Springer.

[6]. J.Han and M.Kamber(2006), *"Data Mining: Concept and techniques"*,2nd edition, The Morgan Kaufmann Series.

[7]. Brachman, R. J. & Anand, T. (1996), "*The process of knowledge discovery in databases.*" AAAI Press / The MIT Press.

[8]. Fayyad U.M. et al. (1996), "*Data Mining and Knowledge Discovery: making sense out of data* ",IEEE Expert, Vol-11, No-5, pp. 20-25.

[9]. Kasper, Fauci, Hauser et al. (2015). Part-7: Oncology and Hematology. *19th Edition Harrion's Principles of Internal Medicine.* McGraw Hill Education: New York. P-467

[10]. https://www.cancer.org/cancer/cervical-cancer/about/what-is-cervical-cancer.html

[11]. Denny L. (2012) "*Cervical cancer: prevention and treatment.*" Discov Med. 14: PP-125–131.

[12]. Arbyn M, Castellsague X, DeSanjose S, et al. (2011) " Worldwide burden of cervical cancer. Ann Oncol." 22: PP- 2675–2686.

[13]. Yeole BB, Kumar AV, Kurkureet A, Sunny L (2004). "Population-based survival from cancers of breast, cervix and ovary in women in Mumbai." Asian Pac J Cancer Prev. 5:308–315.

[14]. ICO Information Centre on HPV and cancer (Summary Report 2014-08-22).Human Papillomavirus and Related Diseases in India. 2014

[15]. PrashantNaresh, (Aug-2014). "*Early Detection of Lung Cancer Using Neural Network Techniques*", Journal of Engineering Research and Applications, Vol-4, Issue-8.

[16]. Ravi Kumar, G., Ramachandra.A, Nagamani.K, (Aug-2013). "*An Efficient Prediction of Breast Cancer Data Using Data Mining Techniques*", International Journal of Innovations in Engineering and Technology (IJIET) Vol-2, Issue-4.

[17]. K.Srinivas, B. Kavitha Rani and Dr. A. Govrghan (2010), *"Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks"* International Journal on Computer Science and Enginerring, Vol-2,No-2,pp.250-255.

[18]. A. Sudha (March-2012). "Utilization of Data Mining Approaches for Predictin of Life Threatening Diseases Survivability", International Journal of Computer Applications (0975-8887) Vol-41-No. 17.

[19]. Neha Sharma (Sept-2012). "Framework for Early Detection and Prevention of oral Cancer Using Data Mining" International Journal of Advances in Engineering and Information Technology (IJAET) ISSN: 2231-1963, Vol-4, Issue-2, pp. 302-310.

[20]. Blake, C., & Merz, C.J. (2001) UCI Repository of Machine Learning Databases. [Machine-readable data repository]. University of California, Department of Information and Computer Information and Computer Science, Irvine, C.A. [Available from http://www.ics.uci.edu/~mlearn/MLRepository.html]

[21]. Sushmita Mitra, Pabitra Mitra(July-2000). "*Staging of Cervical Cancer with Soft Computing*", IEEE Transactions on BioMedical Engineering, Vol-47, No-7.

[22]. C.McGregor, C.Christina and J. Andrew (2012), " *A process mining driven framework for clinical guideline improvement in critical care*", Learning from Medical Data Streams 13th Conference on Artificial Intelligence in Medicine (LEMEDS). http://ceur-ws.org, vol-765.

[23]. V. Vapnik (1998). "*The support vector method of function estimation*"
[24]. N. Chistianini and J. Shawe-Taylor (2000), "*An Introduction to Support Vector Machines, and other Kernel-based learning methods*", Cambridge University Press.
[25]. Apte & S.M. Weiss (1997), "*Data Mining with Decision Trees and Decision Rules*", T.J. Watson Research Center, http:// www.research.ibm.com/dar/papers/pdf/fgcsapteweissue_with_cover.pdf.
[26]. Kaur, H., and Wasan, S.K. (2006), "*Empirical study on applications of data mining techniques in healthcare*", I.comput. Sci. 2(2), pp. 194-200.
[27]. Bellazzi, R, and Zupan, B.(2008), "*Predictive data mining in clinical medicine: current issues and guidelines*", Int. J. Med. Inform. 77:pp. 81-97.
[28]. Ubeyli, E.D.(2007)," *Comparison of different classification algorithms in clinical decision making*", Expert syst 24(1): pp.17-31.
[29]. Potter, R., (july-2007), "*Comparison of classification algorithms applied to breast cancer diagnosis and prognosis, advances in data mining*", 7th Industrial Conference, ICDM- 2007, Leipzig, Germany, pp.- 40-49.
[30]. Romeo, M., Burden, F., Quinn, M., Wood, B., and McNaughton, D. (1998), "*Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer*", Cell. Mol. Biol. (Noisy –le-Grand, France) 44(1): 179.
[31]. Brickly, M., Shepherd, J. P., and Armstrong, R.A.(1998), "*Neural networks: a new technique for development of decision support systems in dentistry*", J.Dent. 26(4): pp. 305-309.
[32]. Einstein, A. J., Wu, H. S., Sanchez, M., and Gil, J.(1998), "*fractal characterization of chromatin appearance for diagnosis in breast cytology*", J. Pathol. 185(4): pp. 366-381.
[33]. J. Fox (1997), "*Applied Regression Analysis, Linear Models, and Related Methods*".
[34]. U. Fayyad, G. Piatetsky- Shapiro and P. Smyth (1996), "*The KDD process of extracting useful knowledge from volumes of data. Commun.*", ACM, Vol-39, no-11,pp. 27-34.
[35]. Agrawal, R., Imielinski, T., and Swami, A.(1993), "*Mining association rules between sets of items in large databases*" , Proceedings of the ACM SIGMOD International Conference on the Management of Data. ACM, Washington DC, pp. 207-216.
[36]. Agrawal, R., and Srikant, R.(1994), "*Fast algorithm for mining association rules*", Proceedings of the 20th international Conference on Very Large Data Bases (VLDB'94). Morgan Kaufmann, Santiago, pp. 487-499.

