International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-6S2, August 2019

Knowledge Discovery in Clinical Data

Taufik Muchtar, Ismail Suardi Wekke, Muhammad Shuhufi, Abid muhtarom, Phong Thanh Nguyen

Abstract*: The information about patients and their medical condition are stored in a large clinical database. In this data the different patterns and relationship give new medicinal learning. To find this hidden knowledge several methods and techniques are developed. The research proposed a data mining technique in huge clinical database for searching the relationships. This data mining technique is known as Knowledge Discovery in Databases. This research defines different methods and process of data mining in clinical database.

***Keywords* : Clinical databases, methodologies, data query, cleaning, and data analysis, data warehousing.**

I. INTRODUCTION

In medical informatics Knowledge discovery in clinical databases is very major issue to concern [1]. the vast majority of medicinal information like laboratory data and record of patients are stored in a computer, and this clinical database is very large in this way it is very difficult for medical expert to manage such huge databases. So a technique that is based on computer should develop to manage this troublesome circumstance [2].

The process which uses to find the important and useful patterns and relationship from the data is known as Knowledge Discovery in Databases (KDD). Many medical organizations putting away a lot of information about the medical condition of their patient and other information. As there is number of cases increasing day by day, and amount of clinical information has amassed, space specialists utilizing manual examination not very useful and it will be very difficult to be familiar with that data. In the manual investigation of information, Data visualization approach can use. Because a large organization has a huge database so they receive several matches for a simple question so the human factor turns into a bottleneck.

For knowledge discovery data dealing abilities and improved information provide new learning opportunities. In this decade research on knowledge discovery in databases in Interdisciplinary manner has risen. In the health care field the data of pattern recognition should connected with skillful people. Data mining is the process which has automated pattern recognition. for finding the hidden patterns the data mining have several methods that is applied to KDD and it is very difficult to find that patterns with previous measurable strategies. On the basis of how the pattern keeps the inconspicuous cases the patterns are calculated. Data repositories, Databases and data warehouses are getting to be pervasive but it is required skills and knowledge to get the benefits from this stored data [3].

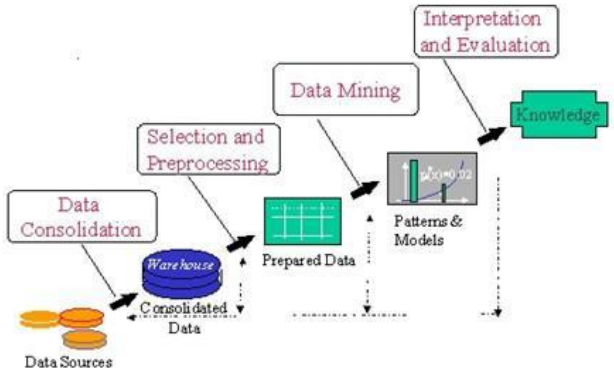

Figure 1: KDD process

II. LITERATURE REVIEW ON CDM

In the zone of data mining, there have been an extraordinary number of studies and concentrates. in uncovering huge clinical examples from patient records and construing already obscure learning [4][5][6] every one of the stages in information mining like Outlier Detection, Clustering, Classification and Feature selection plays important role. The following segments present different reviews and survey on mining clinical data.

2.1Clinical Data Mining

In the biomedical space a few audits on clinical information mining were created during the previous ten years. Several studies are based on the approaches used for clinical data mining and the differences between them. It is not only based on according to the data mining algorithm as discussed in [7], it is also not centered on a particular information mining calculation, for example, in [8]. The study of [9] described that the approaches of data mining on clinical data not only depend on specific type of data like textual data. Paper [10] [11] define the temporal aspect of data. [12] proposed that data mining process not only depend on the information about specific disease or task like pharmaco epidemiology. [13] Presented review on the available data mining tools. In [14] Bellazzi et al define a review of methodological of the predictive data mining in clinical database.

Manuscript published on 30 August 2019.
***** Correspondence Author (s)
Taufik Muchtar, Politeknik ATI Makassar, Indonesia. E-mail: taufik_muchtar@yahoo.co.id
Ismail Suardi Wekke, Sekolah Tinggi Agama Islam Negeri Sorong, Indonesia
Muhammad Shuhufi, Universitas Islam Negeri Alauddin Makassar, Indonesia
Abid Muhtarom, Islamic University of Lamongan, Indonesia
Phong Thanh Nguyen*, Department of Project Management, Ho Chi Minh City Open University, Vietnam. E-mail: phong.nt@ou.edu.vn

III. DATA MINING MODELS IN CDM

For exact and precise basic leadership the Clinical data mining investigation creates viable and advantageous learning [15]. in clinical datasets with bountiful applications Different kinds of mining models have been utilized to define latest trends and pattern [16] [17]. In the following part review of some healthcare data mining models are given

3.1 Models based on Feature Relevance

Clinical data require more consideration in data analysis and storage because Clinical information are commonly voluminous in nature [18] [19]. In data mining Feature relevance analysis is a stage that that empowers scientists to sift through specific indicators of sicknesses from further investigation under the appearance of being less contributory to the location of an ailment [20]. The health record of a patient store the data about patient address, location and ID with the laboratory reports. It shows the condition of patient's health [21].

3.2 Clustering Models

In the area of text domain the Clustering is a generally studied data mining approach. In real world scenarios Clustering has different applications. In the field of data mining it is one of the primary exploratory inquires. The data that is available in documents is directing and summarizing by using Clustering of textual data [22].

3.3 Outlier Detection Models

In 2009 Chandola et al. [23] defined that to translate significant data in medical field outlier detection is very essential. In patient record to detect anomalous patterns outlier detection techniques are used. This record can have important data like indications of any new disease. in 1994 Barnett and Lewis defined that outlier can be described as “an observation, or subsets of observations, appearing to be inconsistent with the remainder of that set of data”. In 2004 Silva [26] proposed that when an element is compared to a standard then it is called outlier.

3.4 Classification Models

To learn a model from a database of named data occasions and, after using learned model it characterize a test case into one of the classes. the anomaly detection method that is depend on Classification work in a comparable two-stage defined by Chandola et al., in 2009 [23]:

1. Training phase this phase study a classifier and use data that is labeled as training data
2. Testing phase this phase using the classifier classifies a test instance as normal or anomalous

3.5 Association Models

it is proposed in [27] that in Association rule(X) Y is characterized over a set of transactions T , the X and Y are consider as sets of items. the set T can be consider as clinical record of patient in a Clinical setting, and the items can be observations, measurements, diagnosis report and symptoms of that patient [27] [28]. S is considering as a set, it described number of transactions in T that contain all members of the set S [29]. In the data set that is mined support the rule (X(Y)) [30].

IV. DATA MINING PROCESSES

To use and recognize the shrouded information in information mining there require three conditions [31].

- The information that is finding from shrouded information ought to have wide view instead of explicit view.
- The information that is incorporated information ought to be separated.
- The information which get from shrouded information ought to be sorted out in the manner so it can use for basic leadership.

The procedure of information mining partitioned in to four stages. The as of now abridged information that find from information distribution center comprise the change of the data. What's more, they use to give the helpful information. The procedure of information mining incorporates after strides as given underneath [4]:

1. Selection of information
2. Transformation of information
3. Mining that information
4. Result understanding

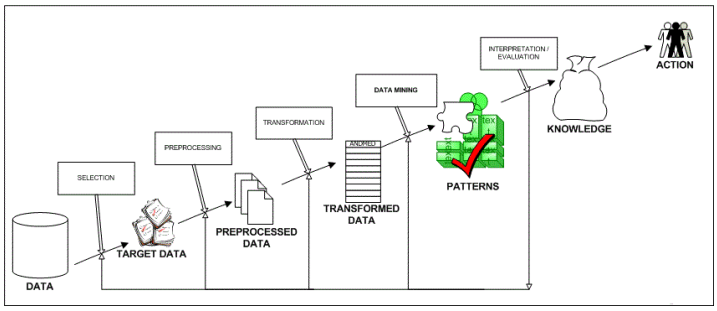

Figure 2: Data Mining process

V. DATA MINING TOOLS

There are several data mining tools are available. Some of them are as follows:

1. SPSS (Statistical Package for the Social Sciences modeler)
2. ANN (Artificial Neural Networks)
3. K-means clustering
4. SNP(Single Nucleotide Polymorphism)
5. RST (Rough Set Theory)

There are many open source data mining tools are also available. Six best tools are:

1. R-Programming
2. Orange
3. Knime
4. NLTK (Natural Language Toolkit)
5. Weka
6. Rapid miner

VI. DATA MINING APPLICATIONS IN HEALTHCARE

There are several applications of data mining exists in health care. The tools of data mining used to predict the health care problem and produce effective results from verified the data. In different healthcare problems data mining tools calculate the accuracy and efficiency on different level. Some medical issue that are calculated and measures using data mining tools are as follows:

1. Cancer
2. Tuberculosis
3. HIV/AIDS
4. Diabetes Mellitus
5. IVF
6. Blood
7. Dengue
8. Kidney dialysis and many more

The table below show different techniques and tools to calculate the accuracy and efficiency in different disease [32].

Table 1: in Healthcare techniques and tools

S No	Type of disease	Data Mining Tools	Data Mining Techniques	Algorithm	Accuracy Level in %
1	Heart Disease	ODND, NCC2	Classification	Naïve	60
2	Cancer	WEKA	Classification	Rules Decision Table	97.77
3	HIV /AIDS	WEKA 3.6	Classification and Association Rule Mining	J48	81.88
4	Blood Bank Sector	WEKA	Classification	J48	89.9
5	Brain Cancer	K-means clustering	Clustering	MAFIA	85
6	Tuberculosis	WEKA	Naïve Bayes Classifier	KNN	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	82.6
8	Kidney Dialysis	RST	Classification	Decision Making	75.97
9	Dengue	SPSS Modeler		C 5.0	80
10	In Vitro Fertilization (IVF)	ANN, RST	Classification		91
11	Hepatitis C	SNP	Information Gain	Decision Rule	73.20

VII. CONCLUSION

The procedure which uses to locate the significant and helpful examples and relationship from the information is known as Knowledge Discovery in Databases (KDD). Numerous medicinal associations securing a great deal of data about the ailment of their patient and other information. the data about patients and their ailment are put away in a huge clinical database. In this information the various examples and relationship give new therapeutic learning. To locate this concealed learning a few strategies and methods are created. The exploration proposed a data mining method in immense clinical database for looking through the connections. This information mining procedure is known as Knowledge Discovery in Databases. This exploration characterizes various strategies and procedure of data mining in clinical database.

ACKNOWLEDGMENT

This article has been written with financial support of RUSA–Phase 2.0 grant sanctioned vide Letter No. F. 24-51/2014-U, Policy (TNMulti-Gen), Dept. of Edn. Govt. of India, Dt. 09.10.2018..

REFERENCES

1. Jonathan C. Prather, David F. Lobach, Linda K. Goodwin, Joseph W. Hales, Marvin L. Hage and W. Edward Hammond, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse.
2. Shusaku Tsumoto, Wojciech Ziarko, Ning Shan and Hiroshi Tanaka, "Knowledge Discovery in Clinical Databases based on Variable Precision Rough Set Model".
3. Abbas Heiat,"Knowledge Discovery and Data Mining in Healthcare: Challenges and Issues"
4. David Hanauer, MD(2007), MS Mining clinical electronic data for research and patient care: Challenges and solutions, Clinical Assistant Professor University of Michigan, USA.
5. Bennett CC and TW Doub. (2010) "Data mining and electronic health records: Selecting optimal clinical treatments in practice". Proceedings of the 6th International Conference on Data Mining. Pages:313-318.
6. M.F. Ochs et al. (eds.)(2010), "Clinical Research Systems and Integration with Medical Systems", Biomedical Informatics for Cancer Research, DOI 10.1007/978-1-4419-5714-6_2, © Springer Science Business Media, LLC 2010.
7. Smyth P(2000),"Data mining: data analysis on a grandscale. In: Statistical Methods in Medical Research", pages:309-327.
8. Patel JL and Goyal RK(2007),"Applications of artificial neu-ral networks in medical science". Curr ClinPharmacol,pp. 217-26.
9. Meystre SM, Savova GK, Kipper-Schuler KC and Hurdle JF(2008),"Extracting information from textualdocuments in the electronic health record: a review of recent research", Yearb Med Inform pp.128-44.
10. Zhou L and Hripcsak G.(2008),"Temporal reasoning with medi-cal data—a review with emphasis on medical natu-ral language processing" J Biomed Inform,vol:40, pages:183-202.
11. Stacey M, McGregor C(2007),"Temporal abstraction inintelligent clinical data analysis: A survey",no.1, pp. 1-24.
12. Hennessy S(2006),"Use of health care databases in pharmacoepidemiology", Basic Clin- PharmacolToxicol.
13. Zupan B and Demsar J(2008), "Open-Source Tools for DataMining. Clinics in Laboratory Medicine", vol:28,pages:37-54.
14. Bellazzi R and Zupan B(2008)"Predictive data mining inclinical medicine: Current issues and guidelines",vol:77,issue:2,pages:81-97.
15. Gregory Piatetsky-Shapiro and Pablo Tamayo,"Microarray Data Mining: Facing the Challenges" SIGKDD Explorations. Vol:5, Issue:2.
16. Weiss and Indurkhya, "Predictive Data Mining", Morgan Kaufmann Publishers.

17. Riccardo Bellazzi and Blaz Zupanb(2008), "Predictive data mining in clinical medicine: Current issues and guidelines"", international journal of medical informatics, pages:81–97.
18. G. Bontempi(2005), "Structural feature selection for wrapper methods". In Proceedings of ESANN 2005, European Symposium on Artificial Neural Networks.
19. Jiang et.al, "Feature Mining Paradigms for Scientific Data", Copyright © by SIAM.
20. Archana Venkataraman, Marek Kubicki, CarlFredrik Westin and Polina Golland(2010), "Robust Feature Selection in Resting-State fMRI Connectivity Based on Population Studies", 978-1-4244-7028-0/10/$26.00 ©2010 IEEE.
21. Taranath N.L., Dr. Shantakumar b Patil, Dr. Prema jyothi Patil and Dr. C.K. Subbaraya, "A Review on Clinical Data Mining".
22. Effat Naaz, Divya Sharma, D Sirisha and Venkatesan M(2015), "Enhanced K-means Clustering Approach for Health Care Analysis Using Clinical Documents", IJPCR, January 2016, Vol:8, Issue:1, pages:60-64.
23. Chandola, V., Banerjee, A., and Kumar, V. (2009)"Anomaly Detection: A Survey",ACM Computing Surveys 41.
24. Laurikkala, J., Juhola, M., and Kentala, E. (2000),"Informal Identification of Outilers in Medical Data", Intelligent Data Analysis in Medicine and Pharmacology.
25. J. Han and M. Kamber(2000), "Data Mining; Concepts and Techniques", Morgan Kaufmann Publishers.
26. Shomona Gracia Jacob and R Geetha Ramani(2012),"Data Mining in Clinical Data Sets: A Review", vol:4,paper no:6.
27. Ogundele I.O, Popoola O.L, Oyesola O.O and Orija K.T(2008),"A Review on Data Mining in Healthcare",Vol:7, Issue:9, international Journal of Advanced Research in Computer Engineering & Technology (IJARCET).
28. D.Usha Rani(2017),"A Survey on Data Mining Tools and Techniques in Medical Field",Vol:8,Issue:5,Pages: 51-54,Special Issue.