ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

A Survey on Knowledge Discovery of Healthcare Dataset using Graph based approach.

M S Saravanan[1]*, and R Sai Manoj Kumar[2].

[1]Professor, Department of CSE & IT, Saveetha School of Engineering, Saveetha University, Chennai, India.
[2]II CSE – B, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha University, Chennai, India.

ABSTRACT

In this paper, we are going to discuss about Healthcare Dataset using Graph based approach. In recent technology innovation we have lot of facilities are emerging to solve the Healthcare issues of an individual. The massive dataset of healthcare is not able to handle to take decision on patient by doctors. Hence the knowledge discovery on healthcare dataset will be represented with graph based approach. The graph based approach can be easily understood by the doctors to predict the further treatment process of a patient. In this paper we surveyed nearly twenty five research findings and presented some interesting observations for the future researchers.

Keywords - Healthcare, Dataset, Graph, Patient, Doctors.

Corresponding author

ISSN: 0975-8585

INTRODUCTION

Healthcare is the maintenance or improvement of health via the diagnosis, treatment, and prevention of disease, illness, injury, and other physical and mental impairments in human beings. Health care is delivered by health professionals in allied health professions, chiropractic, physicians, physician associates, dentistry, midwifery, nursing, medicine, optometry, pharmacy, psychology, and other health professions. It includes the work done in providing primary care, secondary care, and tertiary care, as well as in public health. Today the level of health care has excelled tremendously. Presently the goal of health care is to have a continuum of care for the patient, one which is integrated on all levels.

Healthcare has ended up one of India's biggest segments - both regarding income and vocation. Healthcare includes doctor's facilities, therapeutic gadgets, clinical trials, outsourcing, telemedicine, restorative tourism, medical coverage and therapeutic gear. The Indian Healthcare services segment is developing at a lively pace.

Healthcare Industry has generated large amount of data, by keeping records, compliance & regulatory requirements, and patient care. Most of the data is stored in hard copy form the current trend is toward rapid digitization of large amount of data. Mandatory requirements are used to improve the quality of healthcare delivery and also to reduce the costs. This massive quantity of data holds the wide range of medical and healthcare functions, including other clinical and disease surveillance.

Literature Survey

The effectiveness and efficient graph-based semi-supervised algorithm namely SHG will meet the lot of challenges in the healthcare services [1]. This reference proposed a new graph-based classification approach method on mining health examination records. It has a few significant advantages. First, Health examination records are represented as a graph that associates all relevant cases together. This is particularly helpful for displaying anomalous results that are frequently meagre. Second, multi-wrote connections of information things can be captured and actually mapped into a heterogeneous diagram. Health examination items are represented as different types of nodes on a graph, which enables our method to exploit the hidden graph structures of individual classes to achieve higher performance.

The various studies on the healthcare services can fight against diabetes, sharing and benchmarking diabetes care is essential to influence health policy and improve outcomes and quality of life for people with diabetes. As more and more data is collected, Diabetes measurement will become an even more powerful resource for inspiring and driving change in diabetes care. The fundamental goal of the Diabetes measurement is to measure, share, and improve diabetes outcomes. There are so many ways to get involved in reversing diabetes trends, from collecting data to sharing better practice models to improving public visibility and advocating for the quality of diabetes care at the global, national, clinic, and patient levels. [2]

Discovery of Healthcare Datasets using Data Mining tools

Due to exponential growth of data, it is unstructured in nature and hence it takes a process and method to extract useful information from the data and transform it into understandable and usable form [3]. This is where data mining comes into picture. Plenty of tools are available for data mining tasks using artificial intelligence, machine learning and other techniques to extract data. In this paper we have discussed the popular data mining tools such as RapidMiner, WEKA, R-Programmer, Orange and KNIME [5].

Rapid Miner

It is written in the Java Programming language, offers advanced analytics through template based frameworks. Users hardly have to write any code. Offered as a service, rather than a piece of local software, this tool holds top position on the list of data mining tools. In addition to data mining, RapidMiner also provides functionality like data preprocessing and visualization, predictive analytics and statistical modelling, evaluation, and deployment. What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts [6].

Waikato Environment for Knowledge Analysis (WEKA)

The original non-Java version of WEKA primarily was developed for analysing data from the agricultural domain [7]. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modelling. Its free under the GNU General Public License, which is a big plus compared to RapidMiner, because users can customize it however they please. WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection. WEKA would be more powerful with the addition of sequence modelling, which currently is not included.

R-Programming

What if I tell you that Project R, a GNU project, is written in R itself? It's primarily written in C and Fortran. And a lot of its modules are written in R itself. It's a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity substantially in recent years. Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, and others [8].

Orange

Python is picking up in popularity because it's simple and easy to learn yet powerful. Hence, when it comes to looking for a tool for your work and you are a Python developer, look no further than Orange, a Python-based, powerful and open source tool for both novices and experts. You will fall in love with this tool's visual programming and Python scripting. It also has components for machine learning, add-ons for bioinformatics and text mining. It's packed with features for data analytics.

KNIME

Data processing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis. Written in Java and based on Eclipse, KNIME is easy to extend and to add plugins. Additional functionalities can be added on the go. Plenty of data integration modules are already included in the core version.

A survey on Data Mining approaches for Healthcare

The purpose of this section is to provide an insight towards requirements of health domain and about suitable choice of available technique. Following are the guideline for using different data mining techniques.

Before applying classification technique there is a need to recognize the redundant and inappropriate attributes because these attributes act as a noise and outlier which in turn slow down the processing task. These attributes also had an adverse effect on the performance of classifier. Statistical methods are used for recognizing these attributes. On the other hand the most relevant and useful attributes can be recognized by feature selection methods which in turn enhance the performance and accuracy of classification model.

Big data analytics has the potential to transform the way healthcare providers use sophisticated technologies to gain insight from their clinical and other data repositories and make informed decisions. In the future we'll see the rapid, widespread implementation and use of big data analytics across the healthcare organization and the healthcare industry. To that end, the several challenges highlighted above, must be addressed. As big data analytics becomes more mainstream, issues such as guaranteeing privacy, safeguarding security, establishing standards and governance, and continually improving the tools and technologies will garner attention. Big data analytics and applications in healthcare are at a nascent stage of development, but rapid advances in platforms and tools can accelerate their maturing process.

ISSN: 0975-8585

The design of the system requires the complete understanding of the problem domain. As per the above discussion there is need to efficiently diagnosis the presence of heart disease in an individual. The main objective of this review is to build Intelligent Heart Disease Prediction System that gives diagnosis of heart disease i.e. Heart Attack using historical heart database. Originally 13 attributes were used for Heart Disease Prediction but the same work can also be achieved with the help of less number of attributes. So the required efforts will be reduced. By considering patients basic information and other attributes value, it will be beneficial to improve the accuracy of existing algorithm approaches. It is also possible to classify the diagnosed results into four categories such as Healthy, Mild Attack, Moderate Attack and Severe Attack. More ever the results may also be stored in the database connected with the system for calculating the efficiency and for the record maintenance of the patients. A unique Patient ID can be generated by the system for each patient who may play important role in overall processing of the system. We are having several options for data mining classification techniques namely Neural Networks, Decision Trees, and Naive Bayes can be used as Classifiers.

Operations of Healthcare Services

The healthcare services are starting with situational analysis used to analysis the patient condition and first aid diagnosis with all type of pre testing process, then the duration of the treatment will be decided with Gap analysis and here fixed the goal with various treatment levels of process, then the future remedy will be provided with Future State process and here the shaping the new process for long time follow up will be suggested to the patient care, then the implementing the planning will be allowed to enable the change for some special cases and finally the Control and Measure is the state of approach used to tack the success of the treatment for future analysis and prediction for some seasonal diseases. The process transformation of treatment process is shown in the Figure 1.

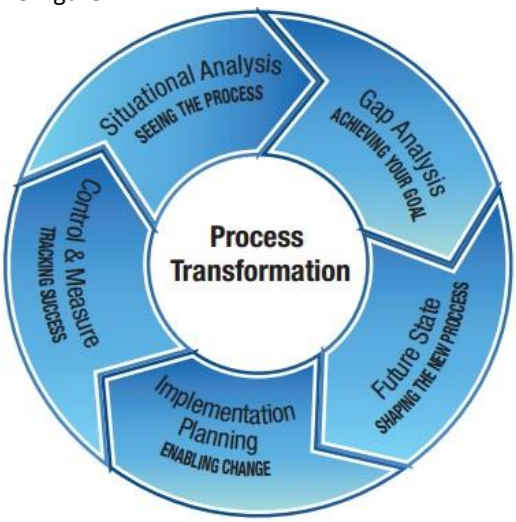

Figure 1. Healthcare Treatment Process Transformation with various States of Approach

Healthcare data Representation and Analysis

The healthcare data are stored in various formats after the data collection from the healthcare centres and also the collected data are needed to be filtered to remove the noise. The data collected from the healthcare centres are unstructured in nature; therefore the data need to be converted into structured. One of the sample structured healthcare data is shown in Table 1.

In the table 1, the healthcare data is shown with the twenty six thousand and seven hundred and seventy one records various patients and collected seventy three thousand six hundred and forty two records with fifty five various test cases, twenty six mental cases and fifty five profiles of the same.

ISSN: 0975-8585

Table 1. Healthcare data Representation in the form of Table

Dataset		# people	# nodes					# links to Record nodes				Density
			Record	Test	Mental	Profile	Total	Test	Mental	Profile	Total	
Real	GHE@10class	26,771	73,642	55	26	55	73,778	601,062	119,952	523,387	1,244,401	0.1242
Synthetic	(100,100)	1,100	3,013	55	26	55	3,149	28,071	4,611	22,552	55,234	0.1348
	(300,300)	3,300	9,054				9,190	84,052	13,982	68,053	166,087	0.1349
	(500,500)	5,500	15,092				15,228	139,736	23,134	113,231	276,101	0.1345
	(1000,1000)	11,000	30,201				30,337	280,412	46,655	227,932	554,999	0.1351
	(1000,3000)	13,000	35,463				35,599	323,101	55,263	265,067	643,431	0.1334
	(1000,5000)	15,000	40,695				40,831	365,005	63,974	301,661	730,640	0.1320
	(1000,10000)	20,000	53,674				53,810	469,575	85,167	393,486	948,228	0.1299
	(1000,15000)	25,000	66,841				66,977	575,360	106,694	485,914	1,167,968	0.1285
	(1000,20000)	30,000	79,979				80,115	682,022	128,357	579,269	1,389,648	0.1278

So for the above set of data types the links between the records need to be identified, hence it is not possible to represent in the form text and summary writing, so an alternative method of Graph Based approach is required to predict the process of treatment for a particular patient. The following section provides the Graph based representation of the above table with various decisions.

Knowledge Discovery of Healthcare Dataset using Graph based approach

Healthcare doctors require a great deal of data to make their wellbeing related exercises and practices with medication solutions which can cure patient's illness. Every day doctors are seeing loads of patients who are experiencing distinctive sorts of sicknesses. By confirming records they are giving sure medicines or drugs, Instead of utilizing records for distinguishing maladies they can utilize chart which will be simple for specialists to recognize specific illness. The Example Healthcare process for the dataset for the Continuum healthcare alliance is shown in the figure 2.

Figure 2. A Sample Healthcare Process from the Continuum Healthcare Alliance

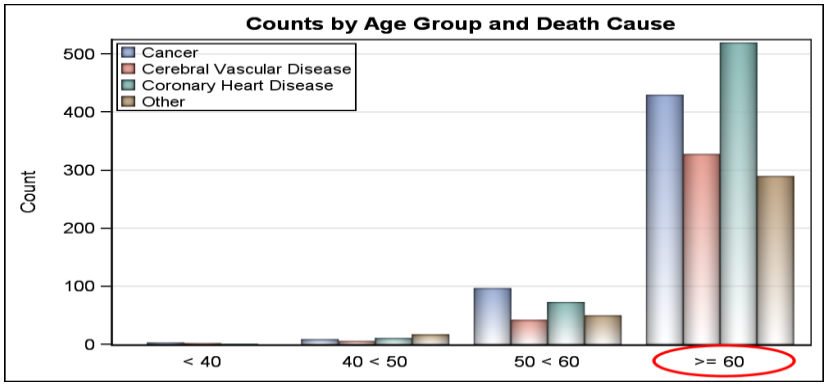

Figure 3. Graph Based Approach for Healthcare Dataset

The figure 3 shows that how many persons are affected with diseases at certain age with the relationship is represented using graph based approach. If age is less than 40 their number of disease affected person count is 0 to 10. If their age is in between 40 to 50 then the number of people affected by cancer is 10 and numbers of people affected by Cerebral vascular disease is 5 and the number of people affected by coronary disease is 15 and 20 people are affected by other diseases. If their age is in between 50 to 60, the number of people affected by cancer is 100 and numbers of people affected by Cerebral vascular disease is 40 and the number of people affected by coronary disease is 70 and 50 people are affected by other diseases. If there is greater than or equal to 60 then the number of people affected by cancer is 430 and numbers of people affected by cerebral vascular disease 530 and the number of people affected by coronary disease is 520 and 290 people are affected by other diseases.

In this paper we observed that Healthcare system is rapidly developing. Doctors are using Graph based method to identify the illness. For the future generation we are introducing new type of Healthcare dataset which will be helpful for them.

CONCLUSION

The reason is to give an understanding towards Health care Dataset using graph based approach. Healthcare Datasets are represented as a graph that associates all relevant cases together. This is especially useful for modelling abnormal results that are often sparse. This work shows a new way of predicting risks for patients based on their annual Healthcare Data. By integrating data from multiple available information sources, more effective prediction may be achieved.

REFERENCES

[1] L. Chen *et al*., "Mining Health Examination Records—A Graph-Base Approach," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 9, pp.2423-2437, Sept.12016.doi: 10.1109/TKDE.2016.2561278.

[2] Ezaz Ahmed, Y.K. Mathur, Varun Kumar. " Knowledge Discovery in HealthCare Datasets Using Data Mining Tools" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 3, No.4, 2012.

[3] Wael Ahmad AlZoubi , "Mining Medical Databases Using Graph based Association Rules" International Journal of Machine Learning and Computing, Vol. 3, No. 3, June 2013.

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

[5] K. Nachimuthu , "Extracting Medical Health Records in a Graph Based Approach" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391.

[6] Gondkar Mayura D, Pawar Suvarna E, "A Survey On Data Mining Techniques To Find Out Type Of Heart Attack" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. V (Jan. 2014), PP 01-05.

[7] M.S.Saravanan, Shafiya Banu, "Building Private Cloud Infrastructure and Related Issues for Healthcare System", Published in International Journal of Applied Engineering Research by Research India Publications, India, Vol.10, Issue.4, March'2015, pp.3040-3045, ISSN:0973-4562.

[8] J. M. Wei, S. Q. Wang, and X. J. Yuan, "Ensemble rough hyper cuboid approach for classifying cancers," IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 3, pp. 381–391, 2010.