LITERATURE REVIEW: BEST KNOWLEDGE ENGINEERING PRACTICES FOR PREDICTION OF HEART DISEASES

Vidwan Arun Chandra Pasumarthi
(5763903)
ISIT919 Assignment 1

Abstract:

In this paper, I have intended to stress the focus on literature available on knowledge engineering in health care. The knowledge-based systems are growing in nature and magnitude in various fields but its development in health care is in a real slow pace. So, systematic literature is conducted in this field by studying academic articles. To narrow down the review, we have selected Cardiology related discipline and also the mortality is high due to heart diseases according to the World Health Organisation (WHO). The main challenges in the medical field are due to the legacy data which is available from ages. The best techniques to study and clean the data are addressed in this paper which is observed from the literature review. The data mining techniques, objectives, and machine learning algorithms are addressed as per their advantages in the medical field. The objectives and techniques vary according to the type of disease, type of data available and type of treatment. In this paper, all three categories are explained based on heart-related issues.

Keywords: challenges, knowledge engineering, data mining, heart diseases.

Introduction:

As the global IT modernization has been taken place for the last two decades, the use of Information Technology techniques and methods has been increased exponentially. As a result, a lot of raw data has been accumulated in the storage systems. This led to the development of many databases with enormous features. yet the data in those systems remains raw which means can't be used for many applications. This raw data has to be converted into useful information which can be applied in various fields and findings. The extraction of useful information from the raw data and using as knowledge is known as knowledge discovery. From the literature found: "*knowledge discovery is the non-trivial extraction of implicit, previously unknown and potentially useful information from data*" (Esfandiari et al., 2014). Data mining can be considered as a mathematical strategy of knowledge discovery (Banaee, Ahmed and Loutfi, 2013). In the 21st century, the use of data mining techniques has been predominantly increased across many fields such as education, fraud detection in financial services, medical fields, prediction in stock markets, etc… (Esfandiari et al., 2014)

The medical field is one undeniable and most prominent sector in human life. Using the concept of knowledge discovery with data mining techniques has been escalated in this field since the last decade. Many hospitals do have traditional Information Systems which

could have a record of patients, conditions, and symptoms. But these systems lack in decision making and in the prediction of the similar condition (Palaniappan s., 2008). These support systems can't find the predictors of the situations or conditions causing the diseases (Palaniappan s., 2008). The data consolidated over the centuries in this field can be used to find predictors of many chronical diseases like cancer, heart disease, diabetes, etc… but it can't be humanly processed. Moreover, the data present in the medical fields are considered as high in volume, too noisy, inconsistent and imbalanced (Benhar, Idri, and Fernández-Alemán, 2018) and also contains too many outliers (Idri et al., 2018). To overcome these problems, data preparation has to be done before applying data mining techniques. This preparation includes (i)data cleaning in which missing data will be replaced with either 0 or with the average frequency values and also eliminate outliers. It includes (ii)data transformation in which the data will be converted as required for the model (data mining) we are training. It includes (iii)data reduction which reduces the dimensionality of the data. It includes (iv)data balancing in which the data will be classified into classes by eliminating outliers and extra features (Benhar, Idri, and Fernández-Alemán, 2018). Application of Knowledge Discovery of Data (KDD) may vary from disease to disease based on the mortality rate and severity of the problem. The aim of the KDD is to decrease the treatment cost and time and increase efficiency by providing accurate data for the decision-making systems.

From the stats observed, heart diseases are becoming the sole reason for too many deaths around the globe. According to the World Health Organisation (WHO), around 12 million people die due to heart diseases and this count is predicted to escalate to 23.6 million by 2030 (Bindu D.C, 2017). In China Coronary Heart Disease (CHD) is one of the leading causes of deaths (Xing, Y., Wang, J., 2007). In the U.S. one person dies due to cordial diseases in every 34 seconds and this is one of the leading cause of deaths in many countries (Soni et al., 2011). Introducing Knowledge discovery concepts in case of cardiology problems could really be helpful for the physicians and doctors to estimate the situation and identify the risk of recurrence. It also helps in building automated systems and decision-making systems which are helpful in the absence of expertise (Soni et al., 2011). Speaking to the core, the machine learning models have to be trained with previous data so that they can analyze, predict, classify when the new is data collected. Not only through hospitals even there are wearable sensors which collect the data from your body. The signs like blood pressure, sugar levels, oxygen saturation levels, respiration rate, heartbeat rate are collected and processed to diagnose any heart-related issues (Banaee, Ahmed and Loutfi, 2013). The entire treatment for any kind of disease can be divided into phases like screening, diagnosis, treatment, prognosis, monitoring and treatment (Esfandiari et al., 2014). The main objectives of data mining are classification, clustering, regression, and association. These objectives imply at different phases of the treatment. There are many algorithms of machine learning that are used at a specific phase of the treatment and for a specific purpose. The most common among them are support vector machines, decision trees, artificial neural networks. Each of these algorithms works to fulfill different objectives. For example, support vector machines (SVM) are used for classification, Artificial neural networks (ANN) are used for prediction, decision trees are used to predict possible outcomes. Using all these techniques together we can diagnose and can possibly control the deaths related to heart diseases.

The following paper is comprised of three sections. The immediate section talks about the search strategy and the keywords used. The later section contains results which I got through the comparison of results from papers collected. In this section, I am going to compare the most used algorithms and accuracy rate of the models in treating heart diseases.

Research methodology or strategy:

The methodology I have selected is the literature review. The key concepts in this paper do cover health care, data mining, challenges, and knowledge engineering.

The keywords used are challenges, data mining, knowledge engineering, and heart diseases. Putting all the keywords together would end up in a large volume of papers. To avoid it, I have used quotation marks for the keywords and also concatenation.

So, the keywords are framed as "knowledge engineering" + "heart diseases", "knowledge engineering" + "challenges" and similar forms.

The databases selected are science direct, Scopus, Springer, IEEE. Also used google scholar to redirect back to the original databases. All the search is done with a limitation implied to last 10 years.

Databases	Keywords combinations	Total selected
Science Direct	KE + heart diseases	3
Scopus	Challenges + KE	2
Springer	KE + health care	1
IEEE	Data mining + healthcare	4

Results:

I am going to compare the results from different papers collected. A comparison table between the machine learning algorithms used by the authors in different cases is formulated pointing the best algorithm according to the results. The second portion of this section will be the challenges in implementing these knowledge engineering techniques in health care or in a medical environment.

NOTE: All these result comparisons are related to the heart diseases in the medicine. The algorithms and results vary for other chronical diseases.

To make it clearer about the algorithms, here is the short description of the techniques.

The data mining techniques are broadly classified into supervised and unsupervised learning. In supervised learning, the output is known, and we try to form a relationship between the input and output, also objectives like classification, regression come under this category. In unsupervised learning output is unknown and objectives like clustering, association rules come under this category. Usually, the type of algorithm used in medical fields is a genetic algorithm as the medical field is divided into diagnosis, signal processing, imaging, etc… there are specific algorithms to achieve specific objectives. These algorithms are used to train and deploy the data mining models using the available data which is divided into a train set and test set.

The following is the comparison table:

Papers referred	Data pre-processing. (winning)	Training model		
		Data mining objective. (winning)	Data mining technique(winning)	
			Having heart disease	Prediction through intelligent systems (IHDPS). no heart disease
Benhar, Idri and Fernández-Alemán, 2018	Reduction	Classification	✓	X
Idri et al., 2018	Reduction	X	X	X
Bindu D.C, 2017	X	Classification	Naïve Bayes	X
Kadi, Idri and Fernandez-Aleman, 2017)	X	Classification	SVM	X
AbuKhousa, E. and Campbell, P., 2012	X	Classification	Naïve Bayes	Decision tree
Palaniappan s., 2008	X	Classification	Naïve Bayes	Decision tree
Xing, Y., Wang, J., 2007	X	Classification	SVM	X

Table 1: Collective information on DM Objectives and techniques used in heart-related issues.

These results are the collective information from the papers collected on data mining techniques of heart-related diseases.

To build a knowledge-based system, the process has to go through three phases: data pre-processing, training model and data post-processing. The final data can predict the occurrence or risk factors of having heart diseases. In the medical field, data preparation is more commonly used in cardiology discipline which stands by 35% (Benhar, Idri, and Fernández-Alemán, 2018) and comes under data pre-processing. In the data preparation step, data reduction is the most used process. It is stated in the paper (Benhar, Idri, and Fernández-Alemán, 2018) that from the literature collected from 97 articles, data reduction is the most

selected technique in data preparation by 67%. The data reduction has the advantage over feature selection when compared to other preparation techniques (Idri et al., 2018). It is very clear from the table that classification is the most suited data mining objective to structure the data related to cardiology in the medical field. There are studies that are related to the medical field which use prediction algorithms as well. Naïve Bayes is the most used classification algorithm to classify the data and analyze it. Support vector machine algorithm comes second in place. Both these algorithms are good at classifying the type of heart disease. Decision trees algorithm is used in building Intelligent Heart Disease Prediction Systems (IHDPS). These systems use the symptoms of the disease and predict the chances of occurring heart diseases. Symptoms are collected from the medical data.

Challenges in implementing knowledge-based systems in the medical field:

As I have mentioned in the introduction, the data in this field is enormous and most imbalanced. There can possibly be many enormous missing values as the data is collected from previous medical records, hospitals, through interviews and patient records. The main challenge in the medical field is that the data should be accurate, or the output will go terribly wrong. So, in the medical field most challenges are in data pre-processing level as the data needs to be cleaned up. The medical data challenges are categorized mainly into domain challenges, general challenges and collection challenges (Esfandiari et al., 2014). Domain challenges can be like mathematical disruption, may or may not be legal or confidentiality issues. General challenges are the messy data and the collection challenges are due to the inconsistent standards in the data (Esfandiari et al., 2014). One another challenge for data mining in the medical field is that the dataset collected to train the model in most cases is public data set which may not be legible for the accuracy of the model. Few other challenges are with the data mining models. They have their own error margins. Few models can outperform the other for similar dataset. So, we have to choose the models according to the requirements.

Discussion and conclusion:

As mentioned earlier, diseases related to cardiology are the most severe when compared to other chronicle diseases. To deal with the case, a new system has to be evolved to overcome the mortality rate. Also, the traditional informational systems in the hospitals need an upgrade as the mortality rate regarding the chronicle diseases kept increasing for the last few years. The data in the medical field is growing in volume as well. The requirement for doctors and physicians has also been increased. So, basically, to use the data available and fill the gap in the shortage of physicians and doctors, we need decision-making systems. These systems have to assist in emergencies or during lack of expertise. Knowledge engineering is one of the bases on which expert systems and decision-making systems can be built. To apply knowledge engineering in the medical field the data mining techniques are required as it the part of artificial intelligence. The challenges in utilizing the legacy medical data should be addressed by the data pre-processing techniques and the right models have to train as the margin error in the medical field is very narrow. Another challenge is selecting the right algorithm for the right purpose. All these challenges have to be overcome by building proper and efficient knowledge-based or decision-based systems using knowledge engineering. The most used objectives and algorithms in predicting heart diseases are identified and listed in the results section.

Reference:

AbuKhousa, E. and Campbell, P., 2012, March. Predictive data mining to support clinical decisions: An overview of heart disease prediction systems. In *2012 International Conference on Innovations in Information Technology (IIT)* (pp. 267-272). IEEE.

Banaee, H., Ahmed, M. and Loutfi, A. (2013). Data Mining for Wearable Sensors in Health Monitoring Systems: A Review of Recent Trends and Challenges. *Sensors*, 13(12), pp.17472-17500

Benhar, H., Idri, A. and Fernández-Alemán, J. (2018). A Systematic Mapping Study of Data Preparation in Heart Disease Knowledge Discovery. *Journal of Medical Systems*, 43(1).

Bindushree, D.C. and Rani, V.U., 2017, August. A review of using various DM techniques for evaluation of performance and analysis of heart disease prediction. In *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)* (pp. 686-690). IEEE.

Esfandiari, N., Babavalian, M., Moghadam, A. and Tabar, V. (2014). Knowledge discovery in medicine: Current issue and future trend. *Expert Systems with Applications*, 41(9), pp.4434-4463

Idri, A., Benhar, H., Fernández-Alemán, J. and Kadi, I. (2018). A systematic map of medical data preprocessing in knowledge discovery. *Computer Methods and Programs in Biomedicine*, 162, pp.69-85.

Kadi, I., Idri, A. and Fernandez-Aleman, J. (2017). Knowledge discovery in cardiology: A systematic literature review. *International Journal of Medical Informatics*, 97, pp.12-32.

Palaniappan, S. and Awang, R., 2008, March. Intelligent heart disease prediction system using data mining techniques. In *2008 IEEE/ACS international conference on computer systems and applications* (pp. 108-115). IEEE.

Soni, J., Ansari, U., Sharma, D., and Soni, S. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, 17(8), pp.43-48

Xing, Y., Wang, J. and Zhao, Z., 2007, November. Combination data mining methods with new medical data to predicting the outcome of coronary heart disease. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)* (pp. 868-872). IEEE.