International Journal of Trend in Scientific Research and Development (IJTSRD)
International Open Access Journal | www.ijtsrd.com

ISSN No: 2456 - 6470 | Volume - 2 | Issue – 6 | Sep – Oct 2018

A Survey on Various Disease Prediction Techniques

C. Leancy Jannet[1], G. Sumalatha[2]
[1]Student, [2]Assistant Professor
Department of CT, Sri Krishna College of Arts and Science, Coimbatore

ABSTRACT

An analysis of various diseases have been predicted using multiple data mining and text mining techniques. In this article we are going to discuss about 6 prediction techniques. Using gene expression pattern we predict the disease outcome and implementation of pathway based approach for classifying disease based on hyper box principles, we also present a novel hybrid prediction model with missing value imputation (HPM-MI) which analyze imputation using simple k-means clustering. A technique based on CCAR (Constraint Class Association Rule) has been used for reducing time consumption in prediction of a particular disease. We have discussed about text mining technique and their applications. Another technique has also been studied about hyper triglyceride mia from anthropometric measures which diverge according to age and gender. Using multilayer classifiers for disease prediction we can achieve high diagnosis accuracy and high performance.

Keyword: *Prediction, Genes, Data Mining, Text Mining, Hyper triglyceride mi a, Missing Values, Hmv and Classifiers*

1. INTRODUCTION

In this article we are going to discuss about 6 techniques for predicting disease. In the first technique diseases can be predicted by gene expression pattern. For selective distinctive genes feature selection algorithm is enrolled. For classification 2 approaches are used network based approaches and pathway based approach and through hyper box representation diseases are classified [1]. The second technique is predicting the disease using MVI we first evaluate 11 missing data imputation techniques practically and then find the perfect method for grasping missing values from dataset using k-means clustering[2]. The third prediction technique is mining CARs (class association rules) which locates relationship among item sets and class labels. These item set constraints minimize CARs and lessen the search space and enhance the performance, first a tree structure is initiated for efficient mining, second 2 theorems for soon trimming rare item set, lastly efficient and fast algorithm for mining CARs 2 approaches are used pre processing and post processing [3]. The fourth technique is text mining which help researchers in assessing scientific literature. Information can be withdrawn using co-occurences based method and NLP-based methods and text mining tools are discussed [4]. Fifth technique is hyper triglyceride mia by anthropometric measures based on data mining. Many diseases can be predicted by the change in hyper triglyceride mia it varies according to age and gender [6]. Sixth and final technique is multilayer classifier for disease prediction. A huge number of prediction models can be build from data mining techniques. Here the concept of machine learning is extended. Machine learning are additionally classified into supervised and unsupervised learning.

2. RELATED WORK

2.1. Pathway Based Approach:

Through gene expression diseases can be predicted where samples of genes are sketched as specimen of different disease states for understanding the disease phenotype. The computation of gene expression estimate disease outcome. By using feature selection algorithm we can pick a batch of genes which are differently expressed. In network based approaches disease module based methods suspect that all cellular element that are owned to the identical topological, working or disease module have a inflated chance of having same disease. Greedy search is executed over a PIN (People of Information Network) to pinpoint a number of gene modules whose mean countenance is supreme. PIN data is usually undependable and clattering and it also has a high false positive rate. In pathway based approaches biological pathways are unreliable and organised troupe of molecular interaction network. Pathway level disease classification approach based on hyper-box principles where given a microarray gene expression portrait and a number of biological pathways/gene set the classification exactness of each pathway/gene sets is assessed by using only the adherent genes in the pathway. By using psoriasis and breast cancer datasets prediction accuracies are greater than 85% superior than sets of genes that are selected unplanned. Hyper box disease classification (Hyper DC) analyze disease specimen accurately when comparing with other prominent classification methodologies considering to SNR classification rates. Hyper DC show inflated SNR. Actual strength of Hyper DC rely on illustrative power. Main advantage of Hyper DC is it is flexible [1].

2.2. Hybrid Prediction Model with Missing Value Imputation:

Exact prediction in huge number of missing values in dataset is a tough task. Many hybrid models to face this problem they have removed the missing instances from dataset which is commercially known as case deletion. Hybrid prediction model with Missing Value Imputation (HPM-MI) scrutinize different imputation technique by applying simple k-means clustering and use the best one to dataset. Missing values happen because of many reasons due to mistake in manual data, equipment mistakes, or inaccurate measurements. Missing values can cause many problems like loss of efficiency, convolutions in managing and examining data. It may also lead to bias decisions. We have analyzed 11 Missing Value Imputation techniques they are case deletion, Most Common Method (MC), Concept Most Common (CMC), K-Nearest Neighbor (KNNI), Weighted Imputation with K-Nearest Neighbor (WKNN). K-Means Clustering Imputation (KMI), Imputation with Fuzzy K-Means Clustering (FKMI), Support Vector Machines Imputations (SVMI), Singular Value Decomposition Imputation (SVDI), Local Least Squares Imputation (LLSI), Matrix Factorisation, Selecting best MVI method is based on accuracy it is attained. We have selected clustering as beginning for selection of best imputation technique. K-Means Clustering seperates datasets into groups so that instances in one set are similar to each other and as dissimilar as possible from the objects in other groups. CMC method has given less number of wrongly classified instances in diabetes as well as in hepatitis dataset. Case deletion is best in hepatitis dataset [2].

2.3. CCAR: With Item set Constraints:

Class Association Rules (CAR's) are used to construct a classification model for prediction and narrate association between item set and class labels. The initiation of mining association rules with item set constraints,3 major approach has been proposed. First method is post-processing method, first finds repeated item sets using a priori algorithm, second approach is pre-processing method which filters out the ones which do not convince the item set constraint, the third approach is constrained item set filtering which tries to assimilate the item set constraints into original mining process because it needs to generate only the frequent item set that pleases the constraints .In post-processing approach CAR-mining algorithm is used. This strategy is not very efficient because all CAR's must be generated and frequently a large number of candidate CAR's need to be tested. Advantage of pre-processing approach is the size of the filtered dataset is very much lesser than the actual dataset. So the mining time can be notably reduced. The authors manifested that their method is higher than the post-processing method. In our proposed method the item set constraint is thrust as intense "inside" estimation as possible. Instead of using all rules, only rules which please the item set constraints are formed to speed up the process. Proposed algorithm for mining CAR's is tree structure which includes only nodes having constrained item set and two theorems for soon trimming infrequent item sets [3].

2.4. TEXT MINING:

One of the major dominant entry point to scientific literature sources for biomedical research is pub Med which give entry to more than 24 million scientific literature. Fetching of relevant information from literature database fusing these information with experimental output is time taking and it also requires some careful attention so text mining is introduced. Text mining reply to many research questions extending from the discovery of drug targets to drug repositioning. Definition of text mining by Marti Hearst is "the discovery by computer of new, previously unknown information unknown information from different written resources, to reveal otherwise 'hidden' meanings". The first and the foremost step in text mining is to fetch suitable textual resources for a given subject of interest. This process is called as information retrieval. After IR the resulting document set can be examined by search algorithms for the occurrence of particular keywords of interest. For example a particular gene should be acknowledged in the text not only by its gene symbol, but also by the synonyms and previous names this is called as NER concept. After IR and NER technical algorithms can be used to discover links between concepts in the text. Recently the most used approaches to extract information from text are co-occurrence based methods and Natural Processing (NLP) based methods. Comparison of these 2 approaches NLP based method has more advantages. Some of the applications of TM to biomedical problems are genome and gene expression annotation, drug repositioning, adverse events, electronic health records, domain specific databases[4].Proteins are the molecules that ease most biological process in a cell. Some of the text mining tools are Bio RAT (Biological Research Assistant for Text Mining, eFIP (Extracting Functional impact of phosphorylati on), FACTA+(Finding Associated Conccepts with text analysis), Gene Ways, Hit Predict, In Print, I2D (Inter logo us Interaction Database), iHOP (information Hyperlinked Over Project), IMID(Integrated Molecular Interaction Database), Negatome, open DMAP, PCorral (Protein Corral), PIE the search, Poly search, PPIExtractor, PPIfinder, PPLook, STRING[5].

2.5 Hyper triglyceride mi a From Anthropometric Measures:

The excellent indicator for the prediction of hyper triglyceride mi a from anthropometric measures of body shape which remains a matter of debate. A total of 5517 subjects have participated in this cross-sectional study (3675 women and 1842 men) aged 20-90 years. When the subject is standing the circumference of 8 particular sites were measured using a flexible non-elastic tape. The BMI was calculated as weight in kilograms divided by square of the height in meters. Various circumference measurements are Forehead Circumference (FC), Neck Circumference (NC), Axilliary Circumference (AC), Chest Circumference (CC), Pelvic Circumference (PC), Rib Circumference (RC), Waist Circumference (WC), Hip Circumference (HC). Male and female data are divided seperately because the difference in body shape with aging may vary according to sex. Waist to Hip Ratio (WHR) were the strongest predictors of hyper triglyceride mi a in Indian men. WHtR (Waist to Height Ratio) was the best predictor of hyper triglyceride mi a in women. Rib to Forehead Circumference ratio (RFCR) was the best predictor in men. However in the age group of 20-50 years the best predictor of hyper triglyceride mi a were rib circumference and WHtR in 51-90 year group in women and RFCR in 20-50 years group and BMI in 51-90 year group men. The best predictor of hyper triglyceride mi a varies according to gender and age [6].

2.6. HMV: Medical Decision Support Framework:

Decision Support System(DSS) help decision makers to collect and interpret information and construct a foundation for decision making . Medical DSS play an important role in medical by guiding doctors with clinical decisions. Data mining in medical area is a process of uncovering hidden patterns and information from huge medical datasets, examine and use them for disease prediction. The proposed HMV overcome the disadvantages of conventional performance by using a group of seven heterogeneous classifiers. HMV framework removes noise from medical dataset by using clustering approach. The selection of optimal set of classifiers is an crucial step. The HMV satisfied 2 conditions accuracy and prediction diversity to achieve high quality. The HMV ensemble framework is done on 2 heart disease dataset, 2 diabetes dataset, 2 liver disease dataset, 1 hepatitis dataset, 1 park in son’s disease dataset. In all these disease datasets HMV has produced highest accuracy in comparison with other classifiers. Multilayer classifier is introduced to enhance the prediction. Predictions done by proposed DSS is parallel with prediction performed by panel of doctors. Heterogeneous classifier ensemble model is used by combining different type of classifiers and achieved a high level of diversity. Naive Bayes (NB) is probabilistic classifier which shows higher prediction accuracy and classification. Eager evaluation methods are decision trees, QDA (Quadratic Discriminant Analysis), LR (Linear Regression), SVM, and Bayesian classification. Lazy evaluation method is KNN combining lazy and eager evaluation algorithm (hybrid approach) results in overcoming the limitation of both eager and lazy methods. Eager method may suffer of missing rule problem when there is no matching exists. In this scenario it adopts default prediction. Proposed framework is constructed on three modules. The first module is based on data acquisition and pre processing which gathers data from different data warehouse and pre process them. In second module individual classifiers training is executed on the training set and they are used for predicting unknown class labels for test instances. The third module is prediction and evaluation of proposed ensemble framework. It is also performed on real time blood CP datasets which was taken from PIMS hospital to discover healthy and disease patients and the result of the samples again showed higher disease prediction accuracy it also guides practitioners and patients for the prediction of disease based on the symptoms of disease [7].

3. COMPARITIVE STUDY:

Predictions	Advantage	Disadvantage
1st prediction	High accuracy	Time consuming
2nd prediction	Accuracy	Imbalanced classification
3rd prediction	Efficiency	Duplicate content
4th prediction	Efficient	uncertain
5th prediction	Simple	No cause effect relationship
6th prediction	High accuracy	Inflexible

Table 1 com paritive study on various predictions

4. CONCLUSION:

Accuracy plays an requisite role in the medical field as it is related to the life of an individual. In the analysis of these six prediction techniques HMV: DSS using multilayer classifier is considered to be the best prediction technique among these because it has higher prediction accuracy and it also help and guide practitioners in predicting the disease.

REFERENCES:

1. LingjianYanga,a,1,ChrysanthiAinalib,c,1,,AristotelisKittasb,FrankO.Nestlec,LazarosG.Papageorgioua,*,SophiaTsokab,*, "Pathway level disease data mining through hyper-box principles, JID:MBS, [m5G;November 10, 2014;15:25]

2. Archana Purwar1 and Sandeep Kumar Singh2 ,"Hybrid Prediction Model with missing value Imputation for medical data", Expert Systems with Applications(2015)

3. Dang Nguyen a, b, Bay Vo a, b, n, BacLe c," CCAR: An efficient method for mining class association rules with item set constraints", Engineering Applications of Artificial Intelligence 37(2015)115–124.

4. Wilco W. M. Fleurena, b, Wyn and Alkema a, c,*, "Application of text mining in the biomedical domain", Methods 74 (2015) 97–106.

5. Nikolas Papanikolaou, Georgios A. Pavlopoulos, The odosios The odosiou, Ioannis Iliopoulos *, "Protein–protein interaction predictions using text mining methods", Methods xxx (2014) xxx–xxx

6. Bum Ju Lee, Jong Yeol Kim*, "Indicators of hyper triglyceride mi a from anthropometric measures based on data mining", Computers in Biology and Medicine 57 (2015) 201–211.

7. [1]Saba Bashir, 2Usman Qamar, 3Farhan Hassan Khan, 4Lubna Naseem 1,2,3 Computer Engineering Department, College of Electrical and Mechanical Engineering National University of Sciences and Technology (NUST), Islamabad, Pakistan 4Shaheed Zulfiqar Ali Bhutto Medical University, PIMS, Islamabad, Pakistan, "HMV: A medical decision support framework using multi-layer classifiers for disease prediction", Journal of Computational Science (2016).