Predictive Research For Mental Health Disease

Shaifali Chauhan , Ankur Garg

Abstract: Many people are suffering from some kind of mental illness and this number is increasing day by day. Despite major revolutions in medical science exact identification of factor that leading to mental illness is still unknown to the world. Due to its ambiguous nature, mental state of person is a major focus on research these days. With the emergence of smart phones, PCs, internet of things. The amount of data human kind produce everyday is huge and only accelerating. These data are stored in a semi structured way and used to get meaningful relationships and trends in data. Data mining techniques can be efficiently used on this data to find hidden patterns between different attributes of data. This paper describes the prototype to use data mining technique namely Random forests classification to determine person's mental state based on attributes such as age, gender, life style, education, Occupation, personal income, vision, sleep, mobility, hypertension, diabetes. The system will predict whether a patient is suffering from mental illness or not.

Index Terms: Data mining, Random Forest, Decision tree, Knowledge Discovery

I INTRODUCTION

Highlight The World Health Organization predicted that depression would be the world's leading cause of disability by 2020[1]. The World Health Organization defines mental health as "The state of well-being in which the individual realizes his or her own abilities to cope up with the normal stresses of life and work productively and fruitfully and be able to make contributions to his or her community"[2].

Our mind record everything that happened to us and react according. Various factors create delusion for mind and start leading to mental illness. Treating mental illness efficient always a been tricky use case due to lack of proper recognition of symptoms and also high cost of health care contributes more towards this as people are reluctant to consult an experts until or unless they are facing problem in day to day activities. The health care industry a huge amount of data which is not being stored in a structured way and if it is being stored in a structured way then no mining is being to done to identify early symptoms or diseases or environmental factors of a person that will eventually lead to mental illness. The Data mining technique can help in health care for a effective treatment and decision making.

A. Data Mining

This age is called age of internet and the data being collected is increasing day by day. We should be using this data to extract meaning information with the help of data mining. As per Jiawen Han in his book Data Mining: Concepts and Techniques[3]. The data mining is "Extraction of interesting, non-trivial, implicit, previously unknown and potentially useful patterns or knowledge from huge amount of data".

The process of knowledge discovery using data mining is known as Knowledge Discovery from Databases(KDD) and has the following steps -

1. Data cleaning - remove noise and inconsistent data
2. Data integration - combine data from different sources
3. Data transformation - In a format appropriate for mining
4. Data mining -generally, in this intelligent methods applied in order to extract data patterns.
5. Pattern evaluation-in this step, data patterns are evaluated.
6. Knowledge representation-generally, in this, knowledge is represented.

- Steps involved in KDD Process is shown in fig1

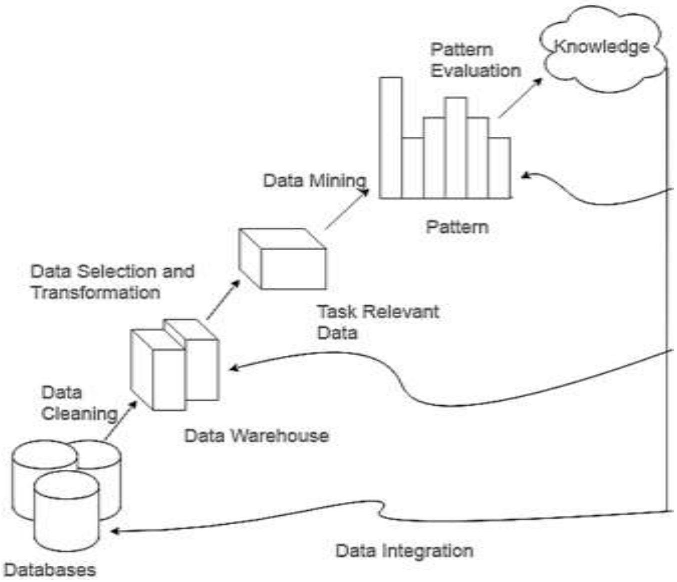


Fig 2.1 Steps involved in the KDD process

B. Machine Learning

As data mining is an integral part of knowledge discovery, data mining used various learning methods to find patterns between different attributes of data. Some learning methods are -

Revised Manuscript Received on July 18, 2019.

Shaifali Chauhan, Computer Science and Engineering, M.I.E.T, Meerut, Uttar Pradesh India.

Ankur Garg, Computer Science and Engineering, M.I.E.T, Meerut, Uttar Pradesh India.

- Learning by classification : Classify based on previous data set
- Learning by Association : Relations and patterns between existing data set
- Clustering : Group of similar thing to find predictive patterns

The Data mining algorithms are classified into two categories : descriptive and predictive. Both these models approaches are important in healthcare to reduce unnecessary experiments and improved accuracy.

C. Mental Disorder

Mental Disorder is a problem which arises when the brain is not working properly, due to this one can not behave as a normal person. There are a number of types of the mental disorder or we can say in the depression, manic depressive illness, dementia, dementia praecox and disorders of anxiety. A person's behavior changes and these changes or symptoms are like mood changes, may be increase temperament etc. The main reason which is observed is stress due to stress there are a number of factors or problems come in one's life.

D. Mental Disorder Symptoms

A number of factors can contribute to risk for mental illness,

Such as-

- Your genes and family history
- Your life experiences, such as stress or a history of abuse, especially if they happen in childhood
- Biological factors such as chemical imbalances in the brain
- A traumatic brain injury
- A mother's exposure to viruses or toxic chemicals while pregnant
- Use of alcohol or recreational drugs
- Having a serious medical condition like cancer
- Having few friends, and feeling lonely or isolated

E. Role of Data Mining in Analysis

As we know that information plays a big role for analysis. So that we must have a large data for analysis. It cannot be done by A human because if human did this analysis then there is some chance of getting error and these errors results in wrong pattern and wrong predictions. That's why we must have some technique which improves our performance of analysis, and the answer is data mining. Actually the data is coming from different sources and when we integrate it then it will be very big in size. To handle it and for making decision on it is a big task and can be done by the data mining .Very first we have to collect data and then we must understand which kind of data is of our need. After selection of data we must remove the missing values and the noisy data from the selected data. Next we have to classify it so there are a number of classification algorithm for that classification is very important part of the data mining process in which we have two types of data set one is training data set and another is test data set, in the training data set the data is known to us and we have to apply classification algorithm on it and making a good classification rules and create a model on it. This model is very useful to classify a test data set in test data set the data is unknown because it is coming from the running environment . so we have to apply our model on it and get the result.

II. RANDOM FORESTS AND DECISION TREES

Decision trees are a type of model used for both classification and regression.[4]. A tree is set of conditions with different answers leading to different leaf of the tree. The Depth of tree is important as it will indicate that how many question were answered before we reached to out prediction. Only thing should be kept in mind while developing a decision tree that the it should not be overfitting. We would like to minimize the error and variance from decision trees. For this purpose we have random forests. Random forest is simply a collection of decision tress whose result will define the final aggregated result. Random tree is a strong modeling technique to avoid error due to overfitting and variance.

III. EXPERIMENTAL RESULTS

We will be analyzing dataset taken for Kaggle repository for a proof of concept. To analyze the data, we will be using two data mining tools - WEKA and RATTLE. We will also calculate various statistics based on observation recorded from both these tools to compare their accuracy.

A.WEKA

Waikato Environment for Knowledge Analysis (WEKA) [5], is a very popular open source software written in Java and developed at the University of Waikato, New Zealand, in 1997. It is available under the GNU General Public License [6]. WEKA is developed by the University of Waikato in New Zealand, it is a system which implement the data mining various algorithm. WEKA can be used for the machine learning technique. The collected data are used and after that we apply different algorithm on the data set, it apply algorithm for preprocessing of data, data classification, data clustering, regression, association rules and as well as it also gives a visualization tool for the users.

B. WEKA Installation

We are using WEKA 3.7.4 for data analysis. When we install it then WEKA will shows as

Waikato Environment for Knowledge Analysis
Version 3.7.4
(c) 1999 - 2011
The University of Waikato
Hamilton, New Zealand

Fig WEKA after installation

C. WEKA File Formats

WEKA is using a default format namely ARFF attribute relation file format. IT is an ASCII text file. It may be that data base in any format and we can convert it into ARFF format. For example our database is maintained in MS excel sheet so first we have to convert it into ARFF format for this we have to convert it into CSV file format and after that it is loaded into the WEKA and easily converted into ARFF file format. We have converted our database into ARFF file format because it will be very helpful in the analysis of data.

D. WEKA Interfaces

There are 7 interfaces in WEKA. Using these interfaces, user can easily mine useful information from data. These following are the interfaces-

- Explorer
- Preprocess
- Classify
- Cluster
- Associate
- Select Attributes
- Visualize

E. RATTLE

Rattle is abbreviated as R Analytical Tool To Learn Easily. Rattle is an analytical tool which is written in the R language and it is a graphical data mining application. Rattle will work after the installation of R. We can analyze the data in a very efficient way and the easy way. By the graphical interface the working in the Rattle is very user friendly and there are many data mining algorithm by which the researcher can easily can input the data in any format and analysis it, to get the desired result.

F. RATTLE Installation

To run the Rattle we have to install first Gnome and then Glade Libraries separately. These packages are simple packages installed in the Linux and OSX .For the Ubuntu we install Glade-3 and the Gnome packages. For the window based system if one wants to install these libraries then one should use this link http://downloads.sourceforge.net/gladewin32.

Full instructions are available from http://rattle.togaware.com [8].

Once you download or install these libraries we have to restart R console to check whether the R has been installed or not and after the restart of R it can find new libraries also. After the installation of R we have to install one more libraries which is RGth2 and same time install R packages.

First install the RGth2

Install . packages (" RGth2")

Then install Rattle Packages

Install . packages ("rattle")

The collected data are used and after that we apply different algorithm on the data set, it apply algorithm for preprocessing of data, data classification, data clustering, regression, association rules and as well as it also gives the visualization tools for the users. We are using rattle 5.2.0 for data analysis.

G. RATTLE File Formats

In Rattle there are a number of file format which we can use for analysis. The Data tab is the starting point for Rattle and where we load our dataset. Rattle is able to load data from various sources. Support is directly included for comma separated data files(.csv files as might be exported by a spreadsheet which use commas to separate variable values in a record), tab separated files (.txt, which are also commonly exported from spreadsheets and use the tab character to separate columns, rather than commas), a common data mining dataset format used by Weka (.arff files), and from an ODBC connection (thus allowing connection to an enormous collection of data sources including MS/Excel, MS/Access, SQL Server, Oracle, IBM DB2, Teradata, MySQL, Postgress, and SQLite) [7]

H. Rattle Interfaces

There are 9 tabs in rattle. We are understating it one by one.

To work with the Rattle the procedure is summarized as:

1. Firstly Dataset must be Load and then variables should be selected;

2. Then for the distribution of the data it must be explorer distributions;

3. After this distribution test should be performed

4. Now for the selection of needed model data should be transformed

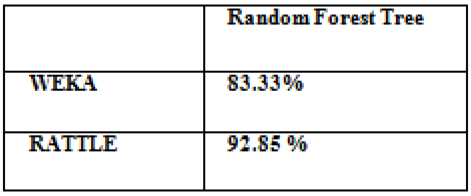
5. Now Models should be build;

6. after this model should be Evaluate and score the datasets;

7. for the Review the work Log tab will help and show the data mining process.

IV. RESULT

We have forest classifier algorithm on our data. Percentage of accuracy is coming from the classifier which is displayed in Table 4.1 which is coming applied our random from the output of our classifiers. This table is constructed on the basis of output by the WEKA tool and Rattle tool on applying random Forest Tree classification algorithm.



	Random Forest Tree
WEKA	83.33%
RATTLE	92.85 %

Table: Accuracy by WEKA and RATTLE Tool

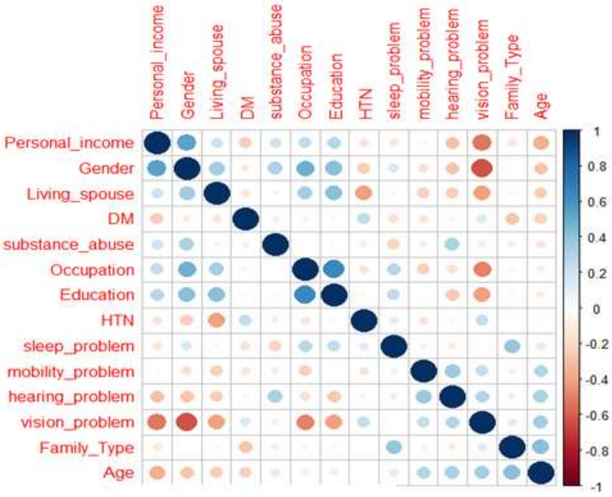


Fig: Correlation Geriatrics.arff using Pearson

This correlation matrix between attributes is important to design trees. Random rules on attributes will result in poor accuracy as compared to the rules that are inferred from correlation matrix.

V. CONCLUSION

In this paper, we have talked about how data mining can help in an efficient and effective health care management .We have also talked about an effective approach in the form of decision trees to predict a person mental state by taking reference from previous instances. Data mining has a potential to improve clinical decisions. Our motive is to increase the number of aware people by which needy person can get advantage of this. And needy person can enjoy this beautiful world. The work can further be enhanced after getting real life dataset and perform a data modeling using tools like WEKA and RATTLE to determine the prediction accuracy of various classification algorithms. In Dataminng, We were able to identify and predict whether a person is normal or abnormal. We found desired results in both tools but in comparision of both tools RATTLE tool has given better results then WEKA in terms of accuracy.

VI. FUTURE WORK

Day by day the role of mobile computing is growing due to the data and we know that the data is coming from various resources and this data is also huge in nature. For this purpose we can collect the data by develop an application which is based on web. And by this we can store the information and also spread the information who need such type of information. We can analyze this data which is based on the mental disorder. And for the analysis point of view we can apply the data mining algorithm on it. Such type of application easily run on the laptop, mobile phone and other information related devices. And one can access and provide the data to the desired centers.

There are a number of classification algorithms, and we can apply these classification algorithms and can check which the best classifier give the best results. After that we can also apply clustering algorithm for the better result. After applying these algorithms a person can insert data by their mobile device and can get the result back.

VII ACKNOWLEDGMENT

Want to express my gratitude to my honored guide Dr Ankur Garg, Professor, Department of Comp Science and Egg., MIET, Meerut for their strict supervision and guidance without which this task would not have been accompolished.

I would like to thank my helpers and my colleagues, who have facilitated me during this research work. I am enormously grateful to my parents and my coworker for being columns of strength, for their unfailing moral support, and inspiration. I acknowledge and grateful one and all who have been instrumental in helping me to complete this research work.

REFERENCES

1. Lopez AD, Murray CCJL{1998): The Global Burden of Disease,1990-2020, Nature Medicine vol 4,pp 1241-1243
2. The World Health Organization(2011a). political Declaration of the High-level Meeting of the General Assembly on the Prevention and Control of Non-Communicable Diseases. 66th Session of the United Nations General Assembly, New York:WHO
3. Jiawei Han and Micheline Kamber. Data Mining : Concepts and Techniques. Morgan Kaufmann Publishers, 2nd edition,2006
4. Niel Liberman. Decision Tress and Random Forests. Towards data science.com
5. Deswal BS, Pawar A. An Epidemiological Study of Mental Disorders at Pune, Maharashtra. Indian J Community Med. 2012;37(2):116-21.
6. G. Holmes; A. Donkin and I.H. Witten (1994). "Weka: A machine learning workbench". Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia.URL:http://www.cs.waikato.ac.nz/~ml/publications/1994/Holmes-ANZIIS-WEKA.pdf. Retrieved 2007-06-25.
7. http://datamining.togaware.com/survivor/Loading_Data.html.
8. Graham J Williams, Rattle: A Data Mining GUI for R, The R Journal Vol. 1/2, December 2009, ISSN 2073-4859