OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

STUDY OF VARIOUS DATA MINING TECHNIQUES IN PREDICTION OF DEPRESSION

Ms. Suchita Sinhal, Research Scholar, CS & IT, The IIS University, Jaipur
Dr. Ruchi Nanda, Sr. Assistant Professor, CS & IT, The IIS University, Jaipur

Abstract

Data mining and Data analytics are promising research fields for their attempts to predict and analyze data from different perceptions and summarize it into significant information in order to identify hidden patterns from a huge dataset. Data mining identifies patterns which were earlier undetected by using statistical approach whereas data analytics focuses on comparing the patterns discovered with other patterns to solve business problems. These are equally significant and connected fields that cannot exist without each other. Healthcare organizations are exploiting these technologies for improvising their clinical and business processes. The healthcare data related to mental health is quite complex and uneven. The state of depression and mental disorders are neglected due to unpredictable symptoms and treatments based on assumptions. Depression is one of the most common problems that affect a large population today and is noticeable in any age group.

This paper provides the description of various data mining techniques such as Artificial Neural Networks, Decision Trees, Fuzzy classifiers and Bayesian classifiers applied to massive volume of depression data. This paper comprehensively presents a literature review on these techniques utilized by the researchers in the prediction of depression. It helps in investigating the factors responsible for depression. These factors are perceived and mathematically modeled. The paper provides an insight into some of the commonly implemented data mining techniques used to detect depression as well as classify the depression in different states and make a comparative study of such techniques based on their performance and accuracy. An estimate of the data size required to predict and detect depression is presented that can help in future research work. It explains why some techniques are more commonly used as compared to others by highlighting the rationale. This paper also helps in finding out the gap in the existing prediction techniques which helps researchers to further improvise the techniques by using some advanced methods and functions that can be applied for the prediction of depression.

Keywords- Data-Mining, Health-Care, Mental-Health, Depression, ANN, Bayesian Classifiers, Decision Trees, Fuzzy Classifiers

Introduction

Today there is an abundance of data which has created the condition of data rich and information poor leading to widening of the gap between data and information. Data Mining helps in the extraction of knowledge from large amounts of data. Data Mining tools helps in finding some interesting patterns. The patterns so discovered can be compared to other patterns in order to generate an insight, by data analytics. Data analytics and data mining provides us with bringing the data from raw state to result, with the main difference being that data mining takes a statistical approach to identify patterns while data analytics is more broadly focused on generating intelligence geared towards solving business problems. As the healthcare industry moves deeper into value-based care, organizations are utilizing

these strategies to improve transparency into their business and clinical processes. Data analytics and data mining are equally critical for business intelligence, and neither can exist without the other. Globally the healthcare issues specifically related to mental health issues and psychiatric expert system building are the least to be bothered about and very limited research work is done here The state of depression and mental disorders are neglected due to unpredictable symptoms and treatments based on assumptions. Moreover people are also unaware and have a very little interest about the mental illness. The amalgamation of more than one illness makes the situation even more complex. Depression is one of the most common problems that affect a large population today and is noticeable in any age group. Depression often precedes and may cause directly or indirectly many chronic conditions such as high blood pressure, diabetes, insomnia. Serious depression in a person can affect the family life and destroy the life of the affected person if not treated on time. Depression is associated with loss of appetite, headaches, body pain, mood disorders, irritability, frustration, unnecessary anxiety, loss of self esteem, fear of death, fatigue, inability to experience pleasure, feeling of guilt etc. The focus of this paper is on the study of data mining techniques such as ANN, Bayesian Classifiers, Fuzzy logic, to detect and diagnose Depression. The patterns that emerge from data mining will not only improve understanding of this disease, but can give us new insights into prevention and treatment. This approach balances the fact that no two cases of mental illness are the same as all patients as individuals are different.

Objectives

- The objective of the research is to enhance the existing data mining concepts to detect and predict depression in all age groups based on certain factors.
- To develop an intelligent data mining algorithm that involves less computational effort and faster learning capability of depression data.

Techniques Implemented in Data Mining

ANN in Depression

Artificial Neural Network (ANN) for depression classification was considered by Bhuvana, 2014. A neural network is constructed by highly interconnected processing units/nodes which perform simple mathematical operations. Neural networks are characterized by their topologies, weight vectors and activation functions used in the hidden layer and output layer, the topology refers to the number of hidden layers and connections between nodes in the hidden layers. The network models can be static or dynamic. Static networks include single layer perceptrons and multilayer perceptrons. A perceptron or adaptive linear element refers to a computing unit. This forms the basic building block for neural networks. The input to a perceptron is the summation of input pattern vectors by weight vectors.

OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

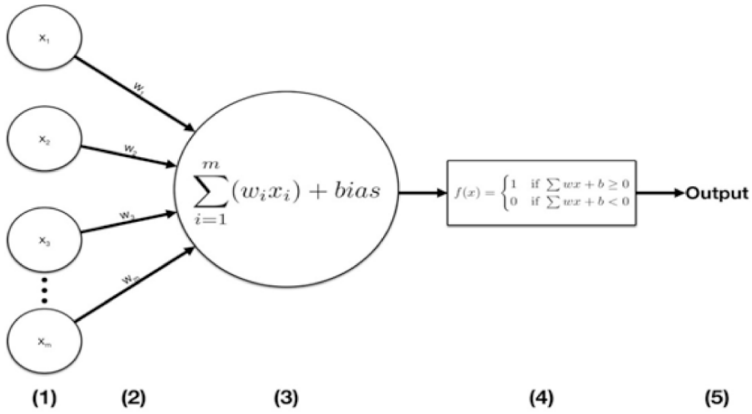

Figure 1 : Single Layer Perception Model (Bhuvana (2014)

In a multilayer perceptron model information flows in a feed forward manner from input layer to the output layer through hidden layers. The number of nodes in the input layer and the output layer is fixed. It depends on the number of input variables and output variables in a a pattern. The number of nodes in the hidden layer and the number of variables in the hidden layer are variable and this number is found by trial and error. The activation function which is used to train the ANN is the sigmoid function which is given by

$$F(x) = 1/\ 1 + \exp(-x)$$

Where F(x) is a non linear differentiable function,

A multilayer feed forward network that uses an exponential activation function, is called Radial Basis Function (RBF). The RBF uses the concept of distance between patterns and the various centers of the patterns. The number of nodes in the hidden layer of the RBF network is equivalent to number of centers, used to find the distance. A bias value of 1 is appended to the hidden layer nodes for convenience of weight processing. The final weights are obtained between hidden layer and the output layer. These weights are used for classification of depression.

Back Propagation Algorithm (BPA) is used for training ANN topology. BPA uses the concept of forward propagation to find the error of ANN for a pattern. BPA uses reverse propagation for updating weights. When the weights are optimal, so that there will be a close mapping among features of depression data with outputs in the output layer of ANN. The BPA uses the steepest-descent method to reach a global minimum. The ANN using BPA has one input layer, one output layer. The number of hidden layers can be one or more than one. The number of layers and number of nodes in the hidden layers are decided by pruning. The connections between nodes in the adjacent layers are initialized with random weights. The training process consists of phase-1 (forward) and (phase 2) reverse propagation. A pattern from the training set is presented in the input layer of the network and the error at the output layer is calculated. The error is propagated backwards towards the input layer and the weights are updated. This procedure is repeated for all the training patterns.

Strong Support for ANN in Depression Classification (Bhuvana (2014); Sau and Bhakta (2017)) The purpose of using ANN for depression classification is due to the following reasons:

1. The working concepts of ANN are based on statistics, like using linear summation between layers, to propagate information from input to output layers.

2. ANN uses transformation function like sigmoid function for back propagation algorithm, exponential function in Radial basis function network, Hyperbolic tangent function in Echo state neural network to squash output values from neurons.
3. ANN uses objective function for finding optimal weights between layers for mapping inputs (information of depression) to outputs (category of depression).
4. ANN can classify depression even if there is slight change in collected data.
5. The ANN can be trained with minimum patterns.

Because, the working properties of ANN are based on statistical concepts, ANN assures correct classification of depression.

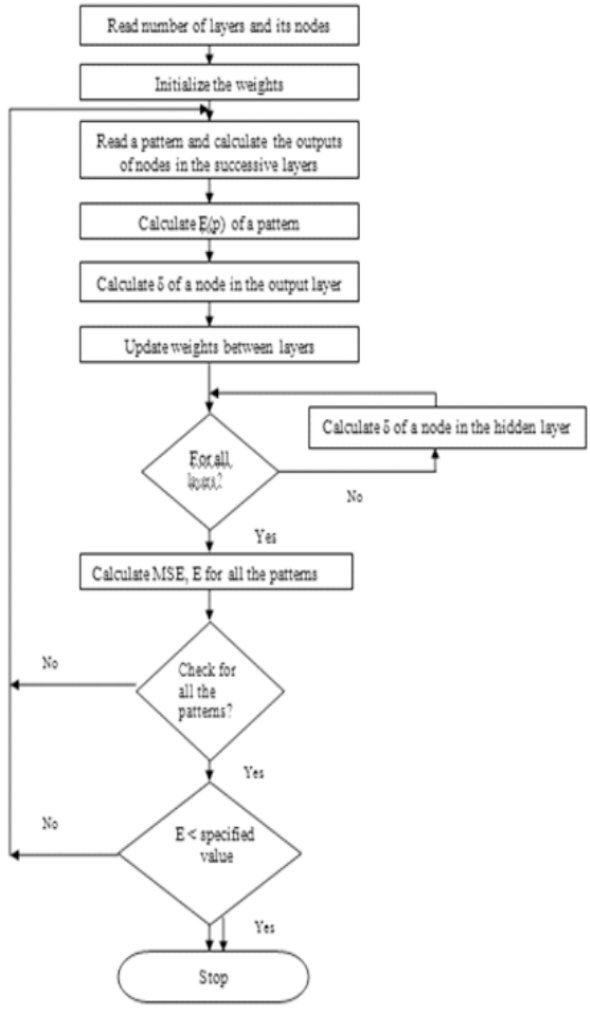

Figure 2: Flowchart for BPA (Bhuvana.R,2014)

Fuzzy Logic in Depression

Medical data related to mental health is characterized by uncertainty.(Gursel, 2016) Fuzzy Logic is a term that is associated with uncertainty, vagueness, fuzziness, having common feature as not being

clear enough to be processed by hard computing techniques. Fuzzy logic deals with approximate values instead of certain values. Any Fuzzy logic system has basically 4 components- a fuzzifier, an inference engine, knowledge base and defuzzifier. Fuzzifier is responsible for fuzzification process that involves converting a fixed object into a fuzzy set. Inference Engine- after fuzzification the resultant fuzzy sets are processed in the inference engine according to the rules of the rule base. The inference engine is the processing unit of the system. Knowledge Base- this is the most important part of FL system. The performance of the FL system depends on the knowledge base. Defuzzifier- normally the output of the inference engine is also a fuzzy set which is not useful in the real world. It needs to be transformed into the useful and understandable value which is done by defuzzifier.

A review of literature is presented that depicts different data mining techniques used for the prediction and diagnosis of depression and its different states in different age groups of people.

As per Sau and Bhakta (2017), Depression is one of the most important causes of mortality and morbidity among the Geriatric population. As the brain experiences aging it becomes more vulnerable to depression. They studied various socio-demographic factors and morbidity attributes to predict depression among geriatric population using Artificial Neural Network. Multilayer Perceptron (MLP), a feed forward ANN model was built in WEKA for classification with input variables i.e., age, gender (Male/Female), living spouse (Yes/No), type of family (Nuclear/Joint), Literacy (Literate/Illiterate), occupational status (presently working/ presently not working), personal income (Yes/No), and substance abuse (Yes/No), hearing problem (Yes/No), visual problem (Yes/No), mobility problem (Yes/No) and sleeping problem (Yes/No). The ANN model was made with seven interconnected neurons in one hidden layer. Primary data was collected by interviewing 105 elderly people and this data was used as a training data set for data mining in Weka. Then, the predictive model was build using ANN as a classifier in the Weka. Output was classified as - depressed and non-depressed. ANN model was trained and tested on the primary data set with 10-fold cross validation method.

Depression is a chronic mood disorder. Chattopadhya(2017) considered the severity of real world depression cases and implemented a fuzzy neural hybrid model for the diagnosis of the same. Principal Component Analysis, a method used to extract hidden features from multidimensional data was used here. It helped to extract the significant features by reducing the number of symptoms. 14 symptoms of depression have been considered in this study which denotes independent factors some of which are- feeling sad, weight loss, insomnia, loss of appetite, lack of thinking, loss of pleasure etc. A data of total 302 adult 'psychotic depression' cases have been collected in this study. Data has been finally normalized using max-min normalization technique. An input vector matrix was created with significant symptoms and then a fuzzy neural hybrid model was developed and tuned with back-propagation neural network algorithm.

Bhuvana et al.(2015) studied the use and implementation of two ANN algorithms : back propagation algorithm and radial basis function for categorizing and identifying the type of depression a patient has based on 21 various inputs to the algorithm. Hamilton depression rating scale is a multiple item questionnaires which was used here to collect information from the affected patients. The scale provides an indication of depression and as a guide to evaluate recovery. Depression data of 1800 patients based on Hamilton rating scale were collected through oral discussion. The information not being precise was presented into the input layer of ANN algorithm and processed with stored

weights. The outputs received from the output layer of the ANN were used to retrieve the information from the database as well as information provided by patients provides a complete information based on which the doctors could suggest remedies. Radial Basis Function Neural Network consisted here of an input layer, an output layer and a single hidden layer. A distance measure is used to associate the inputs to outputs. RBF is capable of performing approximations with an additional advantage of not being involved in repeated training.

Mukhurjee et al. (2014) stated depression as a serious disease that affects a large fraction of the global population. The diagnosis of depression faces a problem due to uncertainty of widely varying symptoms. An attempt was made to model a three layered fully connected neural network taking into account ten common symptoms namely - feeling sad, loss of pleasure, weight loss, insomnia, hypersomnia, loss of appetite, psychomotor agitation were considered for this work. The grade of each symptom and the corresponding probability of depression were assigned [0,1].

The model consists of 3 layers- the input layer, the hidden layer and the output layer. There are 10 neurons in the input layer corresponding to the 10 symptom units. The hidden layer had variable number of neurons ranging between 3 and 15 The output layer has only neuron that gives the final output as degree of depression. The weight sets [v] and [w] connect the outputs of input layer to outputs of hidden layer and the outputs of hidden layer to the inputs of output layer. In this case weight sets [v] and [w] are finely tuned to optimize the system with the help of a back-propagation algorithm. A radial basis function neural network was built in the second approach that used linear transfer functions and a radial basis function as its transfer function.

Bhakta and Sau(2016) studied depression among elderly population which is one of the emerging problem of public health. They used Weka for prediction based on machine learning classifiers. In this paper 5 machine learning classifiers are compared with respect to three test options. The machine learning classifiers that are used are- Bayes Net, Logistic, Multilayer Perceptron, SMO and Decision Table. All these classifiers are used on three different test options-using training and testing set, Cross- Validation and percentage split.

WEKA was used as machine learning platform into which the training data set was loaded, then filters are applied on this data set for optimized result. After pre-processing, result is fed into the classifiers and corresponding test options are selected. Prediction output format is chosen to view the result in human readable format. After all these setup prediction is started and result is collected in a text file. From the values of the metrics it was clear that SMO was more accurate and précised model for prediction of depression among senior citizens.

Chattopadhya et al.(2008) made an attempt to develop fuzzy logic based expert system which may be able to reason like doctors for screening adult psychosis. Among several techniques of fuzzy classifier (FC)-design , clustering to classifier technique(CCT) has been adopted. The key objective was to develop a classifier system that is able to determine the chance of occurrence of adult psychoses.

Conclusion

Depression has affected 2-3% of the global population and increasing day by day due to stress which is quite alarming. There has been very limited research work in this field due to the uncertainty related to the data. As a result it becomes very necessary for a quick and immediate automated detection

and diagnosis for a patient. This can be achieved only with the help of an intelligent data mining concept for the vast amount of data related to depression cases. Such Data mining intelligent algorithms can better classify the collected data and generate interesting patterns which can help in the better understanding of the disease, moreover the patients will be able to receive higher personalized treatments. ANN, Back-propagation and Fuzzy logic have been used so far for the detection and diagnosis of Depression. The alarming increase in the rates of depression cases however calls for the development of a full-fledged expert system tool to achieve the required amount of satisfaction and standardization. A hybrid model can be developed as an expert tool.

A number of different data mining techniques and their implementation have been studied so far in the field of medical health with special case to depression. Further application of technique will depend on the availability of data and the ease of detecting such depression cases.

The comprehensive study of application of various data Mining techniques implies that ANN and fuzzy logic are some of the commonly implemented techniques used in the prediction of depression and produce accurate results as compared to other methods.

ANN with further enhancements can be applied to deal with the unevenness present in data associated with mental health due to the adaptive nature of ANN.

References

- Bhakta, I., & Sau, A. (2016). Prediction of depression among senior citizens using machine learning classifiers. International Journal of Computer Applications, 144(7), 11-16.
- Bhuvana, R., Purushothaman, S., Rajeswari, R., & Balaji, R. G. (2015). Development of combined back propagation algorithm and radial basis function for diagnosing depression patients. International Journal of Engineering & Technology, 4(1), 244.
- Bhuvana, R. (2014). Development of artificial neural networks for data mining to diagnose depression.
- Chattopadhyay, S. (2017). A neuro-fuzzy approach for the diagnosis of depression. Applied Computing and Informatics, 13(1), 10-18.
- Chattopadhyay, S., Pratihar, D. K., & De Sarkar, S. C. (2008). Developing fuzzy classifiers to predict the chance of occurrence of adult psychoses. Knowledge-Based Systems, 21(6), 479-497.
- Concepts, D. M. (2006), Technique, Jiawei Han and Micheline Kamber. University of Illinois at Urbana-Champaign.
- Daimi, K., & Banitaan, S. (2014). Using data mining to predict possible future depression cases. International Journal of Public Health Science (IJPHS), 3(4), 231-240.
- El-Nasr, M. S., Yen, J., & Ioerger, T. R. (2000). Flame—fuzzy logic adaptive model of emotions. Autonomous Agents and Multi-agent systems, 3(3), 219-257.
- Gürsel, G. (2016). Healthcare, uncertainty, and fuzzy logic. Digital Medicine, 2(3), 101.
- Ilgen, M. A., Downing, K., Zivin, K., Hoggatt, K. J., Kim, H. M., Ganoczy, D., ... & Valenstein, M. (2009). Identifying subgroups of patients with depression who are at high risk for suicide. The Journal of clinical psychiatry, 70(11), 1495.

- Koh, H. C., & Tan, G. (2011). Data mining applications in healthcare. Journal of healthcare information management, 19(2), 65.
- Mukherjee, S., Ashish, K., BaranHui, N., & Chattopadhyay, S. (2014). Modeling depression data: feed forward neural network vs. radial basis function neural network. American Journal of Biomedical Sciences, American J. Biomed. Sci, 6(3), 166-174.
- Sau, A., & Bhakta, I. (2017). Artificial Neural Network (ANN) Model to Predict Depression among Geriatric Population at a Slum in Kolkata, India. Journal of clinical and diagnostic research: JCDR, 11(5), VC01.
- Shin, H., Park, H., Lee, J., & Jhee, W. C. (2012). A scoring model to detect abusive billing patterns in health insurance claims. Expert Systems with Applications, 39(8), 7441-7450.
- Tsai, H. H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. Expert systems with applications, 39(9), 8172-8181.
- Yu, S. C., & Lin, Y. H. (2008). Applications of fuzzy theory on health care: an example of depression disorder classification based on FCM. WSEAS Transactions on Information Science and Applications, 5(1), 31-36.