ABSTRACT

Data mining is a concept of finding very close relationship among various features when a query is generated. The amount of exactness of the results obtained for a query is based on the type of algorithm used. Data mining is used to store and manage the multidimensional database system like depression data with 21 features.

In this research work, depression data of different patients have been collected from the hospital. To increase the data additional data have been simulated using the mat lab software.

In this research work, huge amount of data have been collected and generated and treated as raw data. Relevant features are extracted. The extracted features are stored in an acceptable format, so that it can be used for the proposed algorithms in this work. Training and testing of the proposed algorithms give an output in the form of classification to which the patient belongs. If the research work has to be converted into commercial aspects then front end screen and back-end data storage software should be known based on the usage of the customer.

The psychological depression data analysis is considered in this research work. Depression is one of the most common psychological conditions, affecting nearly everyone either personally or due to an affected family member. Depression can interfere with normal functioning and frequently cause problems with work, social and family adjustment. Serious depression can destroy the family life and the life of the depressed person.

A serious loss, illness, relationship problems, work stress, family crisis (like bereavement), financial setback, or any unwelcome life change can trigger a depressive episode. Very often, combinations of biological, psychological and environmental factors are also involved in the development of depressive disorders as well as other psychological problems. Some of the depression features considered are depressed mood, feelings of guilt, suicide, insomnia early, insomnia middle, insomnia late, work
and activities, retardation, psychomotor, agitation, anxiety, anxiety somatic, somatic symptoms, somatic symptoms general, genital symptoms, hypochondriasis, loss of weight, insight, diurnal variation, depersonalization, derealization, paranoid symptoms and obsessional and compulsive symptoms.

In order to evaluate the efficiency of the proposed algorithms in this research work, psychological depression data have been collected from general patients as well as from patients who have undergone different types of counseling from a hospital. The range of values a patient can have is considered based on the Hamiltonian rating scale and expert psychologists decision. Many other evaluating scales can also be considered. Additional data have been generated considering the range of values and also considering the behavioral patterns of data that have been collected from the patients. This ensures that the data generated conform to the actual depression data.

The raw data have been categorized into four classes. Further the data have been separated into training and testing. This is achieved by normalizing the entire set of data. Subsequently, mean for each feature is found. Then summation of squared differences for each pattern is obtained. Redundant patterns are removed. The entire data set is reshuffled in the ascending order based on the summed value for each pattern. Sample patterns that belong to lower range and upper range of summed values are treated as training patterns, and the remaining patterns are considered as testing patterns. There are 21 features used as inputs, and one feature is used as target output for the proposed algorithms. In this work, ANN approach has been proposed for depression diagnosis.

Artificial Neural Networks (ANN) is an algorithm whose outputs are modeled independent of inputs. It is useful for input-output mapping, pattern classification, pattern association, and pattern recovery. Based on the type of data and its analysis, the topology of the ANN and the weight updating algorithms can be decided. The network is made up of input layer, output layer, and one or more hidden layers. The layers are connected with weight vectors. Activation functions are used in the hidden layer(s) and output layer for squashing the outputs of neurons. The topology refers to the number of layers and connections among nodes among layers. The activation functions that can be used are sine, cosine, tanh, log, signum, exponential and sigmoid. 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.

A supervised Back Propagation Algorithm (BPA), Radial Basis Function (RBF), Optimal Discriminant Plane (ODP), Functional Update Back Propagation Algorithm (FUBPA), polynomial processing of inputs, split network, and dynamic analysis for pruning of nodes in the hidden layers have been proposed as different algorithms for classification of psychological depression data mining.

The ANN has been trained with BPA. The topology of the ANN is 21 nodes in the input layer, 10 nodes in the hidden layer and one node in the output layer. The number of cycles required for mean squared error (MSE) to reach the desired minimum value is very large. The MSE does not reach the desired minimum due to some local minima whose domains of attraction are as large as global minimum. The algorithm converges to one of those local minima and hence learning stops prematurely or the value diverges.

The concept of functional update method for BPA indicates that there will be weight updating of the connection between layers, if one node in the output layer is misclassified. The meaning of misclassification is valid when the difference between the target value of the pattern and the output of network in the output layer for the same pattern is greater than a specified value for the purpose of training the inputs.

The network can be trained with any number of nodes in the hidden layer. The correct number of nodes in the hidden layer can be found by dynamically increasing or reducing the number of nodes in the hidden layer. The condition specified is that for the convergence of ANN to be faster, the respective number of nodes are given in the hidden layer (by trial and error simulation method), so that it will converge to the desired MSE or generalize with optimum number of iterations. The optimum represents neither too low nor too high iterations for the network. This in turn will converge to the desired
MSE. By using dynamic analysis, the optimum number of nodes in the hidden layer can be found with minimum computational effort.

The configuration of the network is fixed based on the dynamic analysis and pruning. To reduce the computational effort, the single configuration network is split into two networks. The number of nodes in the hidden layer of each network can be the same or different. Each network is trained separately with initial random weights till the network reaches a different MSE. The split networks, network 1 and network 2, are combined to form the original configuration. The final weights used are obtained from the two networks. This combined network is further trained till the desired classification performance of the network is obtained.

When the patterns are orthogonal to each other then there will be perfect learning of data. In this research work, two important pre-processing methods for the data have been considered such as Polynomial pre-processing and optimal discriminant plane methods.

Polynomial pre-processing involves changing the input dimensions of the pattern into different numerical values. The input vector is outer produced to form a matrix. The input vector is outer produced to obtain a matrix. That is given a column vector ‘v’, an outer produced matrix is obtained by ‘v×v^T’. The different portions of the numerical values are considered as non-linear (NL). The pattern data generated are NL1, NL2, NL3, NL4, NL5, and NL6. Pre-processed patterns are presented to the input layer of the network. Due to this operation the number of nodes in the input layer will either increase or remain the same. The increase in the number of nodes in the input layer will create a smooth surface of the input domain. The pre-processed input vector forms a polynomial discriminant function.

Optimal discriminant plane method involves converting higher dimensional vector into a lower dimensional vector. The 21 features (21 dimensions) are converted in to 2 dimensions. The 2-dimensional vector is presented in the input layer of the network. Due to this, the size of the network is reduced from 21 nodes to 2 nodes in the input
layer. In order to convert 21 inputs into 2 inputs, the 21-dimensional vectors are mapped onto a 2-dimensional space by using a transformation. The transformed 2-dimensional vector does not represent any individual feature instead it is a combination of 21 features with no dimensional quantity.

As an alternative to the BPA, Radial Basis Function is considered. The concept of distance measure is used to associate the input and output pattern values. RBF are capable of producing approximations to an unknown function ‘f’ from a set of input-output data set.

Comparisons of the classification performances and computational efforts of the different weight updating algorithms with different training methods are presented. Several algorithms developed in this work may find other application areas, and it can be applied to other medical diagnosis systems.

This research work has carried out the analysis and modification of existing algorithms and proposed their implementations for depression diagnosis. The existing algorithms have been suitably modified either in the form of pre-processing the inputs, deciding the optimal topology of the network, or training the network in binary. There is not much work that has been done on depression diagnosis using the proposed algorithms. Hence this research work is a solid contribution in the field of psychological depression diagnosis using data mining. The users can be benefited from this research work in the following aspects.

The relevance of mined solution for the given depression data is almost exact. The server can train the new input depression data rapidly and hence there will not be any slackness in the server in meeting the queries of the doctors. This research work is relevant in the present day scenario where even small children are stressed out and many adults at some point in time in their lives go through depression.