International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 4 Issue: 1
ISSN: 2454-4248
241 – 246

A Data Mining Analysis Over Psychiatric Database for Mental health Classification

Shivangi Jain
Department of Computer Science and Engineering
Bhabha Engineering Research Institute
Bhopal, India
sphivangijainap@gmail.com

Dr. Mohit Gangwar
Department of Computer Science and Engineering
Bhabha Engineering Research Institute
Bhopal, India
mohitgangwar@gmail.com

Abstract:Data mining approach help in various extraction unit from large dataset. Mental health and brain statistics is an important body part which is directly connected with the human body. There are many symptoms which can observe from the mental health care dataset and especially with psychiatric dataset. There are many health disease associated with such symptoms i.e. Anxiety, Mood disorder, Depression etc. Diseases such as mental retardation, Alzheimer, dementia and many other related with such symptoms. A proper classification and finding its efficiency is needed while dealing with different set of data. A classification of these disease and analysis requirement make it working for user understanding over disease. In this paper different classification algorithm is presented and classification is performed using J48 (C4.5), Random forest (RF) and Random Tree (RT) approach. The classification with precision, recall, ROC curve and F-measure is taken in as computation parameter. An analysis shows that the Random tree based approach find efficient result while comparing with J48 and Random forest algorithm.

Keywords*:Data mining, Psychiatric disorder, J48 (c4.5), Random forest, Confusion matrix, Mental healthcare.*

*****

I. INTRODUCTION

Today in Dataset there exist data objects that do not comply with the general behavior or model of the data. Such data objects, which are heavy different from or inconsistent with the remaining set of data, are called outliers. An outlier is a data set which is different from the remaining data. Outlier is also referred to as deformity, deviants or anomalies in the data mining and statistics literature. In most applications the data is created by one or more generating processes, which could either reflect activity in the system or observations collected about entities. When the developing process behaves in a casual way, it results in the creation of outliers. Therefore, an outlier often contains useful information about anomaly characteristics of the systems and entities, which impact the data generation process. The recognition of such unusual characteristics provides useful application specific insights [1].

II. PSYCHIATRIC DISORDERS

A mental illness is a disease of the brain that causes mild to severe disturbances in thought and/or behavior, resulting in an inability to cope with life's ordinary demands and routines. Some of the more common disorders are: clinical depression, bipolar disorder, dementia, schizophrenia and anxiety disorders. Symptoms may include changes in mood, personality, personal habits and/or social withdrawal. Mental health problems may be related to excessive stress due to a particular situation or series of events. As with cancer, diabetes and heart disease, mental illnesses are often physical as well as emotional and psychological. Mental illnesses may be caused by a reaction to environmental stresses, genetic factors, biochemical imbalances, or a combination of these. With proper care and treatment many individuals learn to cope or recover from a mental illness or emotional disorder [2]. Persons with mental illness usually exhibit a *cluster* of symptoms – not just one or two symptoms – that are *persistent* and interfere with *daily life and work*. This listing of warning signs and symptoms of mental illness is to be used as an educational and information tool – not as a diagnostic instrument [3].

III. LITERATURE REVIEW

In this section different literature survey approach which works towards mining disease parameters towards mental illness and Psychometric diseases using Data mining approaches. This survey define how the different algorithm given and perform for the data detection for diagnosis.

In this paper [4] Author described the work perform by different model to find Interpretation of Neuropsychiatric Diseases. They have performed work with integration of Rule based reasoning (RBR), Case based reasoning (CBR) and Artificial neural network (ANN) which works to find solution on Neuropsychiatric disease. They have worked and showed five neuropsychiatric diseases with 38 symptoms grouped into six categories. Different model for the disease symptom detection is given by them. Feature selection process is performed with ANN along with RBR

 ISSN: 2454-4248
241 – 246

and CBR approach. The paper described that EEG signal and other resources get generate data periodically.

They have also stated that the combination of EEG parameters, neuroimaging parameters, physiological, psychological and cognitive parameters are required for the detection of neuropsychiatric diseases. Thus it can be determined the disease can be determine using EEG parameters. Paper gives the description of previously performed diseases with psychological analysis. Their model applied to diagnose and analyze them. The work computed by them on Weka tool with real time EEG samples gathered from medical research lab and their hybrid approach outperform best with efficient parameter computation. They have stated that CBR approach predicts the order of confidence of disease.

In this paper [5] author worked with genetic algorithm for diagnosis of Alzheimer's disease (AD). They have worked with genetic algorithm steps to determine early stage, middle stage and complex stage of the disease. This disease is related with mental illness which posses different disorder and multiple parameters to compute it. The algorithm works in a process that it first collects all the sample population from the user input dataset. Further it generates the fitness function and finding genome evaluation fitness of each gene. A new population reproduction based on the cross over and mutation technique with the input population. The "best" solution is returned when fitness function reaches target value. The fitness function having function is to evaluate the quality of each proposed solution. The implementation was performed with C language, also the regression function using R language. The proposed algorithm by the author outperforms best while comparing existing ROC approach. The proposed algorithm also compared with step wise algorithm. Finally author concluded GA for prediction of AD progression by combining the results of a large set of neuropsychological measures from the AIBL study that they had been selected for their sensitivity to cognitive impairment in both MCI and AD.

In this paper [6] machine learning algorithm and its usage in finding mental illness is discussed. They have shown the eight different machine learning approach and perform with 5 mental illness problem , 60 sample cases with 25 attribute detection which participate in mental illness is detected over the research. They have used Multilayer Perception, Multiclass Classifier and LAD Tree classified for better efficient results. Main considerable disorder Attention problem, Academic Problem, Anxiety Problem, Attention Deficit Hyperactivity.

The following are the monitored points which identified as problem and further analyzed and performed further with enhancements.

1. Previous technique such as ANN, CBR algorithm for the processing model generation but still the obvious problem occur with the technique is in generating better result and sequence derivation and finding better result in processing time is lagging in the traditional Heuristic algorithm. This technique persist better result than existing but still enhancement is required which is provided by the proposed procedure [7].
2. Previous technique basic classification doesn't perform a better data classification due to lacking of number of rules thus a better probability model can't get generated using the technique.
3. In previous technique distribution is used because of that the data of the topics varies which determine the drawback of different entities than proposed work which include rule search and distribution algorithm.
4. In the existing distribution independent proportion among component is found thus there is no relation with the other topics is found, where as in new technique normal distribution is used, which provide relation among the topics and provides a flexible framework for the process.
5. Previous Technique make use of original data usage processing with some filtering process to remove the unwanted data. But still the processing of complete data is required at running end.
6. Algorithm process data is need to be investigate and format processing pre-filtering is needed which can save the time and utilization of memory can be saved over the large data process [8].

IV. PSYCHIATRIC DATASET

In order to perform classification on the mental healthcare dataset. A Sample dataset is collected from hospital which is classified from different tree based classification algorithm. A dataset description, symptoms and their description is presented here to understand disease and symptoms information.

V. PROPOSED METHODOLOGY

In order to perform an analysis, there are three different classification approaches were taken and compare. A dataset which is mentioned in previous section is considered and classification is performed. There are following algorithm which is used for analysis.

1. J48(C4.5) Algorithm:

The J48 is the algorithm which is also known as C4.5 is the tree based algorithm for the data classification. This is also referred as statically classifier help in huge data classification with high computation parameter. The modified J48 decision tree algorithm examines the

 ISSN: 2454-4248
241 – 246

normalized information gain that results from choosing an attribute for splitting the data [9].

This approach can also be defined with the following step:

The following pseudo code is used to build decision trees

1. Check for base cases [Initial Device Form Current Active Directory List]
2. For each attribute a {from the captured packets- device address MAC}

Find the normalized information gain from splitting on a select the 16 bit device from the least most significant bit.

3. Let a best be the attribute with the highest normalized information gain.

2. Random Forest(RF) Algorithm:

This is a decision tree based learning approach which help in classification with high computation time. This help in classification, regression and solve the problem of classification. A tree bagging, bootstrapping aggregation approach is provided with the given scenario [10].

The following pseudo code is used to build classification and decision tree using the random forest algorithm

1. Randomly select "K" features from total "m" features where k<<m
2. Among the "K" features, calculate the node "d" using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat the a to c steps until "l" number of nodes has been reached
5. Build forest by repeating steps a to d for "n" number times to create "n" number trees

3. Random Tree(RT) Algorithm:

Random tree based approach is again a tree based classification approach help in classification. This algorithm follows the stochastic process of mining which follow the uniform spanning tree for the classification and follows the weight based approach [11].

The following is the pseudo code which is followed by the random tree based algorithm which is lightning tree algorithm for classification.

- When inserting or searching for an element in a binary search tree, the key of each visited node has to be compared with the key of the element to be inserted or found.
- The shape of the binary search tree depends entirely on the order of insertions and deletions, and can become degenerate.
- After a long intermixed sequence of random insertion and deletion, the expected height of the tree approaches square root of the number of keys, $\sqrt{n}$, which grows much faster than log n.
- There has been a lot of research to prevent degeneration of the tree resulting in worst case time complexity of log n.

Thus these are the J48, Random Forest and Random tree algorithm help in efficient classification used over Psychiatric disease dataset.

VI. PROPOSED ARCHITECTURE

In order to perform research over psychiatric disease dataset, the following execution is performed which is presented in given flow diagram.A following flow diagram shows the simulation setup using Weka tool and performing pre-processing on collected dataset. A proposed flow diagram shows the data flow.

A figure 1 below is shows as proposed work flow executed by us. A data classification over mental disease and finding its usability is executed.

ISSN: 2454-4248
241 – 246

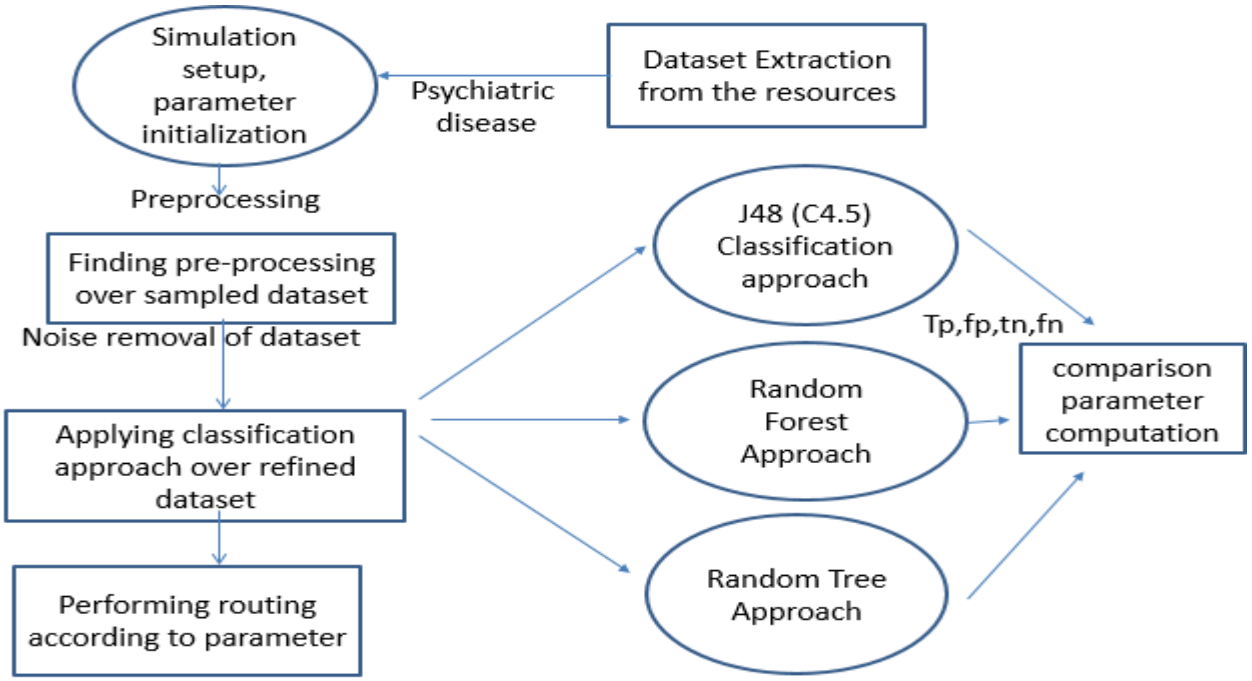

Figure 1: Flow diagram for the proposed work flow which is executed over Weka tool

VII. EXPERIMENT & RESULT EVALUATION

In order to perform simulation over the dataset, Weka tool is used on the i3 machine, 4 GB RAM configuration system. The Weka tool, which have used for analysis. Now next step shows processing done with Weka tool and parameters generator by tool and their analysis is performed.

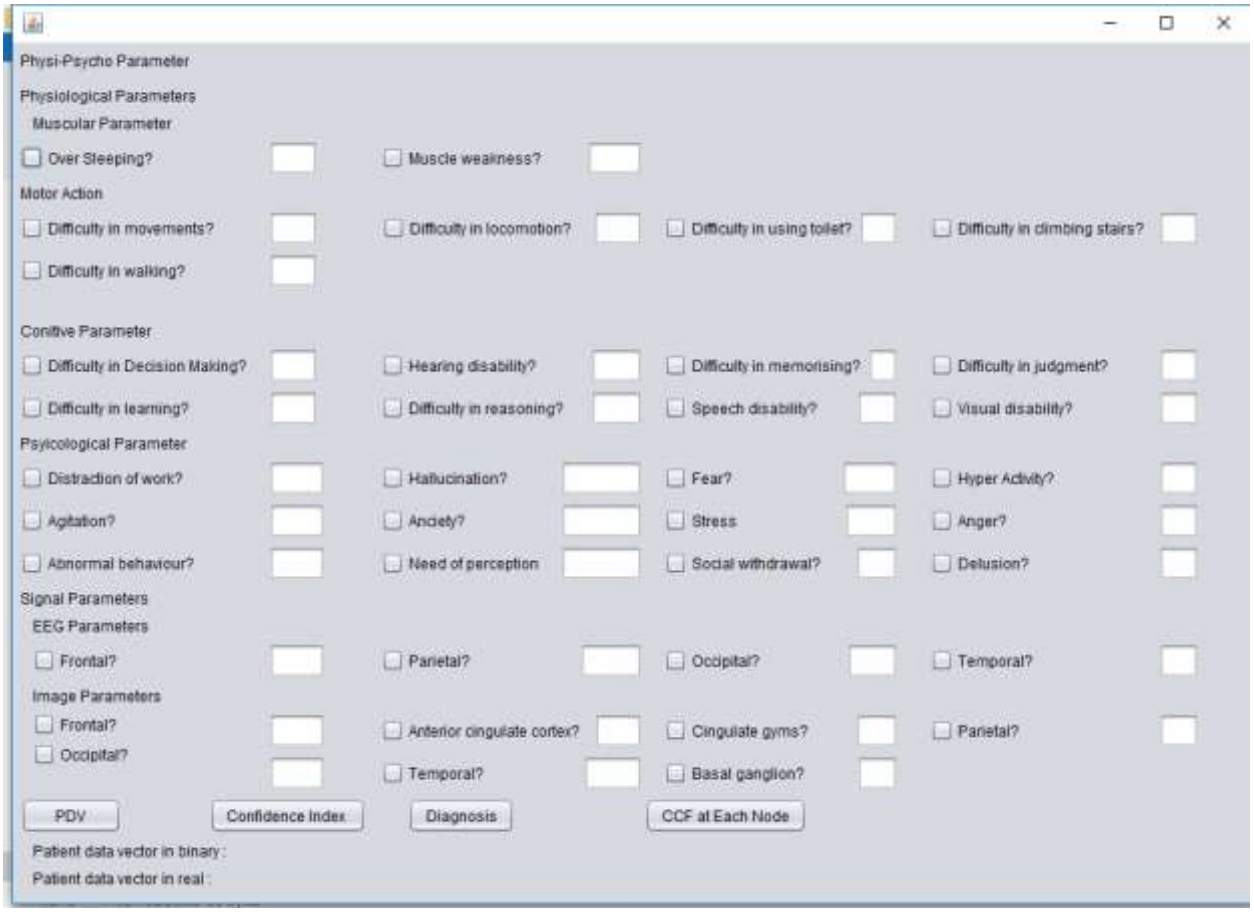

Figure 2: Applying Factors Symptoms and Generating PDV Algorithm

 ISSN: 2454-4248
241 – 246

The above figure 2, consists of several parameters they are physic-psycho parameter, physiological, conative parameter, psychological parameter, signal parameter, image parameter which then subdivided into their respective actions.

Statically Computed Result Comparison: According to the simulation execution with three major classification technique, a comparison table is drawn. The below table present a comparison between three different algorithm using ROC curve, precision, recall and F-Measure as parameter for comparison and computation.

Dataset	Algorithm System	ROC Curve %	Precision %	Recall%	F-Measure
Psychiatric-Dataset 1	J48	99.6	87.2	94.9	90.9
	RF	99.4	90	91.1	90.6
	RT	99.6	86.4	96.2	91

Table 1: A statically comparison analysis over discussed technique

As per the statistical result in table 1, further a comparison is made individually using graph by which a proper monitoring and observation can be made.

Graphical Analysis: A graphical analysis is presented below which shows the efficiency of different classification approach.

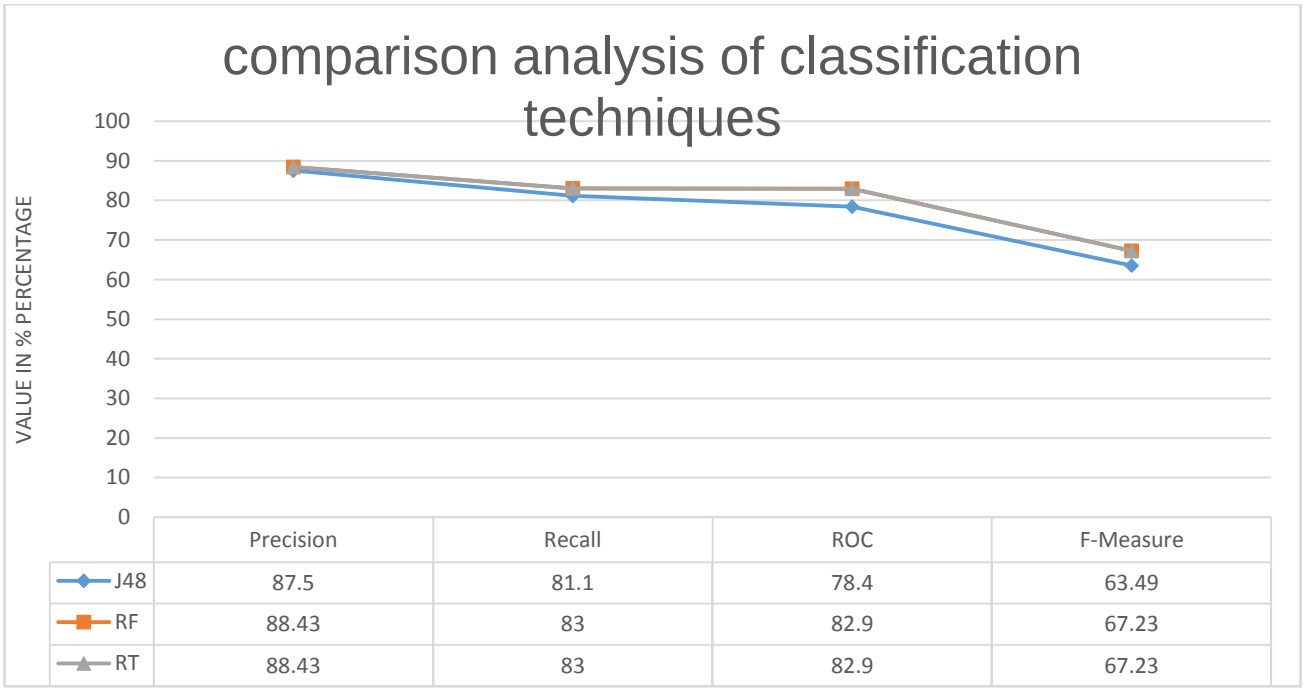

Figure 2: A graphical comparison analysis of classification approach

In the figure 2 above, a classification performance is shown. This graph shows the efficiency of Random tree over Random forest and J48 (C4.5) algorithm.

VIII. CONCLUSION & FUTURE WORK

Data mining and classification is an important concept while dealing with the large mismatch dataset. Psychiatric disorder is an aspect which deals in finding mental disorder, its symptom and classification of medical data according to its condition. It help in getting proper data analysis and decision making accordingly. Previous author worked with limited disease and limited order of algorithm signature. Thus they used technique such as ANN and heuristic based approach for mental disorder relevant entity finding. They have discussed rule based model and finding its usability with matching factor analysis. A set of rule definition is

 ISSN: 2454-4248
241 – 246

takes for classification and apply the rules over dataset. Thus the work is performed in finding dataset, its classification, defining set of rules and further finding an exact common factor which reveals the mental disorder disease. Further a CCF computation of all the disease is get performed over the dataset. Usable data help in finding disease and their impact on mental health disorder. Thus by applying clinical dataset, finding CF and determining the symptom is the major contribution of our work. Classification approach which is C4.5 (J48) and RF (Random forest) is taken for the efficient classification and analysis of technique. These technique belongs to tree based approach which gives data efficiency, better classification and disease symptom finding.

REFERENCES

[1]. MohitGangwar, R. B. Mishra, R. S. Yadav, B. Pandey, "Intelligent Computing Methods for The Interpretation of Neuropsychiatric Diseases Based on Rbr-Cbr-Ann Integration", International Journal Of Computers & Technology, Vol 11, No. 5, 2013.

[2]. American Psychiatric Association. "Diagnostic and statistical manual of mental disorders" (5th ed.). Arlington, VA: American Psychiatric Publishing, (2013).

[3]. Piers Johnson , Luke Vandewater , William Wilson, Paul Maruff, Greg Savage , Petra Graham5 , Lance S Macaulay , Kathryn A Ellis, Cassandra Szoeke , Ralph N Martins, Christopher C Rowe, Colin L Masters, David Ames, Ping Zhang, "Genetic algorithm with logistic regression for prediction of progression to Alzheimer's disease", Johnson et al. BMC Bioinformatics 2014, 15(Suppl 16):S11.

[4]. Francesca Pistollato, Elan L. Ohayon, Ann Lam," Alzheimer disease research in the 21st century: past and current failures, new perspectives and funding priorities", Oncotarget, Advance Publications 2016.

[5]. Ms. Sumathi M.R, Dr. B. Poorna, "Prediction of Mental Health Problems Among Children Using Machine Learning Techniques", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, 2016.

[6]. MohitGangwar, R. B. Mishra, R. S. Yadav, B. Pandey, "Intelligent Computing Method for the Interpretation of Neuropsychiatric Diseases", International Journal of Computer Applications (0975 – 8887) Volume 55– No.17, October 2012.

[7]. Richard G Jackson, Rashmi Patel, NishamaliJayatilleke, Anna Kolliakou, Michael Ball, Genevieve Gorrell, Angus Roberts, Richard J Dobson, Robert Stewart," Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project", Jackson RG, et al. BMJ Open 2017;6:e012012. doi:10.1136/bmjopen-2016-012012.

[8]. Yevgeniya Kovalchuk1 , Robert Stewart, Matthew Broadbent , Tim J. P. Hubbard, Richard J. B. Dobson," Analysis of diagnoses extracted from electronic health records in a large mental health case register", PLOS ONE | DOI:10.1371/journal.pone.0171526 February 16, 2017.

[9]. Ian H. Witten; Eibe Frank; Mark A. Hall (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco. p. 191.

[10]. Prinzie, A., Van den Poel, D. (2008). "Random Forests for multiclass classification: Random MultiNomialLogit". Expert Systems with Applications. 34 (3): 1721–1732. doi:10.1016/j.eswa.2007.01.029.

[11]. Cook, Matthew; Soloveichik, David; Winfree, Erik; Bruck, Jehoshua (2009), "Programmability of chemical reaction networks", in Condon, Anne; Harel, David; Kok, Joost N.; Salomaa, Arto; Winfree, Erik, Algorithmic Bioprocesses, Natural Computing Series, Springer-Verlag, pp. 543–584, doi:10.1007/978-3-540-88869-7_27.