A Novel approach for Psychiatric Patient Detection and Prediction using Data Mining Techniques

Dr. E. Chandra Blessie
Dept. of Computer Science
Nehru College of Management
Coimbatore

Bindu George
Dept. of Computer Science
Nehru College of Management
Coimbatore

***Abstract*— Classification is an important data mining technique with broad application to classify different kind of data used in every field of our day to day life. Several algorithms are developed to extract information and discover knowledge patterns that may be useful for decision support. The medical diagnosis process can be interpreted as a decision-making process during which the physician convinces the diagnosis of a new and unknown case from an available set of medical data. In the last few decades, many researchers have focused on developing effective methods for intelligent psychiatric disorder prediction and decision support system. Psychiatric care remains a notable exception that heavily relies on patient interviews and self-assessment. This is because mental illnesses manifest themselves mainly in the way patients behave throughout their daily life. Several data mining algorithms with better classification accuracy will provide more enough information to identify the psychiatric patients. The objective of this study is to make comparative evaluation of classifiers Naïve Bayes and K-Nearest Neighbor in the context of psychiatric patient dataset. The extracted data patterns can provide useful information to prevent mental illness and assist in delivery of efficient mental health services.**

Keywords*— *Mental disease monitoring; Data Mining; Classification; K-Nearest Neighbor; Naive Bayesian

I. INTRODUCTION

Data mining is a knowledge discovery technique to analyze data, identify hidden patterns and encapsulate it into useful information [1]. Data mining is a group of processes that is based on automated searching for actionable knowledge buried within a huge amount of data to extract information for the purpose of making predictive models for decision making and new discoveries [2]. Predictions and descriptions are principal goals of data mining, in practice [3]. Prediction in data mining involves attributes or variables in the data set to find an unknown or future state values of other attributes [4]. Description emphasize on discovering patterns that explains the data to be interpreted by humans [3]. Data mining problems are generally solved by using various approaches from computer science and from statistics, including hypothesis testing and regression techniques [5]. Performing data mining reveals useful relationship existed among data, and this rule can apply for right decision making [6],[7].

The study of classification involves the discovery of hidden patterns from existing medical data to identify the boundary between healthy and unhealthy individuals [8]. In health domain, the data are created from a variety of devices within a short time span, the characteristics of these data are that they are stored in different formats and created quickly, which can, to a large extent, be regarded as a big data problem. Mental and substance use disorders can have a powerful outcome on the health of individuals, their families, and their communities. Mental, and emotional disorders are an obvious application field for activity recognition. As the symptoms of such diseases manifest themselves in changes of behavior [9], activity aware systems could be used as an important instrument for assisting diagnosis and treatment. Even more, the fact that psychiatrists currently have few objective and reliable alternatives would amplify the value of such a system.

This research paper has presented a psychiatric disordered prediction system using different data mining techniques namely Naïve Bayes (NB), and K-Nearest Neighbor (KNN). The results show that each technique has its own advantages in achieving the objectives of psychiatric disorder diagnosis with high accuracy. The recognition that mental health is costly, and many cases will not become chronic if treated early has led to an increase in research in the last 20 years. The probability of patients getting psychiatric disorder is determined using different attributes such as attitude, habits, behavior, family history, orientation etc.

II. BACKGROUND AND RELATED WORKS

Data mining can be used in medical field to enhance our understanding of health to focus on identifying, extracting and evaluating attributes related to medical data set. Classification data mining methods originate from different research fields and often use various modeling approaches. There are various complications during algorithm development. Many algorithms make decisions by finding associations, classifying and predicting. One possible approach is to develop systems that predict patient state by using predefined algorithms that are initialized based on evidence from scientific or clinical knowledge [10], [11]. This has been the typical approach of systems that recognize patient activities. Different machine learning techniques, such as classification and regression tree method, Bayesian

hierarchical[12] and Support Vector Mechanism[13] have been used for medical diagnosis process based on the extracted features derived from various attributes. Fitness report and demographic details of patients is also useful for utilizing the available hospital resources effectively.

In the area of mental health, most systems deployed to date focus on supporting self-monitoring. Systems that provide patient feedback through questionnaires or text messages are analyzed in [14] and [15]. An automated tool using data mining is proposed by J.Alapont *et al.*, for managing hospital resources such as physical and human resources [16]. Oxtext [17] were developed to log self-reported mood, activities, and quality of sleep in order to monitor depression and state changes. American Health ways system construct a predictive model using data mining to recognize the patients having high risk. Using Predictive model, healthcare provider recognizes the patient which require more concern as compare to other patients [18].

The researchers [19] used the data mining algorithms decision trees, naïve bayes, neural networks, association classification and genetic algorithm for predicting and analyzing various disease from the dataset. Various algorithms with data extraction, human decision is an essential component of each stage of the development and interpretation, including choosing criteria and defining assumptions, optimization functions, and selecting training data[20,21]. A few set of studies that relied on self-monitoring of patients with severe mental illness, specifically bipolar disorder and schizophrenia are presented in [22] and [23]. It revealed that while self-assessment of patient state has a positive effect in increasing emotional self-awareness in patients suffering from depression, anxiety, and stress, the mental health outcomes did not improve significantly.

The researchers [24] uses decision trees, naïve bayes, and neural network to predict heart disease with 15 popular attributes as risk factors listed in the medical literature. One possible approach is to develop systems that predict patient state by using predefined algorithms that are initialized based on evidence from scientific or clinical knowledge [25, 26]. In the area of mental health, most systems deployed to date focus on supporting self-monitoring. Data Mining also helps them to identify the side effects of treatment, to make appropriate decision to reduce the hazard and to develop smart methodologies for treatment.

III. PREPARE YOUR PAPER BEFORE STYLING

In this paper, two classifiers, Naive Bayes algorithm and K-Nearest Neighbour algorithm are used for classification and comparison. Comparison is made on accuracy, sensitivity and specificity using true positive and false positive in confusion matrix generated by the respective algorithm. Also, we can use the correct and incorrect instances that give us a most efficient method for classification by using the confusion matrix.

A. K-Nearest Neighbor Algorithm

The k-Nearest Neighbor algorithm (k-NN) is a technic for classifying objects based on closest training examples in the feature space. The k-Nearest Neighbor algorithm is a type of instance-based learning, or lazy learning, where the function is only approximated locally, and all computation is postpone until classification. The k-Nearest Neighbor algorithm is amongst the easily understood of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors. The best choice of k depends upon the data; generally, larger values of k reduce the effect of noise on the classification but make boundaries between classes less distinct.

Once the nearest-neighbor list is obtained, the test object is classified based on the major part of its nearest neighbors There are several key issues that affect the performance of k-NN. If k is too small, then the result can be sensitive to noise points. On the other hand, if k is too large, then the neighborhood may include too many points from other classes. Another issue is the approach to combining the class labels. The easiest method is to take a majority vote, but this can be a problem if the nearest neighbors vary widely in their distance and the closer neighbors more reliably indicate the class of the object. K-NN classification is an easy to understand and easy to implement classification technique.

B. The Naive Bayesian Algorithm

The Naive Bayesian algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in the given data set. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods. The Naive Bayes is a quick method for creation of statistical predictive models [27]. The algorithm, being a simple one, is used in a variety of their fields of science. They encourage researchers to try this algorithm first [28], before applying more difficult and complex solutions. The naivety of the algorithm results from the fact that it assumes the variables (attributes) to be independent. This assumption, being not true in most real-life situations, usually delivers a model with a good predictive power [27]. The learning process relies on counting correlations (combinations) of each of the values of each attribute with each value of a class. The probability of an instance E consisting of attributes' values $n1$, $n2$,...,nk to be classified as belonging to a class H is estimated with the use of Naive Bayesian equation.

IV. DATASET INFORMATION

In order to develop a system, we conducted a challenging real world study , and collecting data from different areas in Kerala. The number of participants included in the study was limited by different factors. All of them underwent one or more changes in their mental state during the study. Standard dataset from different areas of Kerala has been used for training and testing purpose. The database contains 36 attributes, but we have used some of them in order to obtain the accurate results using less number of feature space. The dataset contains a total of 65 instances, of which some are healthy and other instance have psychiatric disease. The survey used uniform sample design, field protocol for data collection and physical measurements to facilitate

comparability across the state and to ensure high quality data. The data was collected using a questionnaire. Two types of questionnaire - one at household level and another for individual level were used for the survey. In the light of success for different data mining techniques, and specifically ensemble techniques, it is very beneficial to consider ensemble techniques for the disease diagnosis and prediction. Therefore, we have proposed an ensemble framework based on majority voting scheme that combines individual classifiers and achieves higher accuracy for diagnosis of highly psychiatric. Table1 shows the selected psychiatric patients dataset attributes.

TABLE 1 INDIVIDUALS RELATED DATA

Variable	Description	Possible Values
FHS	Family History	{ yes, no }
MST	Marital Status	{ single, married, widowed, separated }
INS	Insights	{ yes, no }
PAT	Patients Attitude	{positive, negative, neutral}
HAB	Habits	{yes, no}
AGE	Age classification	{ 14 - 29, 30 – 44, 45 – 60, above 60
ELE	Energy Level	{healthy, unhealthy, average}
RCY	Religiosity	{very high, high, medium, low}
BHV	Behavior	{Exhibitionism, Anxiety, silent, violent}
PQU	Qualification	{middle, High school, secondary, graduate, post-graduate, higher study}
ORA	Orientation	{No,Yes}

Classification accuracy is usually calculated by determining the percentage of tuples placed in a correct class. This avoids the fact that there also may be a cost associated with an incorrect assignment to the wrong class. A confusion matrix explains the accuracy of the solution to a classification problem. This matrix contains information about actual and predicted classifications done by a classification system.

We have performed classification using Naïve Bayes algorithm and K-Nearest Neighbor algorithm on the given dataset in weka tool. The weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. It offers a well-defined framework for experimenters and developers to build and evaluate their models.

The accuracy, sensitivity and specificity of the classifiers are measured to evaluate the performance individual classifiers. Sensitivity indicates the number of persons that are correctly classified healthy in the dataset whereas specificity indicates the proportion of patients that are correctly classified as sick. Accuracy measures the proportion of correct predictions made by proposed framework against actual class label for test data.

TABLE 2 CONFUSION MATRIX OF INDIVIDUAL CLASSIFIERS

Classifier	Class	Psychiatric	Healthy
Naïve Bayes	Psychiatric	13	2
	Healthy	4	46
KNN	Psychiatric	13	2
	Healthy	15	35

The results of sensitivity, specificity and accuracy for the two individual classifiers are shown in the figures.

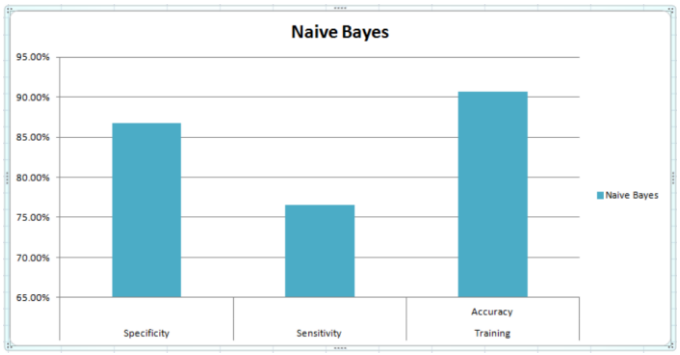

FIG. 1 GRAPHICAL REPRESENTATION OF NAÏVE BAYES CLASSIFICATION

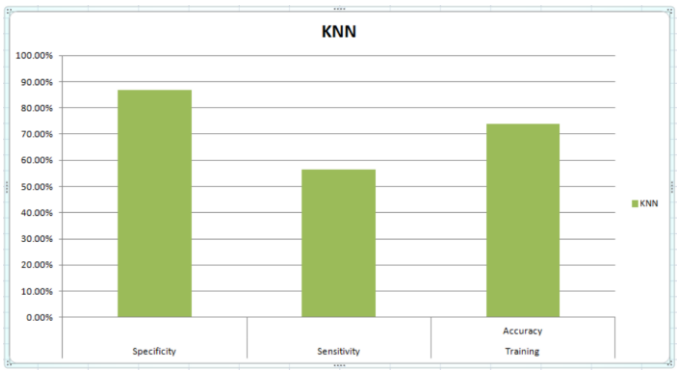

FIG. 2 GRAPHICAL REPRESENTATION OF KNN CLASSIFICATION

Naïve Bayes classification has been able to build a model with greatest accuracy since the model prediction accuracy is 90.71%. The sensitivity and specificity of Naïve Bayes classification is 76.50% and 86.71% respectively. Model accuracies obtained from KNN classifiers is 73.85%.

V. CONCLUSION

Encouraging mental health and preventing mental disorders are one of the important areas to reduce the impact of behavioral health conditions in the world. The main objective of proposed research is to develop an intelligent mental disorder diagnosis and prediction system using various classifiers namely NB and KNN. Today's health care is difficult to imagine without the possibility to objectively measure various physiological parameters related to patients' symptoms. The research undertook an experiment on application of two data mining algorithms to predict the psychiatric disorder and to compare the best method of prediction. In this paper, the accuracy of classification

techniques is evaluated based on the selected classifier algorithms. Based on algorithm analysis, Naïve Bayes classifier has achieved the best performance. K-Nearest Neighbor classifier also showing good results.

The experiment can serve as an important tool for physicians to predict risky cases in the practice and advise accordingly. The model from the classification will be able to answer more complex queries in the prediction of psychiatric disorder condition. This study can help medical practitioners to make intelligent medical decisions more accurately that were not possible with traditional decision support systems. Moreover, the treatment cost can also be reduced by providing effective treatment. There are different issues that influence the performance of applied models including type of problem and type of input data.

REFERENCES

[1] A. K. Sen, S. B. Patel, and D. P. Shukla, "A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-FuzzyIntegrated Approach Two Level," International Journal of Engineering and Computer Science, vol. 2, no. 9, pp. 1663–1671, 2013.

[2] P. Giudici, Applied Data Mining Statistical Methods for Business and Industry, Wiley & Sons, 2003.

[3] R. Rao, "SURVEY ON PREDICTION OF HEART MORBIDITY USING DATA MINING TECHNIQUES," International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 1, no. 3, pp. 14–34, 2011

[4] S. Vijiyarani and S. Sudha, "Disease Prediction in Data MiningTechnique – A Survey," International Journal of Computer Applications & Information Technology, vol. II, no. I, pp. 17–21, 2013.

[5] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Data mining and knowledge discovery in databases, Commun. ACM 39 (1996) 24–26.

[6] Vijayajothi P, Tan SY, Sarinder KD, Amandeep SS. A methodological review of data mining techniques in predictive medicine: An application in hemodynamic prediction for abdominal aorticaneurysm disease. Published by Elsevier, Biocybernetics and Biomedical Engineering, 2014;34(3):139-145.

[7] K.C.Tan, E.J.Teoh, Q.Yu,K.C. Goh.A hybrid evolutionary algorithm for attribute selection in datamining. Expert Systems with Applications, 2009; 36:8616-8630.

[8] Palaniappan, S., Awang, R.: Intelligent Heart Disease Prediction System Using Data Mining Techniques. 978-1-4244-1968-5/08/ ©IEEE (2008)

[9] American Psychiatric Association. (2013, Jun.). Diagnostic and Statistical Manual of Mental Disorders (5th ed.) [Online]. Available: dsm. psychiatryonline.org

[10] A. Honka, K. Kaipainen, H. Hietala, and N. Saranummi, "Rethinking health: ICT enabled services to empower people to manage their health," IEEE Rev. Biomed. Eng., vol. 4, pp. 119–139, Nov. 2011.

[11] M. Morris and F. Guilak, "Mobile heart health: Project highlight." IEEE Pervasive Comput., vol. 8, no. 2, pp. 57–61, Apr.–Jun. 2009.

[12] N. F. Garcia, P. Gomis, A. La Cruz, G. Passeriello, F. Mora, "Bayesian hierarchical model with wavelet transform coefficients of the ECG in obstructive sleep apnea screening", Comput. Cardiol., vol. 27, pp. 275-278, 2000.

[13] A. H. Khandoker, M. Palaniswami, C. K. Karmakar, "Support vector machines for automated recognition of obstructive sleep apnea syndrome from ECG recordings",IEEE Trans. Inf. Technol. Biomed., vol. 13, no. 1, pp. 37-48, Jan. 2009.

[14] E. Granholm, D. Ben-Zeev, P. Link, K. Bradshaw, and J. L. Holden, "Mobile assessment and treatment for schizophrenia (MATS): A pilot trial of an interactive text-messaging intervention for medication adherence, socialization, and auditory hallucinations." Schizophrenia Bull., vol. 38, no. 3, pp. 414–425, 2012.

[15] S. Reid, S. Kauer, S. Hearps, A. Crooke, A. Khor, L. Sanci, and G. Patton, "A mobile phone application for the assessment and management of youth mental health problems in primary care: A randomised controlled trial." IEEE Commun. Mag., vol. 12, no. 1, p. 131, Nov. 2011.

[16] J. Alapont, A. Bella-Sanjuán, C. Ferri, J. Hernández-Orallo, J. D. Llopis-Llopis and M. J. Ramírez-Quintana, "Specialised Tools for Automating Data Mining for Hospital Management", http://www.dsic.upv.es/~abella/papers/HIS_DM.pdf, (2005).

[17] Oxtext. (2014). True colours-improved management for people with biplar disorder. [Online].

[18] M. Ridinger, "American Healthways uses SAS to improve patient care", DM Review, vol. 12, no.139, (2002).

[19] K. Sudhakar, "Study of Heart Disease Prediction using Data Mining," vol. 4, no. 1, pp. 1157–1160, 2014.

[20] Diakopoulos N., Accountability in algorithmic decision making Commun ACM;59:56-62.

[21] Jagadish HV, Gehrke J, Labrinidis A, et al, Big data and its technical challenges. Commun ACM. 2014;57:86-94.

[22] C. Depp, B. Mausbach, E. Granholm, V. Cardenas, D. Ben-Zeev, T. Patterson, B. Lebowitz, and D. Jeste, "Mobile interventions for severe mental illness: Design and preliminary data from three approaches." J. Nervous Mental Dis., vol. 198, no. 10, pp. 712–721, 2010.

[23] E. Granholm, D. Ben-Zeev, P. Link, K. Bradshaw, and J. L. Holden, "Mobile assessment and treatment for schizophrenia (MATS): A pilot trial of an interactive text-messaging intervention for medication adherence, socialization, and auditory hallucinations." Schizophrenia Bull., vol. 38, no. 3, pp. 414–425, 2012.

[24] K. Srinivas, K. Raghavendra Kao, and A. Govardham, Analysis of coronary heart disease and prediction of heart attack in coal mining regions using data mining techniques," in The 5th International Conference on Computer Science & Education, 2010, pp. 1344– 1349.

[25] A. Honka, K. Kaipainen, H. Hietala, and N. Saranummi, "Rethinking health: ICT enabled services to empower people to manage their health," IEEE Rev. Biomed. Eng., vol. 4, pp. 119–139, Nov. 2011.

[26] M. Morris and F. Guilak, "Mobile heart health: Project highlight." IEEE Pervasive Comput., vol. 8, no. 2, pp. 57–61, Apr.–Jun. 2009.

[27] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Data mining and knowledge discovery in databases, Commun. ACM 39 (1996) 24–26.

[28] Hosmer, D.W. and Lemeshow, S. Applied Logistic Regression. 2nd ed. Wiley, New York; 2000