A Survey on Data Mining Algorithms in Prediction of Psychiatric Disorders

E. Chandra Blessie[1], Bindu George[2]

[1]Assistant Professor, Department of Computing (AIML), Coimbatore Institute of Technology, India
chandrablessie[at]gmail.com

[2]Research Scholar, Department of Computer Science, Nehru College of Management, Coimbatore, India
Tessgeorgesh[at]email.com

Abstract: ***The medical diagnosis and automated decision making process by computer algorithms based on data from our behaviors is fundamental to the digital economy. Researchers have been using various data mining techniques and statistical analysis to improve the disease diagnosis accuracy in medical healthcare. Psychiatric disorder is a highly prevalent condition associated with many adverse health problems. Several data mining algorithms with better classification accuracy will provide more sufficient information to identify the severe psychiatric disorders. The objective of the proposed research is the comparison of different data mining algorithms and to predict the acute psychiatric disordered patients more accurately. After feature analysis, models by five algorithms including C5.0, Neural Network, Support Vector Machine (SVM), K - Nearest Neighbor (KNN) and Naïve bayes developed and validated. C5.0 Decision tree has been able to build a model with greatest accuracy 90.77%, KNN, SVM, Neural Network and Naïve Bayes have been 73.85%, 83.08%, 77.93 and 90.71% respectively.***

Keywords: Data mining, Support vector machine, Naive Bayes, Decision tree, K - Nearest Neighbor

1. Introduction

With the advancement of science, the volume of accumulated data in different fields has been increased that it is well known the explosion of information [1]. When analyzing the accumulated data, they could reveal their hidden useful information. Data mining is a process of analyzing and identifying previously unknown and hidden patterns, relationships and knowledge from large datasets that was not possible with traditional techniques [2]. Performing data mining reveals useful relationship existed among data, and this rule can apply for right decision making [3], [4]. Algorithms apply complex techniques and statistics to create models that make decisions. One of the most predictive data mining techniques is classification. Predictive models have the specific aim of allowing us to predict the unknown values of variables of interest given known values of other variables. Neural Network, support vector machine, Naive bayes and decision Tree are different form of classification algorithms [5].

In medical domain, the major challenges that healthcare organizations are facing is provision of quality services. Quality service points to diagnosing a patient accurately and then providing proper treatment. An intelligent computer - based information and decision support system can achieve great accuracy in prediction and classification of psychiatric disorder. Data mining plays an important role in psychiatric field. The study of classification involves the discovery of hidden patterns from existing clinical data to identify the boundary between the psychiatric and healthy individuals [6]. Automated decision making algorithms impact the daily life of those with and without mental illness.

Mental Disorder is a very common mental health problem worldwide. The World Health Organization estimates that 121 million people currently suffer from depression, with 5.8% of men and 9.5% of women experiencing a depressive episode in any given year [7]. In light of these high rates of depression, it is a cause for concern that mood disorders are the most common psychiatric condition associated with suicide [8]. This article utilizes district, block and village level data on diagnoses and treatment of mental illness. The study conducted on various sources of public information on mental health and substance use disorder identifies and treatment. Mental illness is not equally distributed across the district, with higher rates of serious psychological stress and major depressive episodes in rural area, as compared to developed area. For this study real world data's are collected from various hospitals in Idukki district.

2. Related Works

There are many challenges during algorithm development. Many algorithms make decisions by finding associations, classifying and predicting. Different machine learning techniques, such as classification and regression tree method, Bayesian hierarchical [9] and Support Vector Mechanism [10] have been used for medical diagnosis process based on the extracted features derived from various attributes. 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 [11, 12]. Hlaudi Daniel et al. [13] used various classification algorithms for risk prediction in medical field specifically for heart disease. In this paper J48 has better accuracy of 99.07% when compared to other algorithms like Naive bayes – 97.22% and Bayes Net – 98.14%. Peter and Somasundaram [14] used pattern recognition and data mining techniques for risk prediction in medical field specifically for cardiovascular disease. It is concluded that Naive Bayes has better accuracy as compared to other algorithms. The proposed technique requires the input attribute set in ASCII file format and use of only numerical attributes. Aditya Methaila et al. [15] used various algorithms and combinations for effective heart attack prediction. In this paper Decision Tree has performed with 99.62% accuracy by using 15 attributes compared to Naive Bayes 96.53 % and Artificial Neural Network 88.3%. The researchers [16] used the data mining algorithms decision trees, naive bayes, neural networks, association classification and genetic algorithm for predicting and analyzing heart disease from the dataset.

Many researchers have used decision trees and its combination with other algorithms to solve various biological problems [17]. Srinivas et al. [18] analyzed various data mining techniques including rule - based, decision tree, Naive Bayes and artificial neural network and stated that their proposed system can easily answer the complex queries for the pre - diction of heart attack. The proposed technique used only 15 attributes for prediction, and its accuracy varies from dataset to dataset. Purushottam et al. [19] proposed a heart disease prediction system using various classification algorithms. The proposed system helps medical practitioners in decision making based on various parameters and its accuracy is 86.7%. The accuracy and performance of various well known algorithms on Heart disease data set are SVM – 70.59%, C4.5 –73.53% and KNN – 76.47%. Ghumbre et al. [20] presented a heart disease prediction system using radial - based function network structure and support vector machine. The analysis shows that the results obtained from support vector machine algorithm are equivalently as good as radial - based function network. This technique is affected by data acquisition method used for input of dataset. Jabbar et al. [21] used an associative classification algorithm for the prediction of various diseases. The algorithm uses genetic approach for prediction which results in higher accuracy and interestingness. First, an associative classification is used to classify the dataset with labeled classes and rules are collected from the training dataset. These rules are then organized in a form to construct a classifier. The genetic algorithm solves the optimization problems efficiently. Chitra and Seenivasagam [22] adopted a supervised learning algorithm to predict heart disease in a patient at early stage. The proposed classifier is named as cascaded neural network (CNN) with hidden neurons. Nikhar S., [28] presented a review of disease prediction for healthcare system using data mining techniques. Different data mining classification techniques such as Naive Bayes and Decision Tree are used here with better accuracy. Some of the important data mining approaches in health care are given in Table1.

Table 1: Data Mining Approaches in Healthcare

Author/year/ reference	*Technique*	*Specificity (%)*	*Sensitivity (%)*	*Accuracy (%)*
Bashir Saba et al.2014 [23]	Decision tree	85.71	63.16	72.73
	Naive Bayes	92.86	68.42	78.79
	Support vector	78.57	73.68	75.76
Ghumbre et al.2011 [20]	Support vector	88.50	84.06	85.05
	Radial basis function	82.10	82.40	82.24
Tu et al.2009 [24]	J4.8 decision tree	84.48	72.01	78.9
	Bagging algorithm	86.64	74.93	81.41
Abdar M. et al.2015 [25]	C5.0	90.9	95.23	93.02
	Support vector	90.9	80.95	86.05
	K nearest neighbor	88.63	88.09	88.37
	Neural network	86.36	73.80	80.23
Chitra et al.2013 [22]	Cascaded neural	87	83	85
	Support vector	77.5	85.5	82
Shouman et al.2013 [26]	Gain ratio decision	81.6	75.6	79.1
	Naive Bayes	80.8	78	83.5
	K nearest neighbor	85.1	76.7	83.2
Bashir Saba et al.2014	Naive Bayes	76.82	77.51	76.63
	Support vector	87.94	74.95	78.56

3. Dataset Information

The database is related to the data set of psychiatric patients in different areas of Kerala. The database contains 32 attributes, but we have used only 11 of them in order to obtain the accurate results using less number of feature space. The survey used uniform sample design, field protocol for data collection and physical measurements to facilitate comparability across the state and also 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. Table2 shows the selected psychiatric patients dataset attributes.

Table 2: Patients Related Data

Variable	*Description*	*Possible Values*
Sex	Patients Gender	{Male, Female}
MST	Marital Status	{ single, married, widowed, separated }
LOC	Living Location	{village, town, city}
PAT	Patients Attitude	{positive, negative, neutral}
SAH	Substance Abused Habit	{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}
PFS	Patient family status	{Joint, Individual}
PAI	Family annual income	{ poor, medium, high}

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.

4. Proposed Framework

The proposed framework is based on a novel combination of two heterogeneous classifiers: Decision Tree and Naïve Bayes. Decision tree is like a flowchart where every non leaf node is test on an attribute, every branch of node represents the outcome of test and every leaf node is class label. Root node represents all data at start [29]. Decision tree classifier does not require any domain knowledge and uses tree like graph. It calculates the conditional probabilities for research analysis and chooses the best alternative traversing from root to leaf and indicates unique class separation [30]. The Naïve Bayes classifier focuses on the rule that presence or absence of a disease depends on a feature itself. It assumes that features ate independent of each other. Supervised learning algorithm can be used to train the probability model of Naive Bayes classifier [31].

4.1 Experiments and Results

We have used five classifiers and trained them using psychiatric dataset to classify them as psychiatric or severe psychiatric. The accuracy, sensitivity and specificity of the classifiers are measured to evaluate the performance of proposed ensemble framework with 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. Mathematically:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}}$$

Accuracy measures the proportion of correct predictions made by proposed framework against actual class label for test data. Mathematically:

$$\text{Accuracy} = \frac{\text{True Positives+True Negatives}}{\text{True Pos + False Pos + True Neg + False Neg}}$$

Decision tree generates crisp rules that are used to classify data into psychiatric or critical psychiatric individuals whereas Naïve Bayes and SVM are first used to train the classifier and then trained classifiers classify test data into two classes. The results of sensitivity, specificity and accuracy for the five individual classifiers are given in Table 3.

5. Results and Discussion

This section presents the experimental results and analysis done for this study. Here we aim to provide a comprehensive evaluation over different classification algorithms including: Decision Tree, Support Vector Machines, k - Nearest Neighbors, Neural Network and Naive Bayes. Data divided into train set and test set. The training set is used to build the classifier and test set used to validate it. Model development is conducted in two main steps including model fitness and model accuracy. To calculate the model fitness criteria we used the data of training set; however, to compute the model accuracy measurements, data of testing set is applied which is merely much more valuable to judge about our models accuracy. Related results of these experiments are demonstrated in Table 3.

Table 3: Comparison of proposed Ensemble Technique

Algorithms	*Specificity*	*Sensitivity*	*Training Accuracy*
C5.0	98.99 %	71.40 %	90.77 %
SVM	73.31 %	61.11 %	83.08 %
KNN	86.71 %	56.45 %	73.85 %
Naive Bayes	86.71 %	76.50 %	90.71 %
Neural Network	66.7 %	58.5 %	77.93 %

C5.0 Decision tree has been able to build a model with greatest accuracy since the model prediction accuracy is 90.77%. Model accuracies obtained from other classifiers are different as this value for SVM, KNN, Naïve Bayes and Neural network have been 83.08%, 73.85%, 90.71% and 77.93% respectively.

6. Conclusion

The widespread acceptance and growing dependence on technology and decision making, reflect the numerous positive impacts to society. However, people with mental illness may be most vulnerable to the risk related to errors and basics in algorithms. The rapid adaption of technology has blurred traditional boundaries, leaving unresolved what should be public versus private medical versus non - medical data and human versus machine decision making. The purpose of this study is comparison of different machine learning algorithm on prediction of more psychiatric disordered person with more accurately. In this study, KNN, SVM, C5.0, Neural network and Naïve Bayes were implemented on the given dataset of 14 attributes. Inconsistencies and missing values were also resolved before the model construction. Based on investigated methods, decision tree has achieved the best performance. There are different issues that influence the performance of applied models including type of problem and type of input data. Decision trees are able to generate understandable rules and can perform classification without requiring much computation and clearly indicate that which fields are most important for prediction or classification.

References

[1] Hamid Bagheri, Abdusalam Abdullah Shaltooki. BigData: Challenges, Opportunities and Cloud Based Solutions. International journal of *Electrical and Computer Engineering (IJECE)*, 2014; 5 (2): 340 - 343.

[2] Rajkumar, M.; Reena, G. S.: Diagnosis of heaer disease using data mining algorithm. Glob. J. Comput. Sci. Technol.1**0** (10) (2010)

[3] Vijayajothi P, Tan SY, Sarinder KD, Amandeep SS. A methodological review of data mining techniques in predictive medicine: Anapplicationinhemodynamicpredictionforabdominalaorticaneurysmdisease. *Published by Elsevier*, Biocybernetics and Biomedical Engineering, 2014; 34 (3): 139 - 145.

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

[5] Nikola K, Elisa C. Spiking neural network methodology for modelling classification and understanding of EEG spatio - temporal data measuring cognitive processes. *Information Sciences*, 2015; 294: 565 - 575.

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

[7] World Health Organization (2001), “Mental and Neurological Disorders, ” Fact sheet No.265.

[8] Jamison, K. R. (2000), “Suicide and Bipolar Disorder, ” *Journal of Clinical Psychiatry*, vol.61, Suppl 9, pp.47 - 51

[9] 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.

[10] 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.

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

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

[13] Masethe, H., Masethe A., “Prediction of heart disease using classification algorithms”, World congress on Engineering and computer science, October 2014, (WCECS) vol.2.

[14] Peter, T. J.; Somasundaram, K.: An empirical study on prediction of heart disease using classification data mining techniques. In: IEEE - International Conference on Advances in Engineering, Science and Management (ICAESM - 2012) March (2012)

[15] Methaila, A., Kansal, P., Arya, H. and Kumar, P., “Early heart disease prediction using data mining techniques” *Computer Science & Information Technology Journal*, 2014, pp.53 - 59.

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

[17] Abuhaiba, I. S. I.: Efficient OCR using simple features and decision trees with backtracking. Arab. J. Sci. Eng.**31** (2), 223–244 (2006)

[18] Srinivas, K.; Rani, B. K.; Govrdhan, A.: Applications of data min - ing techniques in healthcare and prediction of heart attacks. Int. J. Comput. Sci. Eng. (IJCSE) **02** (02), 250–255 (2010).

[19] Purushottam, Kanak Saxena, Richa Sharma, “ Efficient Heart Disease Prediction System”, Procedia Computer Science, Volume 85, 2016,

[20] Ghumbre, S.; Patil, C.; Ghatol, A.: Heart disease diagnosis using support vector machine. In: International Conference on Com - puter Science and Information Technology (ICCSIT’) Pattaya (2011)

[21] Jabbar, M. A., Chandra, P., Deekshatulu B. L., Heart Disease Prediction System using Associative Classification and Genetic Algorithm, International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies - ICECIT, (2012)

[22] Chitra, R.; Seenivasagam, V.: Heart disease prediction system using supervised learning classifier. Bonfring Int. J. Softw. Eng. Soft Comput.3 (1), 1–7 (2013)

[23] BashirSaba, Qamar Usman and Javed Muhammad.: An Ensemble based Decision Support Framework for Intelligent Heart Disease Diagnosis. International Conference on Information Society, - Society 2014.10.1109/i - Society.2014.7009056.

[24] Tu, M. C.; Shin, D.; Shin, D.: Effective diagnosis of heart disease through bagging approach. In: 2nd International Conference on Biomedical Engineering and Informatics, 2009. BMEI’09, pp.1–4. IEEE (2009)

[25] Abdar Moloud, Rostam Niakan Kalhori, Sharareh, Sutikno, Tole Subroto, Imam and Arji, Goli.: “Comparing performance of data mining algorithms in prediction heart diseses”.5.1569 - 1576. (2015).

[26] Shouman, M.; Turner, T.; Stocker, R.: Integrating clustering with different data mining techniques in the diagnosis of heart disease. J. Comput. Sci. Eng.20 (1) (2013)

[27] Bashir Saba, Qamar Usman, Khan Farhan and Javed, Muhammad. MV5: A Clinical Decision Support Framework for Heart Disease Prediction Using Majority Vote Based Classifier Ensemble. ARABIAN JOURNAL FOR SCIENCE AND ENGINEERING.10.1007/s13369 - 014 - 1315 - 0. (2014).

[28] Nikhar S., Karandikar A. M, “PREDICTION OF HEART DISEASE USING DATA MINING TECHNIQUES” - A Review, International Research Journal of Engineering and Technology; Volume: 03 Issue: 02, Feb - 2016.

Author Profile

Dr. E Chandra Blessie received her MCA from Manomaniam Sundaranar University, Tuticorin; M. Phil in Computer Science from Alagappa University, Karaikudi and Ph. D from from Karunya University, Coimbatore. She is currently working as an Assistant Professor in the department of Computer Science, Coimbatore Institute of Technology, Coimbatore. A noted python trainer, she has authored a book on " A practical Approach for Python Beginners" and has authored a commendable number of research papers in national/ international and reputed journals. Ms. Chandra Blessie was a recipient of ' Best Active Participation - Woman Member 2012 - 2013' award honoured by the Computer Society of India. Her current research interests include Mining Big Data, Cloud Computing and Preprocessing Techniques for Data Mining.

Bindu George received her MCA from Madurai Kamaraj University, Madurai; M. Phil in Computer Science from Bharathiar University, Coimbatore and MBA from Madurai Kamaraj University, Madurai. She is currently doing her Ph. D. in Nehru College of Management, Bharathiar University, Coimbatore. Her research interests include Data Mining and Machine Learning.