UCT JOURNAL OF RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY 2019(02)

Available online at http://journals.researchub.org

Predicting Effective Factors in Schizophrenia Using Data Mining

Dr. Dinesh Mavaluru[1]*****, *Dr. Jayabrabu Ramakrishnan*[2], *Dr. Azath Mubarakali*[3]

[1]*College of Computing and Informatics, Saudi Electronic University, Saudi Arabia,*
[2]*College of Computer Science and Information Technology, Jazan University, Saudi Arabia*
[3]*College of Computer Science, Department of CNE, King Khalid University, Saudi Arabia*

ARTICLE INFO

Article history:
Received 04 Dec. 2018
Accepted 60 Jan. 2019
Published 13 Feb. 2019

Keywords:
Data Mining, Schizophrenia, Decision Tree, Support Vector Machine, Neural Network

ABSTRACT

Data mining is a technique for discovering new knowledge from databases and the use of data mining in medicine is considered one of the most widely used fields of data mining. Schizophrenia is one of the most common illnesses that cause many financial and social damages to society due to the loss of individual performance. In this study, we will examine the most effective fields of predictor in schizophrenia and then predict the age of occurring schizophrenia. In this study, some common classification methods such as support vector machine, decision tree and neural network have been used. The results show that the support vector machine model has more efficiency than other models.

INTRODUCTION

Schizophrenia is a severe mental disorder that affects entire personality of individual, but all its symptoms can be controlled with medication. Symptoms of illness is varied in different people, but often include depression, strange behavior, irritability, excessive sleep or inability to sleep, forgetfulness, severe response to crises and even suicide. The mortality rate in schizophrenia patients is twice of normal individuals [1]. Rapid treatment of this illness can improve the pre-consciousness of patients and prevent many clinical symptoms in subsequent attacks. The more the treatment begins faster, the chance of full recovery increases and, in general, the faster treatment is associated with a better outcome for the patient. Data mining is performed with the intention of helping the physician to remove the disorder from the data. Research on data and finding useful results and models for diseases using data mining will help in the prevention and early detection of illness.

According to the above, it seems that a system that can be a solution as decision-making support in medical and psychiatric centers is necessary. To date, many studies have been conducted in relation to the factors affecting schizophrenia; most of them are conducted using statistical analysis and usually the proposed hypothesis has been tested using statistical techniques. But it can investigate the impact of multiple factors and obtain meaningful, frequent and hidden patterns among the collected data using the data mining tool without defining the hypothesis. Then the results of data mining can be used as a source of knowledge for decision support systems for treatment or prevention. In this study, it is tried to investigate the effect of different factors in occurring schizophrenia using data mining.

LITERATURE REVIEW

The volume of data stored in databases is rapidly increasing, and this huge volume of data stored includes valuable but hidden knowledge and any data can play an important role in the decision-making process. The number of data analysts increases with a rate lower than the volume of data stored, which is a reason for using automated methods in extracting knowledge from the data. Data mining is the main step of an extensive process called knowledge discovery from database. This process includes the application of several pre-processing methods to facilitate the application of the data mining algorithm and the post-processing methods to reprocess and promote the knowledge obtained. The essence of data mining is the extraction of knowledge from data using relatively automated mechanisms. But this phrase raises a question what kind of knowledge should be discovered and how this discovery will be realized. In data mining, we seek to discover knowledge with capability of prediction. In fact, man seeks to discover knowledge with a high degree of accuracy. This knowledge discovered should be understood by the user so that ultimately support him/her in the decisions. In most cases, if the discovered knowledge is merely like a black box and performs the prediction without any explanation, the user will not trust its accuracy [8]. By defining data mining as the process of discovering and analyzing data belonging to large collections using automated tool and based on meaningful patterns and rules [7], we understand that data mining is the use of data analysis tool to discover unknowns, valid patterns, and relationships of large collects of data [9].

In a study conducted in 2008, the kinematic features of pen motion in the handwriting of two groups of people with schizophrenia and healthy people were entered into a classification system of support vector machine and classification has been performed with 95.4% accuracy.

**Corresponding author:* d.mavaluru@seu.edu.sa

DOI: https://doi.org/10.24200/jrset.vol7iss02pp29-33

The result of this study indicate that the use of handwriting extracted patterns can replace conventional methods of the diagnosis of illness and brain imaging as a proper method for diagnosis of schizophrenia [3].

In a study using data obtained from Functional Magnetic Resonance Imaging (fMRI), Pouyan et al. presented a method for the diagnosis of schizophrenia. In this study, FMRI images were categorized into two groups of healthy and diseased individuals using Support Vector Machine Classification (SVM) and classification accuracy was 70%. The results of this study show that the proposed method has a proper classification accuracy [10].

Recently, little work has been performed from healthy people using genotypic information to classify patients with brain disorders. Research has shown that the support vector machine algorithm can classify both bipolar and schizophrenic individuals from high-precision individuals using gene expression data [11]. In a study in 2007, the components extracted from the force applied to the pen in three dimensions and for testing, made from pen made by themselves capable of recording force in three dimensions have been used, and a method to detect patients with brain disorders from healthy individuals provided. In this study, the components extracted from the force applied to the pen were categorized into two groups of healthy and diseased individuals, and the precision of classification has been 95.8% using the Support Vector Machine (SVM) classifier. The results of this study show that the proposed method has a proper classification precision [6]. Researchers used the artificial neural network in 2007 to classify patients with neuropathic symptoms. The result indicated the high capability of this technique in data analysis for classification of neuropathic diseases [2].

METHOD

In this study, the samples were selected randomly from the records for prediction, and also the variables or in other words, the attributes of the samples were identified and extracted according to the view of experts. In this study, about 260 cases of schizophrenia patients have been studies and collected in Kerman. The considered variables were extracted and collected from the related cases in accordance with Table (1). The age variable of occurring schizophrenia is considered as the target variable to predict the age of occurring schizophrenia that has three values of 1, 2 and 3.

Table 1. Predictive variables for modeling

Variables	
head injury in childhood	age
Injuries during childbirth	Gender
Age of occurring schizophrenia	Marital status
Age of mother of people with schizophrenia at their birth	History of mental illness in childhood
Environmental factors	Using addictive substance
History of mental illness in family and family members	Birth season
Age of father of people with schizophrenia at their birth	Deprivations of childhood

Preparing the data

The purpose of this step is to increase the quality of the data used so that we can provide a suitable dataset for the data modeling step. Data preparation includes all the activities required to make the final dataset required for the modeling tool from the raw data. These activities include attribute selection, data conversion and clearing, and more. In this step, missing data and out of date data have been cleared. Also, records with less than 50% of the information variables completed were excluded from the dataset.

The proposed algorithm

The algorithms used in data mining techniques try to find and present the closest model to the data characteristics. Models can be predictive or descriptive. In this study, software classification model algorithms including neural network, C&R, C5, and SVM were used to predict the factors affecting schizophrenia and the age of occurring illness. These methods are briefly described below.

C&R Tree Algorithm

This node decides tree that can predict or classify future observations by the help of it. This method uses the recursive partitioning method to separate records prepared into separate pieces on the basis of minimization of impurities at each step, node is pure when all node data place within one of the classifications specified in the target variable. The target or predicted variable can be defined as a domain or a class. (Only two subgroups are possible for each variable)

Algorithm C5.0

This algorithm is capable of making both a decision tree and a set of rules. This model works by branching out the sample based on a field that obtains the most information at each step. The final field should be classified. It is also possible to divide each branch into more than two subgroups at each stage.

Neural Network Algorithm

Neural networks are one of the most widely used predictive methods. A neural network is a simplified model that works based on the function of the human brain. The algorithm aims to simulate some of the simple functions of the human brain. Each node in the neural network is like a neuron in the human brain that the network of connecting these neurons performs complex learning tasks. The process of analyzing data in a neural network is like a black box. Neural networks are very useful in estimating and predicting [5].

SVM algorithm

The SVM algorithm is a robust algorithm based on the theory of statistical learning and it is one of the algorithms that place in the subgroup of prediction algorithms and its basis is the theory of statistical learning. SVM separates samples of different classes using a line called decision boundary. This decision boundary is called the support vector. Each decision boundary is restricted to two hyper plans that their distance from the decision boundary is the same, and the distance between the two hyper plans is the margin of cluster. The purpose of SVM clustering is to find a decision boundary with the maximum clustering margin. The decision boundary can be linear or nonlinear. This algorithm works well for high dimensional data [4].

MODELING

Using various classification algorithms on the dataset, appropriate models are presented that provide high accuracy rules and describe the information provided to the model. We also find acceptable results regarding the effect of various factors on the occurring schizophrenia. In this research, IBM Spss Modeler software was used for data mining. The dataset is divided into two parts of model education data and model test data. For this purpose, 75% of the database records were randomly selected as education data and 25% of the data were used as the model test dataset.

Evaluation of the classification model

The most common criterion for evaluating classification is the accuracy criterion, calculated according to formula (1). But this criterion is tended to the majority class for imbalanced data that has a larger member class. In these cases, in addition to accuracy, criteria such as recall, precision and F measure are used to evaluate the performance, which are calculated using the confusion matrix of Table (2) and using the following formulas.

Table 2. Confusion matrix for positive and negative tuples

Predicted class / Real class	**C1**	**C2**
C1	t_pos	f_neg
C2	f_pos	t_neg

(1) Accuracy $=\frac{t_{pos}+t\_neg}{t_{pos}+t_{neg}+f_{pos}+f\_neg}$

(2) Recall $=\frac{t_{pos}}{t_{pos}+f\_neg}$

(3) Precision $=\frac{t_{pos}}{t_{pos}+f_{pos}}$

(4) $\mathrm{F}-\mathrm{Measure}=2\,\frac{\mathrm{recall}\times\mathrm{precision}}{\mathrm{recall}+\mathrm{precision}}$

RESULTS

In this research, different models have been produced using data mining classification techniques. In this section, we discuss the results obtained from the implementation of models and their evaluation. For this purpose, the models produced are provided to analysis tools and their precision is evaluated.

Produced set of rules

As a result, by implementing the model obtained from the C&R tree, a set of the most effective rules discovered from a software view will be provided. Given that data mining is a tool producing hypothesis, the accuracy and degree of effectiveness of the rules obtained should be examined and interpreted by experts in the field. The rules of the C&R tree after removing the overlaps are as follows:

1. If a person has a mental illness in childhood, the probability of occurring schizophrenia at the age of 15 to 35 years is more (66.66%).

2. If a person's gender is male and has an addiction, the probability of occurring schizophrenia at the age of 15 to 35 years is more (100%).

3. If a person's gender is male, the probability of occurring schizophrenia at the age lower than 15 years is more (75%).

4. If a person's gender is female and married and she has a history of mental illness in her family, the probability of occurring schizophrenia at the age of 15 to 35 years is more (100%).

5. If a person's gender is female and single and she has a mental illness in childhood, the probability of occurring schizophrenia at the age over 35 years is more (100%).

6. If a person's gender is female and single and has no mental illness in childhood, the probability of occurring schizophrenia at the age lower than 15 is more (6.2%).

2-9 Set of rules extracted from tree C5.0

After implementing the C5.0 model, the set of rules obtained from this tree is as follows:
1-If a person is single and has an addiction, the probability of occurring schizophrenia at the age of 15 to 35 is more (92%).

2-If a person is single and has a mental illness in childhood and the members of his/her family have a mental illness and the season of his/her birth is winter, the probability of occurring schizophrenia is more at the age of 15 to 35 years (83.33%)

3. If a person is single and has a mental illness in childhood and the season of his/her birth is winter, the probability of occurring schizophrenia is more at the age of 15 to 35 years (83.33%)

4. If a person is single and she/he has been deprived in childhood and has a history of mental illness in family members, the probability of occurring schizophrenia is more at the age of 15 to 35 years (80%)

5. If a person is married and has a mental illness in his/her family members, the probability of occurring schizophrenia is more at the age of 15 to 35 years (62.96%)

6. If a person is married and has no mental illness in his family members and his gender is male, the probability of occurring schizophrenia is more at the age of 15 to 35 years (85.29%)

7. If a person is married and has no mental illness her family members and his gender is female and his birth season is winter, the probability of occurring schizophrenia is more at the age above 35 years (83%)

Evaluation of the model prepared by C&R

Implementing model C&R shows that the classification precision of this model for educational data is 80.65% and for test data is 68.57% which seems a proper precision. Examining the results of the model analysis shows that the model has correctly predicted 9 cases from 22 individuals belonging to category 1 and wrongly recognized 3 cases from 142 individuals from category 2 and correctly predicted 2 cases from 22 individuals from category 3 according to Table (3).

Table 3. Confusion matrix of C&R model

Current statues	**1**	**2**	**3**
1	9	13	0
2	3	139	0
3	1	19	2

Based on the data in Table (3), the accuracy, recall, precision and F criteria are presented in Table (4).

Table 4. Evaluation Criteria of C&R Model

class	F Measure	Recall	Precision
1	0.61	0.41	0.69
2	0.89	0.98	0.81
3	0.17	0.09	1
Avg	0.52	0.49	0.83

Evaluation of the model prepared by C5

The results of implementing C5.0 model show that the classification precision by this model for educational data is 82.8% and for test data is 67.14%. Examining the results of the model analysis shows that the model has correctly predicted 10 cases from category 1 and correctly predicted 139 cases from category 2 and correctly recognized 5 cases from category 3 according to Table (5).

Table 5. Confusion matrix of C5 model

Current status	1	2	3
1	10	12	0
2	2	139	1
3	0	17	5

Based on the data in Table (5), the accuracy, recall, precision and F criteria are presented in Table (6).

Table 6. Evaluation Criteria of C5 Model

Class	F-Measure	Recall	Precision
1	0.14	0.12	0.17
2	0.8	0.88	0.73
3	0	0	0
Avg	0.31	0.33	0.3

Evaluation of the model prepared by neural network

The classification precision of neural network model for educational data is 90.86% and for test data is 67.14% which seems a proper precision. Examining the results of the model analysis shows that the model has correctly predicted 16 cases belonging to category 1 and correctly predicted 140 cases from category 2 and correctly recognized 13 cases from category 3 according to Table (7).

Table 7. Confusion matrix of neutral network model

Current status	1	2	3
1	16	6	0
2	12	140	0
3	0	9	13

Based on the data in Table (7), the accuracy, recall, precision and F criteria are presented in Table (8).

Table 8. Evaluation Criteria of Neural Network Model

class	F-Measure	Recall	Precision
1	0.8	0.73	0.89
2	0.94	0.98	0.9
3	0.74	0.59	1
Avg	0.83	0.77	0.93

Evaluation of the model prepared by SVM

The results of implementing SVM model show that the classification precision by this model for educational data is 93.55% and for test data is 68.57% which seems a proper precision. Examining the results of the model analysis shows that the model has correctly predicted 19 individuals from category 1 and correctly predicted 138 individuals from category 2 and correctly recognized 17 cases from category 3 according to Table (9).

Table 9. Confusion matrix of SVM model

Current status	1	2	3
1	19	3	0
2	1	138	3
3	0	5	17

The accuracy, recall, precision and F criteria are presented in Table (1) and the overall comparison of the models used in this study is shown in Table (1).

Table 10. Evaluation Criteria of the SVM Model

Class	FMeasure	Recall	Precision
1	0.9	0.86	0.95
2	0.95	0.97	0.94
3	0.81	0.77	0.85
Avg	0.89	0.87	0.91

Table 11. Comparison of the prediction models

F measure	Precision	Recall	Accuracy	Model
0.52%	0.83%	0.49%	80.65%	**C&R**
0.61%	0.83%	0.55%	82.8%	**C5**
0.83%	0.93%	0.77%	90.86%	neural network
0.89%	0.91%	0.87%	93.55%	**SVM**

CONCLUSION

The present study is a study to recognize the effective factors in schizophrenia. In this study, we sought to discover the rules and information to be used by physicians in diagnosis using data mining on data related to schizophrenic patients. In this study, the most effective factors in occurring disease were investigated based on data mining techniques and the age of occurring illness was predicted. Four algorithms were used for this purpose and accuracy, recall, precision and F criteria were used to evaluate the efficiency of each algorithm.

According to the data, the results of investigating the factors affecting schizophrenia indicate that although lifestyle (addiction, deprivation) or hereditary and genetic factors are not ineffective in the schizophrenia, but it is not a strong reason for schizophrenia. More complex factors may probably have been involved in occurring schizophrenia. Of course, because many of our data didn't have quantity, it is not possible to express a definitive conclusion. Therefore, the necessity of research on a completer and more comprehensive dataset is essential for future research.

The result from investigating and comparing the algorithms used on the dataset of the schizophrenia disease shows that support vector machine algorithm with the polynomial kernel has obtained the highest accuracy. The accuracy for the educational data was 93.55% and for the test data was 68.57%. This finding shows the capability of this technique in predicting the age of occurring schizophrenia by applying effective factors in the occurring disease. Therefore, it can predict the age of occurring schizophrenia using the support vector machine algorithm. Predicting the time of occurring illness can always be proposed as an effective factor in the treatment and prevention of mental illnesses. Given that the ability of support vector machine algorithm was proved in predicting the age of occurring illness using factors influencing the illness, it can be identified the age range at risk of illness in the community and acted to treat that age group by constructing a proper pattern. Finally, it is recommended to use the algorithm used in this study on the data of schizophrenic patients from other provinces for comparison. Also, other data mining algorithms to be applied on the data of schizophrenic patients studied.

Acknowledgment

Hereby, the authors appreciate and thank Dr. Sadeghi and his colleagues at Kerman Golestan Salamat Center for their participation in the data collection.

REFERENCES

[1] Badner, J. A., & Gershon, E. S. (2002). Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia. Molecular psychiatry, 7(4), 405-411.

[2] Behrman, M., Linder, R., Assadi, A. H., Stacey, B. R., & Backonja, M. M. (2007). Classification of patients with pain based on neuropathic pain symptoms: comparison of an artificial neural network against an established scoring system. [Comparative Study .]Eur J Pain, 11(4), 370-376.

[3] Borjkhani, M., Tohidkhah, F., & Davandeh, H. (2008). Written Disorders in Schizophrenia and Improvement of Diagnosis Based on Extracted Patterns from Pen Motion Using SVM Method, 15th Iranian Medical Engineering Conference, Mashhad, Iranian Association of Medical Engineering, Islamic Azad University of Mashhad.

[4] Ghazanfari, M., Alizadeh, S., Teymourpour, B. (2011). Data mining and knowledge discovery. 2nd ed. Tehran: publication of university of science and technology.[Persian]

[5] Han, J., & Kamber, M. (2006). Data Mining: Conceptsand Techniques. 2nd ed.San Francisco: Morgan Kufman Publisher.

[6] Dose, M., Gruber, C., Grunz, A., Hook, C., Kempf, J., Scharfenberg, G., & Sick, B. (2007, April). Towards an automated analysis of neuroleptics' impact on human hand motor skills. In 2007 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (pp. 494-501). IEEE.

[7] Michalewicz, Z. (1996). Genetic Algorithms + Data Structures = Evolution Programs. 3rd Ed. Springer-Verlag.

[8] Michie, D., Spiegelhalter, D. J. & Taylor, C.C. (1994). Machine Learning, Neural and Statistical Classification. New York: Ellis Horwood.

[9] Adriaans, P., & Zantinge, D. (1996). Data Mining (New York: Addison Wesley).

[10] Pouyan, A. A., & Shahamat, H. (2013). Diagnosis of schizophrenia using FMRI images, 12th National Conference on Intelligent Systems, Bam, Iranian Intelligent Systems Association, Bam Higher Education Complex.

[11] Struyf, J., Dobrin, S., & Page, D. (2008). Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. BMC Genomics, 9(1), 531. http://dx.doi. org/10.1186/1471-2164-9-531