International Journal of Computer & Information Technologies (IJOCIT)
Volume 04, Issue 02
Jun 2016
ISSN = 2345-3877

Data mining Application for Exploring the Relation-ship between Addiction and Depression

Shiva Farokhbalaghi[1], Rozita Jamili Oskouei [*2]

[1]Department of Software Engineering, Zanjan Branch, Islamic Azad University, Zanjan
[2]Department of Computer Science and Information Technology, Mahdishahr Branch, Mahdishahr Iran[1,2]

Email: sh.balaghi@yahoo.com & rozita2010r@gmail.com

Abstract: In recent years, increasing the number of people with addiction and depression has become one of the major challenges in our country. This has a very important role in increasing the number of deaths, suicides, and murders caused by drug use in the society, in addition to the high economic burden due to the rehabilitation and treatment costs, which are imposed to the economic system of the community. Therefore, creating a method or methods for identifying those at risk of developing any of the listed diseases (addiction or depression) and considering preventive measures is one of the important requirements, which prevents the waste of medical expenses and financial or life loss damages in the community .

In this paper, the data related to the patients including the results of clinical and para-clinical tests, personal information (such as gender, age, education level, place of residence, economic status, etc.) were collected. Then, data mining techniques such as clustering, discovering forum relationships between various variables (properties) were applied on these data. Finally, a model will be presented to any of the diseases (addiction or depression) with a high degree of reliability and accuracy to help clinicians to identify individuals at high risk. In addition, the relationship between the two diseases can be identified by creating forum rules. This model can be used as a medical assistant or worker in remote or disadvantaged areas in the case of implementing in the form of a smart system.

Keywords: Addiction, Depression, Data mining, Clustering, Medical Assistant, Smart System.

1. Introduction

Addiction and depression has already become a serious problem for the country that annually killed a large number of people, especially young people. In addition, it imposes a heavy financial and economic burden (due to costs for the treatment or rehabilitation of patients) to the community. Hence, it is recommended that researchers from different fields seek solutions to help fixing the problem.

In the field of medical and health, prevention is always emphasized rather than treatment because the cost of treating patients is high. Thus, it will be very noticeable if a model to be created to identify people at risk before the disease (addiction/depression) that intelligently can be able to predict the possibility of affecting and providing optimal strategies for prevention (prophylaxis) of them.

Exploration of reality through data mining techniques can provide an opportunity for diagnosis diseases and providing an optimal solution for the treatment, necessary cares to avoid risk factors for various diseases, investigation in the field of public health and disease control etc.

There are unknown aspects to identify the main cause of people's willingness to drugs and depression. In addition, there is no detailed and proprietary study about the infection rates of depressed people to drugs based on gender, age and place of residence and other applicable factors, as well as about which of the two females or males are more vulnerable in terms of affecting the two diseases? Whether some factors such as dietary habits, family economic status, geographical location, etc. are involved in people's vulnerability? And whether any smart system (smart mode) can be designed to help diagnosing the disease at an early stage or identify individuals susceptible to getting the disease? Therefore, there is no reliable answer for them.

In this paper, we investigate the causes of addiction and the factors affecting it, as well as analyzing various factors such as age, gender, family background, social status of the family, financial condition, results of blood tests, etc. In the next step, we create models using data mining and classification techniques. Finally, we compare their efficiency and accuracy and then, we select and introduce the most stringent model. The purpose is presenting a model to help diagnosing risked individual (at the risk for depressive illness) that can provide necessary cares for these people and prevent them to

be affected. Finally, we will create relationships between factors affecting in the development of addiction or depression using techniques for discovering forum rules.

This paper is organized in six sections. The Second section includes related work about data mining applications in depression and addiction. The third section describes data collection and pre-processing steps. The fourth section includes data analysis processes. The fifth section shows evaluation results of our discovered patterns. The sixth section concludes of this paper and discusses about future work.

2. Related Work

In this section, we will evaluate the data mining applications in the diagnosis and treatment of depression. Then we will examine the role of data mining to predict a person's risk of addiction, discovering addicts and depressed people in the early stages of development and then we will choose the best treatment for depressed addicts.

Some researchers have used fuzzy algorithms and neural networks for early detection of depression disease [1]. Based on the results of their experiments, about 92% of all discovered forum rules are unreliable or unused. However, the obtained algorithm could be used with high reliability for detecting the possibility of depression and detecting the depression developing.

Other researchers have conducted other studies to predict the relationship between brain tumors and depression [2]. Several regression models were tested with 37% of the variance in the prediction of depression in people with brain cancer. Other researchers used ontology and Bayesian networks to create a framework to help diagnosing depression [3]. In addition, some studies have been done in order to use current data in social networks for the diagnosis of depression [4]. The results show that applying performed activities by users on social networks can be a way to identify people with depression. In this way, the people who are most active in these networks are less depressed. In addition, students' activities in social networks to explore their behavior pattern and using these patterns to discover their abnormal behaviors were observed and analyzed through data mining techniques [5]. Other studies have shown that investigating female behavior pattern in social networks can give us a model that can be used to predict postpartum depression [6].

Kevin Daimi et al. used data mining techniques such as classification to predict the risk of depression [7]. They selected about fifty factors (properties) for their analysis

that some of these factors are mentioned in the following table.

The symptoms and effects of depression have been studied by various researchers. Biological symptoms of depression were investigated by Matthew et al. [8]. They used data analysis using regression techniques and measuring the variance to find a group of biological signals to help predicting the severity of depression. This technique was not very successful to predict neurological disorders and their severity. Tung et al. emphasized using statistical inference method called negative emotion evaluation model to explore the tendency to depression through posts on the web [9]. For this purpose, data on posts and Chinese forms in Taiwan were collected. They classified each post based on the severity of depression, negative emotions, symptoms, and negative thoughts. Koh et al. examined the relationship between lung cramps diseases and depression. Their results showed that assessment test scores directly related to depression diagnosis and have high accuracy in predicting depression in patients with lung diseases [10]. In other words, there is correlation between eight terms in diagnostic test and depression. Kumar et al. studied the psychological, mental, and physical problems arising from Internet addiction among high school students living in urban areas located in West Turkey [11]. Data were analyzed by analysis of variance and t-test. The percentage of students' addiction in using the Internet in this study had an average of 44.51 ± 17.90. Their results of studying showed a direct relationship between internet addiction and getting a variety of illnesses and physical changes (such as poor sleep, loss of appetite) and disabilities and mental illnesses (such as fatigue, anger and confusion, palpitations, loss of relationships with friends and acquaintances, malaise and the emptiness of life) when they are not connected to the internet. Jang et al. examined the relationship between parental alcoholism and Internet addiction for both males and females. 266 males and 253s female were selected for this study in the range of 11 to 12 years old [12]. More than fifty percent of these people have parents with high college degrees, and 72.4% of these patients had moderate economic situation. About 3.5% of those parents were divorced from each other. There is a direct correlation for boys between the amount of intelligence with the degree of anxiety and confusion. However, there is no correlation between intelligence and ill-treatment with family members. The intelligence and aggression in both the male and female have a positive Correlation.

Yu Chen et al. researched the side effects of drugs in pregnant women. They conducted this research using forum rules to evaluate the relationship between the laws of the drug doses, duration of use, as well as their gestational age [13]. The evaluation results of the proposed tool showed that this tool due to its interactive nature allows the user to extract useful rules. America's National Institute of Addiction Research conducted a research to evaluate the relationship between hyperactivity in children and the possibility of turning to drug addiction [14]. The results showed that there is a significant difference between males and females in dealing with physical and emotional problems. When these diseases occur in women, women mostly try to go to a psychologist or specialist while men turn to use of various medicines and drug occasionally. Charanpreet Kaur et al. examined the use of data mining to diagnose and treat addiction [15].

3. Data Collection & Pre-Processing

This phase includes the following sections:

3.1. Collecting Patients' Data

The required data for this research have been collected from 2011 to 2015 in rehab center and drug treatment of addicts in Zanjan Province managed by Mr. Moradi located in Bisim St. The number of folders was 2500 that some of them were related to addicts who referred just one time in recent years and did not return for checkup, or their information was incomplete. Therefore, these patients were excluded from the list of our patients. In total, 1150 patients were selected for evaluating that 932 of them were selected using the sampling method, simple selection of samples without replacement as the training set. The remaining 350 records were used as a test set to assess the proposed method and model and measuring its accuracy.

Other data such as blood group, blood cholesterol, blood sugar, etc. were collected from patients' test tabs.

3.2. Screening the collected data

The following cases were performed for screening data:

- Removing all records that have missing or unrelated field.
- Removing the name of patients (for the sake of security and hiding personal information)
- Removing virtues that have not been answered by many patients.
- Applying the normalization rules on data

- Restoration and integration of data

4. Data Analysis

Several analyzes are performed on the obtained data. The purpose is searching the relationship between different characteristics and trying to understand the impact of all the risk factors in creating the disease. There are several ways to create a model. Here, we will use three of them (including KNN algorithms, neural network and decision tree algorithm). Finally, we will assess the efficiency and accuracy of each of them in diagnosing depression in addicts.

4.1. Evaluation Method

We used the calculation of specificity, precision, sensitivity, and accuracy (AC), to evaluate the performance of the created patterns and selecting the best pattern to predict the possibility of developing a depressive illness among addicts that their equation is as follows:

$$Ac = \left(\frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}\right)$$

$$\text{precision} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{false positives}}$$

Whereas:

TP: The number of samples that is correctly classified. (Number of healthy people whose lack of illness was not properly diagnosed).

FP: The number of healthy samples (individuals) who were incorrectly put in the patients' category.

FN: The number of patients who have been diagnosed wrongly healthy.

TN: The number of patients who have been correctly diagnosed their depression disease.

5. Experimental Results

All patients' information is stored in an Excel file with raw data to analyze the obtained data in the second step using Rapid miner.

Pre-processing operation was performed before entering data in Rapidminer (listed in Section 3). Then, the operation of symbol or samples selection was done to create the data set of training using Simple Random Sampling (without Replacement). To do this, randomly and without replacement method for selecting samples was used. Finally, the data related to 932 patients were selected as the training set. The data related to 350 patients were selected as the test set.

Both sets of data were stored separately in an Excel file and they are entered into Rapid Miner as needed. According to these figures,

the following conclusions can be deduced about the selected data:

- In the total of 1150 addicts, the information of 932 relatively complete addicts was selected as the training set.
- In total, 20 characteristics related to addicts were collected and evaluated that two characteristics are related to ID and Label. We conduct the following analyzes to discover the relationship between different characteristics.
- Evaluating the relationship between the types of drug with the type of drug of addicts' location: those who overdose drug live in the city of Zanjan. In other words, there was no overdose case around the city of Zanjan in the collected data by us. On the other hand, the highest rate of drug use was related to opium, heroin, and hashish lonely or in combination.
- Surveying the way of using the drug: the most way of using the drug among all studied addicts was fumigant way.
- Surveying the relationship between physical illness and the way of using the drug among all studied addicts: there is a relationship between different physical diseases and the way of using the drug. In addition, addicts who use drug in fumigant way are more prone to physical diseases.
- The relationship between depression and Location: the depression rate in addicts living in the city is more than addicts who are outside of the city.
- The relationship between occupation and addiction: People, who are self-employed, have the highest risk of addiction.
- The relationship between marital status and addiction: More addicts have positive marital status. It is interesting to note that about 63% of them have stated that their spouses are informed about their addiction, but they do not have any problem in their family relationships.
- The relationship between the level of education and addiction: most of the addicts are employees in public or private institutions with higher education degrees and they continue their addiction.
- Checking the statistics of depression among addicts: about 73 percent of addicts have a mental illness such as (psychosis, anxiety, loneliness, feel-

ing the desire to commit suicide, stress, etc.).

- Creating forum relationships between the different characteristics to help identifying risked individuals: Rule Model and FP-Tree methods were used to create forum relationships in Rapid miner. The results showed that the prediction accuracy of the relationships between different factors affecting depression in addicts is equal to 787 from 927, i.e. 85%.

6. Conclusion

In this paper, the data of 1150 addict patients were collected from a drug rehabilitation center. Then, we initially evaluated existing factors in those addictions. Factors such as education level, occupation, physical and mental diseases, etc. were investigated and their relationships were explored as possible. In the next step, we created a pattern to identify depressed people among addicts using three methods of data mining include decision tree C5.0, Neural Networks, and KNN. Then, the identification accuracy of depressed people was separately checked by any of these methods. Finally, it was observed that neural networks have a greater accuracy (93.2%) compared to two other methods to identify depressed or risked individuals.

References

[1] Subhagata Chattopadhyay, "A neuro-fuzzy approach for the diagnosis of depression", Applied Computing and Informatics Journal (2014).
[2] David K. Wellisch*, Thomas A. Kaleita, Donald Freeman, Timothy Cloughesyand Jeffrey Goldman, "Predicting Major Depression in Brain Tumor Patients", Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/pon.562
[3] Yue-Shan ChangǏǧWan-Chun HungǏǧTong-Ying Juang, (2013), "Depression Diagnosis based on Ontologies an Bayesian Networks", 2013 IEEE International Conference on Systems, Man, and Cybernetics.
[4] Munmun De Choudhury, Michael Gamon, Scott Counts,Eric Horvitz, "Predicting Depression via Social Media" , 2013, Association for the Advancement of Artificial Intelligence (www.aaai.org).
[5] Park, M.; Cha, C.; & Cha, M. (2012), "Depressive Moods of Users Captured in Twitter". In Proc. ACM SIGKDD Workshop on Healthcare Informatics (HIKDD).
[6] De Choudhury, M.; Counts, S.; & Horvitz, E. (2013), "Predicting Postpartum Changes in Behavior and Mood via Social Media". In Proc. CHI 2013, to appear.
[7] Kevin Daimi, Shadi Banitaan, (2014), "Using Data Mining to Predict Possible Future Depression Cases", International Journal of Public Health Science (IJPHS), Vol.3, No.4, pp. 231 - 240.
[8] Mathew, RJ. Largen, J., Claghorn, JL., "Biological Symptoms of Depression", *Psychosomatic Medicine*, Vo. 41, No.6, pp. 439-443.
[9] Tung, C., Lu, W., (2012), "Predict Depression Tendency of Web Posts using Negative Emotion Evaluation Model", ACM SIGKDD Workshop on Health Informatics (HIKDD 2012), Beijing, China.
[10] Koh, H., Tan, G., (2005), "Data Mining Applications in Healthcare," Journal of Healthcare Information Management, Vol. 19, No. 2, pp. 64-72.
[11] Kamer Gür, Seher Yurt, Serap Bulduk and Sinem Atagöz, (2014), "Internet addiction and physical and psychosocial behavior problems among rural secondary school students", Nursing and Health Sciences, 2014Wiley Publishing Asia Pty Ltd, pp. 1-8.
[12] Mi Heui Jang and Eun Sun Ji, (2012), "Gender differences in associations between parental problem drinking and early adolescents' Internet addiction", Journal for Specialists in Pediatric Nursing, Vol. 17, pp.288–300.

[13] Yu Chen, Lars Henning Pedersen, Wesley W. Chu1 Jorn Olsen, (2007), "Drug Exposure Side Effects from Mining Pregnancy Data", ACM SIGKDD Explorations Newsletter - Special issue on data mining for health informatics, Volume 9, Issue 1, pp. 22 – 29.
[14] National Institute of Drug Abuse, (2008), Comoridity: Addiction and Other Mental illness", Research Report Series, pp. 1-12.
[15] Charanpreet Kaur & Shweta Bhardwaj, (2014), "DRUG Discovery Using Data Mining", International Journal of Information and Computation Technology.Vol. 4, No. 4, pp. 335-342.

