International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

A Comparative Study on the Performance of Classifiers in Predicting Frequency of Drug Intake: A case study with Ketamine, Heroin, Crack and Meth

Shomona Gracia Jacob
SSN College of Engineering
Sahitya Sridhar
SSN College of Engineering
Nikhilesh Murugavel
SSN College of Engineering

ABSTRACT

Data Mining and pattern recognition has been used in several commercial and academical applications. A wide variety of techniques have been devised to tackle this variety in applications. Classification is a data mining technique that assigns items to target categories. The goal of classification is to predict the target category of the item accurately. Despite the long tradition of data mining research, no classification algorithm has been found to yield maximum accuracy in all applications. In this research a comparative study was performed on the performance of classification algorithms that are conventionally known to have high accuracy - Bayesian Networks, Random Forest, Regression, SMO, J48, Bayesian Networks, Perceptron classifiers and Logistic regression. The frequency of intake of the drugs crack, meth, heroin and ketamine were analyzed with personality measurements including NEO-F FI-R (neuroticism, extraversion, openness to experience, agreeableness and conscientiousness), BIS-11 (impulsivity), ImpSS(sensation seeking) and general demographic information. In addition, correlation based feature selection technique was applied on each algorithm and its results were recorded. The results of this work report that SMO performs best with over 75% accuracy when there are several features contributing to construction of the model and when the number of features is reduced using Feature Subset selection. The accuracy of J48, Logistic regression and multilayer perceptron models increase after feature selection with over 70% accuracy.

INTRODUCTION

Data mining is the process of extracting previously unknown useful relationships and patterns from data. This research work aims at studying the effects of data mining techniques in prediction of the frequency of drug intake in a large population. The drugs that have been studied are Ketamine, Heroin, Meth and Crack. An abusable psychoactive drug is a drug whose influence on mental functions pleasantly induces stupor and insensibility that some people choose to consume it for reasons other than to relieve illness. Drug consumption has an important and negative effect on the health of a human being, society and is considered a serious global problem. A number of reasons contribute to the initial drug usage including economical, psychological, social reasons. These factors are also associated with traits pertaining to an individual's personality. Psychologists state that the personality characteristics of the Big Five Factor Model (FFM) are most exhaustive traits for understanding the nuances of every individual. These personality measurements include neuroticism, extraversion, openness to experience, agreeableness and conscientiousness.

A lot of research studies have shown that these individualistic traits are related to drug consumption. Sutin, Evans and Zonderman [1] established that high Neuroticism and low Agreeableness increased risk of drug consumption and for every standard deviation decrease in Conscientiousness, there was a high in risk of drug consumption. Roncero et al [2] presented a correlation between high neuroticism and presence of psychotic symptoms following cocaine consumption. Dubey et al [3] established that a group of drug consumers had high Neuroticism and Extraversion dimensions, whereas others had significantly higher values of Openness and Conscientiousness dimensions of the Big-Five Factors. No significant difference was obtained on the Agreeableness trait of personality. Raketic et al [4] found that substance-dependent women had high values of Neuroticism and low values on Conscientiousness. Females dependent on a specific drug Opiate, scored highest on Neuroticism and Extraversion and lowest on Agreeableness and Conscientiousness. On the other

hand, females dependent on Alcohol had higher levels of Conscientiousness and lower levels of Neuroticism when compared to opiate-dependent women. Terracciano et al [5] established that cocaine/heroin drug consumers score very high on Neuroticism, and very low on Conscientiousness. Vollrath & Torgersen [6] noticed that the personality traits of Neuroticism, Extraversion, and Conscientiousness are highly related to dangerous health habits. A low score of Conscientiousness, and high score of Extraversion or high score of Neuroticism are associated with highly deleterious habits. Flory et al [7] observed that consumers of alcohol have lower Agreeableness and Conscientiousness and higher Extraversion. Additionally they observed that lower Agreeableness and Conscientiousness, and higher Openness to experience are related to the use of marijuana.

Turiano et al [8] observed that higher levels of neuroticism, extraversion, openness, and lower levels of conscientiousness and agreeableness indicated prolonged consumption of drugs. An increase in the score of neuroticism and openness indicated increased drug usage, while an increase in the score of conscientiousness and agreeableness indicated reduced drug consumption. It has also been observed that a high score of neuroticism has been associated positively with many other addictions such as Internet addiction, exercise addiction, compulsive buying, and study addiction, video game addiction and mobile phone addiction [9].

In the study presented in this paper, the performance of various classification algorithms in predicting the frequency of drug intake in an individual is investigated. WEKA (Waikato Environment for Knowledge Analysis) is used to conduct the experiments. Six different classifiers are selected and evaluated namely, Random Forest, SMO, J48, Bayesian Networks, Multi Layer Perceptron classifiers and Logistic regression.

DATA AND METHODOLOGY

THE DATA

The data set used in this study was collected by an anonymous online survey between March 2011 and March 2012. The survey tool used was Survey Gizmo. The survey yielded 2051 responses. Of the 2051 responses, 166 responses were considered invalid owing to a validity check to verify that the participants were attentive. Nine participants were further removed from the database based on previous other studies [10]. The dataset consists of twelve attributes for each participant in the study, the personality traits, NEO-F FI-R (neuroticism, extraversion, openness to experience, agreeableness and conscientiousness), BIS-11 (impulsivity), ImpSS(sensation seeking) and general demographic information such as ethnicity, gender, age, country and education. Additionally, the data set contains information on the consumption of 18 drugs including alcohol, benzodiazepines, cocaine, heroin, ketamine, methadone, nicotine, and Volatile Substance Abuse (VSA) of which ketamine, heroin, crack and meth were chosen for this study. The dataset was this divided into four sub datasets with for each drug respectively. The frequency of drug consumption was divided into multiple classes - never used, used over a decade back, used in the last decade, used in the last year, used in the last month, used in the last week and used in the last day which are annotated from CL0 to CL6. The default dataset is published online in UCI Machine Learning repository [11]. Table 1 contains the details of the data set.

Table 1. Table indicating details about the dataset for each drug

Dataset characteristics	Attribute characteristics	Associated tasks	Number of instances	Number of attributes	Missing values
Multivariate	Real	Classification	1885	12	N/A

METHODOLOGY

WEKA (Waikato Environment for Knowledge Analysis) is used to conduct the experiments.

Classification

Classification algorithms build models from a portion of the dataset called the training set (output data variable classes are known). These models are then used to predict the class labels for the test data set (output data variable unknown). The quality and performance of a classifier is evaluated by a number of methods such

as accuracy rate, f-measure, Kappa statistic, ROC area under curve and time spent for classification [12]. Since, this study focuses on practical analysis of the data, we use accuracy rate which is defined as the sum of number of true positives and the number of true negatives over the total number of instances in the sample [13]. In this study, the method used for assessing accuracy is 10 fold cross validation. The initial data is partitioned into 10 mutually exclusive subsets each of approximately equal size. Training and testing is performed 10 times. In each iteration, the i^{th} partition is used as the training set and the other partitions are used as test data sets [14]. The classification algorithms studied in this research are Random Forest, SMO, J48, Bayesian Networks, Multi Layer Perceptron and Logistic regression.

J48

It is an open source java implementation of C4.5 for Weka, a data mining tool developed by University of Wakaito. This algorithm is an optimized implementation of C4.5 and outputs a decision tree. A decision tree is a classifier that constructs a tree like structure, which can take multiple courses of action for a single node. It contains a root node, multiple intermediate nodes and a leaf node. Each node contains a decision and on taking decisions sequentially the leaf node is reached, containing the target class. A splitting criterion is used to identify the best node to split at each level of the tree [15].

Random Forest

Random forest classification was proposed by Breiman [16]. The random forest classification method consist of a collection of structured decision trees. Each tree is constructed using a different bootstrapped sample from the initial dataset [17]. In regular trees, a node is split using the best split among all given variables. In random forest, the best split if found using the best among a subset of predictors chosen for that node at random. In standard trees, each node is split using the best split among all variables. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node [18]. The random forest algorithm multiple of decision trees and combines them into a single model. In the case study we constructed a random forest of 10 trees; each constructed using four random features.

Logistic Regression

In this study, we use Multinomial logistic regression (multiple prediction classes) in which, the log odds of the output dependent variables are modeled as linear combination of the input independent variables [19]. The coefficients are computed using Ridge regression estimators which were introduced as an alternative to the ordinary least squares estimator in the presence of multi-collinearity [20].

Bayesian networks

A Bayesian network is a graphical probabilistic model that captures the relationships among the independent attributes. Since, the model takes care of dependencies among the variables it can handle situations where data entries are missing. A Bayesian approach to probability of an event is associated with degree of belief in the event. It is based on Bayes rule of conditional probability [14,21].

SMO

SMO (sequential minimal optimization) is an enhanced algorithm based on Support Vector Machines (SVM). A support vector machine's optimization problem is a Quadratic problem. SMO is an iterative method to solve the optimization problem by breaking the problem into a series of smaller sub problems which are solved analytically. This significantly improves the computational time of the algorithm [22].

Multi Layer Perceptron

The multi layer perceptron network is the most common neural network model. Neural networks contain information processing units that resemble neurons available in the human body except that they are artificial [23]. The MLP network has an input layer with a set of sensory nodes as input nodes, one or more hidden layers of nodes and an output later of nodes. The input nodes are the independent attributes and the output nodes are the splits between the target output classes [24].

Feature Selection

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

Performance of the classifiers can be improved by selecting a few features which contribute to increase its accuracy [25]. The feature selection method used in this study is Correlation-based Feature Subset Selection. In Correlation-based Feature Subset Selection [26], useful attribute subsets are those that contain features which predict the class but are not correlated with any other feature. CFS calculates a measure called the merit of a feature from pair-wise correlations and a formula. A heuristic search technique is then used to search the set of all subsets and the subset with the largest value of merit is chosen and reported. The selected features for prediction of the class label for each drug have been tabulated in Table 2.

Table 2. Table indicating features selected for each drug using CFS

Crack	Heroin
Age Gender Education Country Neuroticism score (NScore) Openness score (OScore) Impulsivity Sensation Seeking (SS)	Age Education Country Neuroticism score (NScore) Openness score (OScore) Agreeableness score (AScore) Conscientiousness (CScore) Impulsivity Sensation Seeking (SS)
Ketamine	**Meth**
Age Gender Country Openness score (OScore) Conscientiousness (CScore) Sensation Seeking (SS)	Country Ethnicity Agreeableness score (AScore) Conscientiousness (CScore)

RESULTS

The experimental results are portrayed in two sections. Primary results focused on evaluating the performance of classifiers when all the input features were included as a part of the dataset [27]. Following this, feature selection was applied to remove a few features and the performance was evaluated again. The results before and after feature selection of attributes are tabulated in Table 3 and Table 4 respectively.

Table 3. Accuracy of each classifier for each drug before feature selection

Drug	Algorithms implemented	Time Taken (Sec)	Correctly classified instances	Incorrectly classified instances	Accuracy (%)
Crack	Bayesian Network	0.07	1558	327	82.6525
	SMO	0.39	1627	258	86.3130
	J48	0.05	1601	284	84.9337
	Random Forest	0.09	1619	266	85.8886
	Logistic Regression	0.71	1621	264	85.9947
	Multilayer perceptron	4.89	1568	317	83.1830
Heroin	Bayesian Network	0.01	1530	355	81.1671
	SMO	0.16	1605	280	85.1459
	J48	0.04	1598	287	84.7745
	Random Forest	0.09	1597	288	84.7215
	Logistic Regression	0.39	1598	287	84.7745
	Multilayer perceptron	3.21	1588	297	84.2440
Ketamine	Bayesian Network	0.01	1531	354	81.2202

International Journal of Engineering Technology Science and Research
IJETSR
www.ijetsr.com
ISSN 2394 – 3386
Volume 4, Issue 9
September 2017

	SMO	0.11	1605	280	85.1459
	J48	0.05	1436	449	76.1804
	Random Forest	0.09	1478	407	78.4085
	Logistic Regression	0.3	1487	398	78.8859
	Multilayer perceptron	3.27	1440	445	76.3926
Meth	Bayesian Network	1.65	1354	531	71.8302
	SMO	0.23	1429	456	75.8090
	J48	0.35	1328	557	70.4509
	Random Forest	0.35	1386	499	73.5279
	Logistic Regression	1.17	1418	467	75.2255
	Multilayer perceptron	4.6	1379	506	73.1565

Table 4. Accuracy of each classifier for each drug after feature selection

Drug	Algorithms implemented	Time Taken (Sec)	Correctly classified instances	Incorrectly classified instances	Accuracy (%)
Crack	Bayesian Network	0.0	1558	327	82.6525
	SMO	0.2	1627	258	86.3130
	J48	0.04	1625	260	86.2069
	Random Forest	0.07	1603	282	85.0398
	Logistic Regression	0.42	1623	262	86.1008
	Multilayer perceptron	3.33	1611	274	85.4642
Heroin	Bayesian Network	0.01	1531	354	81.2202
	SMO	0.11	1605	280	85.1459
	J48	0.05	1596	289	84.6684
	Random Forest	0.09	1478	407	78.4085
	Logistic Regression	0.3	1600	285	84.8806
	Multilayer perceptron	2.81	1588	297	84.2440
Ketamine	Bayesian Network	0.00	1466	419	77.7719
	SMO	0.11	1490	395	79.0451
	J48	0.03	1490	395	79.0451
	Random Forest	0.44	1488	397	73.6870
	Logistic Regression	0.19	1490	395	79.0451
	Multilayer perceptron	3.7	1490	395	79.0451
Meth	Bayesian Network	0.04	1421	464	75.3846
	SMO	3.23	1429	456	75.8090
	J48	0.02	1423	462	75.4907
	Random Forest	0.08	1289	596	69.1373
	Logistic Regression	1.17	1418	467	75.2255
	Multilayer perceptron	2.42	1427	458	75.7029

CONCLUSION

This paper has evaluated the performance of classifiers in predicting the frequency of drug intake. From the tables above, we observe that maximum accuracy for all drugs before and after feature selection is obtained on application of the SMO classification algorithm. The accuracy of SMO is also consistent, and hence we conclude that the number of features used for model construction is irrelevant to its performance.

In future, we intend to apply more feature selection techniques and test the models on different datasets.

REFERENCES

1. Sutin, A. R., Evans, M. K., & Zonderman, A. B. (2013). Personality traits and illicit substances: The moderating role of poverty. Drug and Alcohol Dependence, 131(3), 247–251. https://doi.org/10.1016/j.drugalcdep.2012.10.020
2. Roncero, C., Daigre, C., Barral, C., Ros-Cucurull, E., Grau-López, L., Rodríguez-Cintas, L., … Valero, S. (2014). Neuroticism associated with cocaine-induced psychosis in cocaine-dependent patients: A cross-sectional observational study. PLoS ONE, 9(9). https://doi.org/10.1371/journal.pone.0106111
3. Dubey M.; Gupta, S.; Kumar, B., C. . A. (2010). Five Factor Correlates: A Comparison of Substance Abusers and Non-Substance Abusers. Journal of the Indian Academy of Applied Psychology, 36(1), 107–114.
4. Raketic D.; Barisic, JV.; Svetozarevic, SM.; Gazibara, T.; Tepavcevic, DK.; Milovanovic, SD. (2016). FIVE-FACTOR MODEL PERSONALITY PROFILES: THE DIFFERENCES BETWEEN ALCOHOL AND OPIATE ADDICTION AMONG FEMALES. PSYCHIATRIA DANUBINA,.
5. Terracciano, A., Löckenhoff, C. E., Crum, R. M., Bienvenu, O. J., & Costa, P. T. (2008). Five-Factor Model personality profiles of drug users. BMC Psychiatry, 8, 22. https://doi.org/10.1186/1471-244X-8-22
6. Vollrath, M., & Torgersen, S. (2002). Who takes health risks? A probe into eight personality types. Personality and Individual Differences, 32(7), 1185–1197. https://doi.org/10.1016/S0191-8869(01)00080-0
7. Flory, K., Lynam, D., Milich, R., Leukefeld, C., & Clayton, R. (2002). The relations among personality, symptoms of alcohol and marijuana abuse, and symptoms of comorbid psychopathology: results from a community sample. Experimental and Clinical Psychopharmacology, 10(4), 425–434. https://doi.org/10.1037/1064-1297.10.4.425
8. Turiano, N. A., Whiteman, S. D., Hampson, S. E., Roberts, B. W., & Mroczek, D. K. (2012). Personality and substance use in midlife: Conscientiousness as a moderator and the effects of trait change. Journal of Research in Personality, 46(3), 295–305. https://doi.org/10.1016/j.jrp.2012.02.009
9. Andreassen, C. S., Griffiths, M. D., Gjertsen, S. R., Krossbakken, E., Kvam, S., & Pallesen, S. (2013). The relationships between behavioral addictions and the five-factor model of personality. Journal of Behavioral Addictions, 2(2), 90–99. https://doi.org/10.1556/JBA.2.2013.003
10. Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., & Gorban, A. N. (2015). The Five Factor Model of personality and evaluation of drug consumption risk. Retrieved from https://arxiv.org/ftp/arxiv/papers/1506/1506.06297.pdf
11. UCI Machine Learning repository, University of California, Irvine, Link : https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29
12. Witten, I., & Frank, E. F. (2005). Data Mining: Pratical Machine Learning Tools and Techiniques - 2nd edition. Machine Learning. https://doi.org/10.1002/1521-3773(20010316)40:6<9823::AID-ANIE9823>3.3.CO;2-C
13. Amancio, D. R., Comin, C. H., Casanova, D., Travieso, G., Bruno, O. M., Rodrigues, F. A., & Da Fontoura Costa, L. (2014). A systematic comparison of supervised classifiers. PLoS ONE, 9(4). https://doi.org/10.1371/journal.pone.0094137
14. Han, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann (Vol. 12). https://doi.org/10.1007/978-3-642-19721-5
15. Patil, T. R., & Sherekar, S. S. (2013). Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification. International Journal Of Computer Science And Applications, ISSN: 0974-1011, 6(2), 256–261.
16. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
17. Hastie, T., Tibshirani, R., & Friedman, J. (2011). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics) (9780387848570): Trevor Hastie, Robert Tibshirani, Jerome Friedman: Books. In The elements of statistical learning: dta mining, inference, and prediction (pp. 501–520).
18. Liaw, a, & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(December), 18–22. https://doi.org/10.1177/154405910408300516
19. Starkweather, J., & Moske, A. K. (2011). Multinomial logistic regression. Multinomial Logistic Regression, 51(6), 404–410. https://doi.org/10.1097/00006199-200211000-00009
20. Al-Hassan, Y. M. (2010). Performance of a new ridge regression estimator. Journal of the Association of Arab Universities for Basic and Applied Sciences, 9(1), 23–26. https://doi.org/10.1016/j.jaubas.2010.12.006
21. Heckerman, D. (1996). A Tutorial on Learning With Bayesian Networks. Innovations in Bayesian Networks, 1995(November), 33–82. https://doi.org/10.1007/978-3-540-85066-3
22. Jacob, S. G. (2016). Prediction of Credit-Card Defaulters: A Comparative Study on Performance of Classifiers. International Journal of Computer Applications, 145(7), 975–8887.
23. Haykin, S. (1994). Neural networks-A comprehensive foundation. New York: IEEE Press. Herrmann, M., Bauer, H.-U., & Der, R. https://doi.org/10.1017/S0269888998214044
24. Tsai, C. F., Hsu, Y. F., & Yen, D. C. (2014). A comparative study of classifier ensembles for bankruptcy prediction. Applied Soft Computing Journal, 24, 977–984. https://doi.org/10.1016/j.asoc.2014.08.047
25. Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15), 2429–2437. https://doi.org/10.1093/bioinformatics/bth267
26. Geetha Ramani, R., & Jacob, S. G. (2013). Prediction of P53 Mutants (Multiple Sites) Transcriptional Activity Based on Structural (2D&3D) Properties. PLoS ONE, 8(2). https://doi.org/10.1371/journal.pone.0055401
27. Lim, T., Loh, W., & Shih, Y. (2000). A comparison of prediction accuracy, complexity, and training time of thirty three old and new classification algorithms. Machine Learning, 40(3), 203–229. https://doi.org/10.1023/A:1007608224229