2017 CMBEC40 Conference
Winnipeg MB
May 23–26, 2017

CLASSIFICATION OF UNDETERMINED DEATHS BY POISONING: COMPARISON OF HOMOGENEOUS AND HETEROGENEOUS DATABASES

Alana Esty[1], Monique Frize[1],[2], Corresponding Author, Steven R. McFaull[3], Robin Skinner[3], Jenini Subaskaran[3] and Melinda Tiv[3]
[1]Systems and Computer Engineering, Carleton University
[2]School of Electrical Engineering and Computer Science, University of Ottawa,
[3]Surveillance and Epidemiology Division, Public Health Agency of Canada, Ottawa, Canada

ABSTRACT

Identifying factors that differentiate suicides from unintentional deaths by poisoning is essential for accurate monitoring of suicide in order to help develop prevention measures. The results of our research demonstrated that the use of machine learning techniques such as: Artificial Neural Networks, Decision Trees, and Case-Based Reasoning have enabled us to classify the majority of undetermined cases found in the two databases analyzed. The data originated from the Canadian Coroner and Medical Examiner Database (CCMED): the first dataset included deaths in the Province of Ontario; the second dataset included cases from several other provinces excluding Ontario.

INTRODUCTION

Some articles mention that the misclassification rates of suicides are due to a variety of reasons, from religious motivations to differences in coroners' diagnosis [1], [2]. A first study used data from Ontario, extracted from the Canadian Coroner and Medical Examiner Database (CCMED). This database combines national suicide rates and uses error detection software, in order to ensure the data entering the database is the highest standard of accuracy. This second study used the CCMED database containing cases mainly from Alberta, with some cases from other provinces, excluding Ontario. Ontario is the only province in Canada which follows the strict Beckon test [3]. Both of these studies applied machine learning techniques: case based reasoning tools (CBR), decision trees (DTs) and artificial neural networks (ANNs), to classify undetermined cases. It was expected that this second study which represents a heterogeneous dataset and excludes data from Ontario, would show a lower misclassification rate of suicides due to the absence of this stringent test. Results of the first study were reported in [3].

DATA

There were twenty-nine variables considered for building the classifiers in the first study and also for the second study: a few of these variables include case identification number (a unique identifier), age of individual, sex of individual, alcohol abuse problem, drug abuse problem, prior suicidal behavior (declared intent) and presence of a suicide note (or equivalent such as video, e-mails, texts, letters).

The case identification number was labeled as the primary key in both the suicidal and unintentional datasets and was not used in training or testing the classifier. The ANN classifier used for this work performed better with a 1 or -1 as the output variable [4]; the manner of death was labeled as a 1 for suicide and -1 for unintentional cases.

DATA ANALYSIS

Three main tools were used in this research to classify the undetermined cases into either a suicide or an unintentional case: case based reasoning tools (CBR), decision trees (DTs) and artificial neural networks (ANNs).

Case-Based Reasoning

A case based reasoning (CBR) system was used to replace missing data from the suicidal and unintentional cases before training the ANN. The CBR followed the k-nearest neighbour algorithm to fill these missing cases in order to improve the performance of the classifier. The algorithm defined k as 10; therefore 10 matching cases closest to the missing value were retrieved from the database and the missing value replaced by the mean of those 10 cases [5].

Decision Trees

The software used to develop the DT (Decision Trees) is the C5 (See 5) RuleQuest Research [6]. One of the advantages of using the DT algorithm is that missing cases and an abundance of zeros does not affect the performance of the classifier, unlike the ANN, so there was no need to use CBR for this specific classifier.

Artificial Neural Networks

The ANN was developed using the Fast Artificial Neural Network (FANN) library. A feed-forward, back-propagation ANN was implemented for this work [7]. The FANN library allows users to modify parameters such as the number of hidden layers and the number of neurons and connections of the desired ANN [4]. For the second study, we used 20 variables from the original 29 variables provided in the dataset. Variables with a high percentage of missing values, such as employment problems, smoker, relationship problems (intimate), family relationship problems, financial problems, live alone and marital problems were excluded. Additionally, the case identification and manner of death were excluded for training purposes.

METHODOLOGY

Accuracy Measures

A main goal in this work was to identify the cases of suicides among the undetermined cases. Therefore, a Type II error or a false negative rate (classifier identifies an unintentional case when it is truly a suicide) was thought to be more critical than a Type I error or false positive rate (classifier identifies a suicide when it is an unintentional case). The performance matrix consisted of the true positive rate (classifier correctly identifies a suicide), the true negative rate (classifier correctly identifies an unintentional case), false negative rate and the false positive rate.

Experimental Sets (Second Study)

For the DT classifier, the data was organized into three separate sets. In Set 1 (See Table 1), all 27 variables were used (excluding case identification and manner of death for training purposes). Set 2 (See Table 1) contained only four key variables to train the DT: suicide note, suicide attempt, mental health problems and drug abuse problems. Set 3 (See Table 1) was composed of all of the variables except that seven obvious variables were removed: suicide note, suicide behavior, suicide attempt, crisis, pain, drug abuse problem and mental health problems; the purpose was to assess the classification performance without obvious variables.

The purpose of the distinction of these sets was to see which would be the optimal set to train the undetermined dataset and which would provide the best classification of the data. Set 2 (See Table 1), as mentioned previously, contained only four variables which had the highest weights outputted by the decision tree when classifying the undetermined data as either a 1 or -1. Finally, the last set was to test how well the classifier performed without obvious variables in order to see how this approach affected the classification results.

Similarly, the dataset for the ANN Classifier was divided into three sets. Firstly, all variables were used for Set 1 (See Table 2) except the variables with a high percentage of missing values (greater than 50%) mentioned in an earlier section. The ANN Classifier does not perform well with missing values and replacing missing values only works well when the percentage of values missing is less than 50%. For Set 2 (See Table 1 and 2), the same four key variables that were used to train the DT were used for the ANN: suicide note, suicide attempt, mental health problems and drug abuse problems. For Set 3 (See Table 1 and 2), seven obvious variables were removed: suicide note, suicide behavior, suicide attempt, crisis, pain, drug abuse problem and mental health problems as well as the variables that were excluded due to a high percentage of missing values.

The weights of each of the variables submitted to the ANN were calculated with a modified Garson algorithm [8], [9]. These weights were continuously updated throughout the training period in order to minimize the error between the actual output and the target output data of a suicidal or unintentional classification. Through an automated manipulation of the weights during the training period, the total sum of the errors reduced until it reached around 0.001 (default error of the FANN library [7]). Then the networks were applied to unseen data, using these optimal weights obtained during training. Subsequently, the verification thresholds for the specificity and sensitivity were 90% and 85% respectively. The network weights were saved in a network file if the networks surpassed these thresholds.

5-by-2 Cross Validation

In order to assess the accuracy and validity of the ANN and DT classifier, a 5-by-2 cross validation test was done. This validation test was chosen because it has a high Type II error and an acceptable Type 1 error or false positive rate [10].

Thresholds

When the classifier outputs a value, it is in the range of -1 to 1; in the first study, [3], thresholds of 0.5 and -0.5 were chosen to label the values outputted by the ANN. These same thresholds were applied to the present study. For example, if the value outputted by the ANN was greater than 0.5, a classification of "suicide" was assigned and a classification of "possibly suicide" was assigned if the output was between 0 and 0.5. Whereas, if the value was less than -0.5 a classification of "unintentional" was assigned and between 0 and -0.5 was described as "possibly unintentional". These thresholds were implemented in order to get the most accurate classification of a suicide or of an unintentional case.

RESULTS

Table 1: Second Study Performance Results DT Classifier

Performance Indicators with Decision Tree Experiments			
Description	***Set 1***	***Set 2***	***Set 3***
Number of Variables	27	4	20
Sensitivity	85± 2%	70± 2%	75± 2%
Specificity	90± 1%	90± 2%	86± 4%
Positive Predictive Value	88± 1%	86± 2%	83± 3%
Negative Predictive Value	88± 1%	81± 1%	81± 4%
Accuracy	96 ± 1%	96 ± 1%	96 ± 2%

Set 1: All 27 variables were used (excluding case ID number and manner of death).
Set 2: Four key variables were used to train the DT.
Set 3: Seven obvious variables were removed

Table 2: Second study Performance Results ANN Classifier

Performance Indicators with Artificial Neural Network Experiments			
Description	***Set 1***	***Set 2***	***Set 3***
Number of Variables	20	4	13
Sensitivity	88.3 ± 1%	69.9 ± 2%	78.1 ± 2%
Specificity	89.1 ± 5%	92.3 ± 1%	70.5 ± 1%
Positive Predictive Value	87.5 ± 5%	88.4 ± 1%	69.5 ± 1%
Negative Predictive Value	90.0 ± 1%	78.5 ± 1%	79.2 ± 0%
Accuracy	88.7 ± 3%	82.1 ± 1%	74.0 ± 1%
AUC	95 ± 2%	87.4 ± 1%	82.0 ± 1%

Set 1: All variables were used for Set 1 except the variables with a high percentage of missing values (greater than 50%).
Set 2: Four key variables were used to train the DT
Set 3: Seven obvious variables were removed in addition to the variables with a high percentage of missing values

Table 3: Parameters of ANN

Artificial Neural Network Options and Parameters Experiments	
Description	***Mode***
Hidden Layers	1
Connection rate	0.3
Number of hidden nodes	11

Artificial Neural Network Options and Parameters Experiments	
Description	***Mode***
Weights	2
Training Algorithm	1
Learning rate	0.9
Training Error Function	1
Incremental training momentum	1.2
Quickprop training decay factor	-0.0001
Quickprop training maximum growth factor	1.75
Rprop training initial step	0.2
Rprop training initial step increase factor	1.2
Rprop training initial step decrease factor	0.5

Discussion on the Classification Results

For the second study, Set 1 was also chosen as the best set to train the ANN for its optimal sensitivity and specificity values and was used to classify the undetermined cases using these parameters. All 10 networks were trained and the predictions were averaged to obtain the final classification results. The final classification (with the same thresholds used in the first study of 0.5 and -0.5) resulted in the classification of 1856 cases: 1290 unintentional cases, 112 suicidal cases, 269 possibly unintentional cases and 185 possibly suicidal cases. The misclassification rate of 6% for the second study was computed by dividing the number of suicidal cases determined from the ANN classifier by the total number of undetermined cases.

CONCLUSION

In the first study [3], the misclassification rate of suicides was 37% using a database that contained only cases in the Province of Ontario. The reason for a decrease in the misclassification rate in the second study may be due to the absence of the strict Beckon test, which seems to result in a higher misclassification rate of suicides in Ontario compared to Alberta and cases from other provinces. These two studies have demonstrated that it is possible to use machine learning techniques to classify undetermined cases of deaths by poisoning. Since there is no clear gold standard for this field as there are several undetermined deaths classified by coroners throughout Canada, one may rely on machine learning in the future to classify undetermined cases. Discovering the true rate of suicides and of unintentional deaths can help to develop strategies for prevention of these occurrences

ACKNOWLEDGEMENTS

The authors wish to thank Health Canada for a grant to perform this research and Hasmik Martirosyan for her help in performing the first study with the Ontario database.

REFERENCES

[1] I. M. Tollefsen, Et Al. ”Accidents and undetermined deaths: re-evaluation of nationwide samples from the Scandinavian countries.” *BMC public healt*h vol. 16 pp.1-12, 2016.

[2] S. S. Bakst, T. Braun, I. Zucker, Z. Amitai and T. Shohat, “The accuracy of suicide statistics: are true suicide deaths misclassified?.” *Social Psychiatry and Psychiatric Epidemiology* vol. 51 pp. 115-123, 2016.

[3] M. Frize, Et Al.”Classification of Undetermined Deaths by Poisoning: Suicidal or Unintentional.” XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016. Springer International Publishing, 2016.

[4] S. Nissen, ”Neural networks made simple.” *Software 2.0* vol. 2 pp. 14-19, 2005.

[5] C. Ennett, M. Frize, and C. R. Walker “Imputation of missing values by integrating neural networks and case-based reasoning”. In Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE, pp 4337-4341 2008.

[6] RuleQuest Research webpage. Available at: https//www.rulequest.com/. Accessed June 26th 2016.

[7] Fast Artificial Neural Network library documentation (FANN). Avail-able at: http://leenissen.dk/fann/wp. Accessed July 7th, 2016.

[8] G.D. Garson, “Interpreting neural network connection weights”. *Artificial Intelligence Expert* vol.6 pp. 47-51, 1991

[9] A.T.C Goh “Back-propagation neural networks for modeling complex systems”. *Artificial Intelligence in Engineering* vol.9 pp. 143-151 1995.

[10] T.Dietterich,”Approximate statistical tests for comparing supervised classification learning algorithms.” *Neural computation* vol. 10 pp.1895-1923 1998.