Iliou Machine Learning Data Preprocessing Method for Suicide Prediction from Family History

Theodoros Iliou, Georgia Konstantopoulou, Christina Lymperopoulou, Konstantinos Anastasopoulos, George Anastassopoulos, Dimitrios Margounakis, Dimitrios Lymberopoulos

Iliou Machine Learning Data Preprocessing Method for Suicide Prediction from Family History

Theodoros Iliou[1], Georgia Konstantopoulou[2], Christina Lymperopoulou[1], Konstantinos Anastasopoulos[3], George Anastassopoulos[1,4], Dimitrios Margounakis[4] and Dimitrios Lymberopoulos[3]

[1] Medical Informatics Laboratory, Medical School, Democritus University of Thrace, Alexandroupolis, 68100, Greece, tiliou@med.duth.gr, chlympero@gmail.com, anasta@med.duth.gr

[2] Special Office for Health Consulting Services, University of Patras, Greece, gkonstantop@upatras.gr

[3] Department of Electrical Engineer, University of Patras, Greece, Wire Communications Lab, anastasofpv@gmail.com, dlympero@upatras.gr

[4] Hellenic Open University, Patras, Greece, Margounakis.dimitrios@ac.eap.gr

ABSTRACT

As real world data tends to be incomplete, noisy and inconsistent, data preprocessing is an important issue for data mining. Data preparation includes data cleaning, data integration, data transformation and data reduction. In this paper, Iliou preprocessing method is compared with Principal Component Analysis in suicide prediction according to family history. The dataset consists of 360 students, aged 18 to 24, who were experiencing family history problems. The performance of Iliou and Principal Component Analysis data preprocessing methods was evaluated using the 10-fold cross validation method assessing ten classification algorithms, IB1, J48, Random Forest, MLP, SMO, JRip, RBF, Naïve Bayes, AdaBoostM1 and HMM, respectively. Experimental results illustrate that Iliou data preprocessing algorithm outperforms Principal Component Analysis data preprocessing method, achieving 100% against 71.34% classification performance, respectively. According to the classification results, Iliou preprocessing method is the most suitable for suicide prediction.

Keywords

Data preprocessing; machine learning; data mining; classification algorithms; suicide, family history.

1. INTRODUCTION

Suicide is considered as a major public health problem with increasing trends in many developed and developing countries. Suicide is among the ten leading causes of death for all age ranges [1]. Factors that have been associated mainly with increased risk of suicide are socio-demographic factors, psychiatric morbidity, physical health problems and biological background. Suicidal behavior is defined as the behavior in which a person wants to harm himself in order to put an end to his life. It can be distinguished in suicide attempts, in which the person does not achieve his ultimate goal and suicides in which death is achieved [2]. Suicidal ideation concerns thoughts connected with desire, intention and method of suicide. According to O'Carroll et al. (1996), suicidal ideation refers to self-reported thoughts of committing suicide-related behavior. It includes thoughts of suicidal behaviors. Whereas, attempting suicide is the self-induced act of auto-harming committed with the intention of the person to end his life [3]. Other mental disorders associated with suicide are schizophrenia, substance abuse, alcoholism, personality disorders (especially antisocial and borderline personality disorders), panic disorder, etc. Factors associated with increased rates of suicide include drug abuse, psychological states, cultural, family and social situations, and genetic predisposition [4].

Mental illness and drug use often co-exist [5]. Other risk factors include: previous suicide attempts [6], the immediate access to an instrument which can help to commit suicide, a family history of suicide, or the presence of traumatic brain injury [7]. For example, it has been found that suicide rates are higher in houses where there are weapons, than in those that do not have [8]. Socio-economic factors such as unemployment, poverty, homelessness and discrimination may trigger suicidal thoughts [9]. Approximately 15-40% of victims leave a suicide note [10].

Genetic susceptibility appears to represent 38% to 55% of suicidal behaviors [11]. The war veterans are at greater risk of committing suicide and this is partly due to higher rates of mental illness and physical health problems associated with the war [12]. Family history of suicidal behavior greatly increases the risk for suicide. Also, the presence of parents' psychopathology and particularly depression and substance use were associated with suicide.

Moreover, according to some studies, the suicide victims are more likely to come from families of divorced parents.
This paper is structured as follows: Section 2 provides a review of the related literature regarding machine learning techniques used for classification or prediction. Section 3 describes the basic principles of the data preprocessing procedure and section 4 describes ILIOU preprocessing method. Finally, Section 5 presents the experimental results of this study, while Section 6 concludes this paper and describes future work.

2. RELATED WORK

In other paper we establish a novel data preprocessing method for improving the prognosis' possibilities of a patient suffering from depression to be leaded, to the suicide [13]. For this reason, the effectiveness of many machine learning classification algorithms is measured, with and without the use of our suggested preprocessing method. The experimental results reveal that our novel proposed data preprocessing method markedly improved the overall performance on initial dataset comparing with Principal Component Analysis (PCA) and Evolutionary search feature selection methods. So this preprocessing method can be used for significantly boost classification algorithms performance in similar datasets and can be used for suicide tendency prediction [13]. Traditional approaches to the prediction of suicide attempts have limited the accuracy and scale of risk detection for these dangerous behaviors [14]. They sought to overcome these limitations by applying machine learning to electronic health records within a large medical database. Participants were 5,167 adult patients with a claim code for self-injury (i.e., ICD-9, E95x), expert review of records determined that 3,250 patients made a suicide attempt (i.e., cases), and 1,917 patients engaged in self-injury that was nonsuicidal, accidental, or unverifiable (i.e., controls). We developed machine learning algorithms that accurately predicted future suicide attempts (AUC = 0.84, precision = 0.79, recall = 0.95, Brier score = 0.14). Moreover, accuracy improved from 720 days to 7 days before the suicide attempt, and predictor importance shifted across time. These findings represent a step toward accurate and scalable risk detection and provide insight into how suicide attempt risk shifts over time [15].

In this paper we had responses from 360 students, aged 18 to 24, and were collected during the clinical interview, which included the following questions:
"Is there family history of psychopathology?", "Is there death incident in your family?" "There is abuse incident in your family?", "Is there a chronically ill or disabled person in your family?", "Is there a person in the family who attempted suicide or committed suicide?", "Is there in your family alcoholic or drug user?", “Do you have you thoughts of committing suicide?".

3. DATA PREPROCESSING

The set of techniques used prior to the application of a data mining method is named as data preprocessing for data mining [16] and it is known to be one of the most meaningful issues within the famous Knowledge Discovery from Data process [17, 18]. Since real world data are generally imperfect, incomplete (lacking attribute values, lacking certain attributes of interest, or containing only aggregate data), noisy (containing errors or outliers), inconsistent (containing discrepancies in codes or names) and contain redundancies, is not directly applicable for a starting a data mining process. We must also mention the fast growing of data generation rates and their size in business, industrial, academic and science applications. The bigger amounts of data collected require more sophisticated mechanisms to analyze it. Data preprocessing is able to adapt the data to the requirements posed by each data mining algorithm, enabling to process data that would be unfeasible otherwise.
Data preprocessing includes (figure 1):

- Data cleaning: fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies.
- Data integration: using multiple databases or files.
- Data transformation: normalization and aggregation.
- Data reduction: reducing the features of the initial dataset but producing the same or similar analytical results.

4. ILIOU PREPROCESSING METHOD

In figure 1, a brief description of ILIOU method is presented.

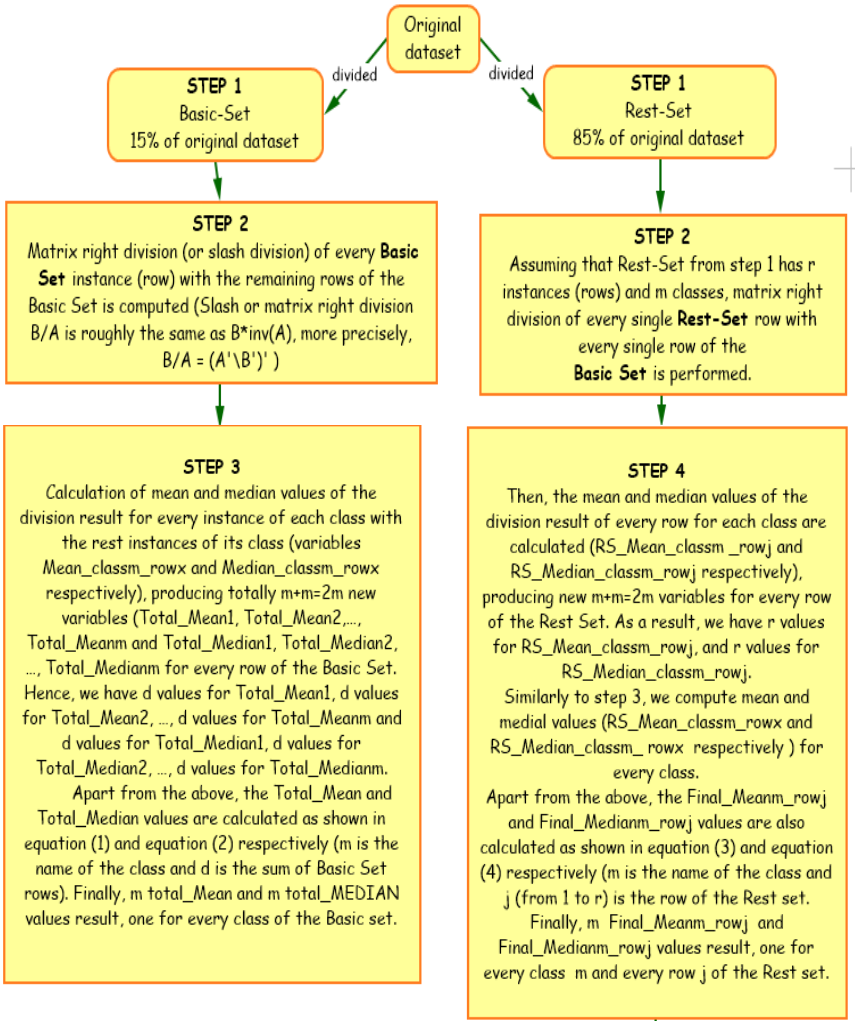

Figure 1. ILIOU preprocessing method

(Equation 1)

$$\text{Total\_Mean}_{m} = \frac{(\text{Mean\_class}_{m}\text{\_row}_{1} + \text{Mean\_class}_{m}\text{\_row}_{2} + \ldots + \text{Mean\_class}_{m}\text{\_row}_{d})}{d}$$

(Equation 2)

$$\text{Total\_Median}_{m} = \frac{(\text{Median\_class}_{m}\text{\_row}_{1} + \text{Median\_class}_{m}\text{\_row}_{2} + \ldots + \text{Median\_class}_{m}\text{\_row}_{d})}{d}$$

(Equation 3)

$$\text{Final\_Mean}_{m}\text{\_row}_{j} = \text{total\_Mean}_{m\ (step3)} - \text{RS\_Mean\_class}_{m}\text{\_row}_{j}$$

(Equation 4)

$$\text{Final\_Median}_{m}\text{\_row}_{j} = \text{total\_Median}_{m\ (step3)} - \text{RS\_Median\_class}_{m}\text{\_row}_{j}$$

5. EXPERIMENTAL RESULTS

For testing the generalization of our reprocessing method we used the repeated 10-fold cross validation technique [19] during the classification procedure. For the classification results we used some well-known classification algorithms in WEKA 3.8 data mining software [20] by their default WEKA parameters. The performance of the classifiers was measured by Precision, Recall, Kappa statistics, Weighted Avg ROC area, Weighted Avg Precision Recall Curve (PRC) area, Matthews correlation coefficient (MCC) and F-measure metrics. The comparison of the aforementioned statistics and metrics of the initial data set, of transformed data set with PCA and with Iliou preprocessing method are illustrated in Table 1-3.

As shown in Tables 1, 2 and 3 and in figure 2, Iliou data preprocessing method has remarkably improved the classification performance (Correctly Classified Instances %) for each classifier compared to PCA and initial dataset classification performance. Iliou method achieved 100% classification performance with Multilayer Perceptron (MLP) classifier, whereas PCA has almost the same classification results (71.34%) as initial data classification (72.47%). PCA achieved maximum classification performance using Naïve Bayes, while the best classification results with initial dataset achieved by J48 and SMO classification algorithms. Hence, Iliou data preprocessing method noticeably outperform PCA for each classification scheme. As we notice the Precision, Recall and ROC values we also observe that ILIOU method has much higher values in these parameters that's why ILIOU method has the best classification performance. According to k and MCC value we also observe that ILIOU method's values are close to 1, which means that the classification results are not accidental.

Table 1. Original data results

Classifiers	Correctly Classified Instances (%)	Pre	Rec	k	ROC	PRC Area	F-Measure	TP Rate	FP Rate	MCC	Root mean squared error
IB1	68.82	0.68	0.68	0.2	0.64	0.64	0.663	0.68	0.43	0.29	0.46
J48	**72.47**	0.73	0.72	0.3	0.63	0.64	0.698	0.72	0.4	0.39	0.44
Ran For	67.41	0.66	0.67	0.2	0.65	0.65	0.644	0.67	0.45	0.26	0.46
MLP	68.53	0.67	0.68	0.2	0.65	0.65	0.658	0.68	0.44	0.29	0.46
SMO	**72.47**	0.73	0.72	0.3	0.66	0.64	0.698	0.72	0.40	0.39	0.52
Jrip	67.41	0.63	0.67	0.2	0.62	0.62	0.656	0.67	0.43	0.27	0.46
RBF	71.34	0.71	0.71	0.3	0.67	0.68	0.694	0.71	0.39	0.36	0.44
Naïve Bayes	71.91	0.71	0.71	0.3	0.69	0.68	0.703	0.71	0.38	0.37	0.47
AdaBoostM1	69.66	0.70	0.69	0.2	0.67	0.67	0.661	0.69	0.44	0.32	0.44
HMM	37.92	0.14	0.37	0	0.50	0.52	0.209	0.37	0.37	0	0.5

Table 2. PCA results

Classifiers	Correctly Classified Instances (%)	Pre	Rec	k	ROC	PRC Area	F-Measure	TP Rate	FP Rate	MCC	Root mean squared error
IB1	54.49	0.57	0.54	0.0	0.55	0.55	0.551	0.54	0.43	0.10	0.67
J48	69.66	0.72	0.69	0.2	0.65	0.64	0.646	0.69	0.47	0.33	0.457
Ran For	67.97	0.67	0.68	0.2	0.65	0.65	0.652	0.68	0.44	0.27	0.46
MLP	68.53	0.68	0.68	0.2	0.65	0.65	0.656	0.68	0.44	0.29	0.46
SMO	71.06	0.72	0.71	0.3	0.63	0.62	0.676	0.71	0.43	0.36	0.53
Jrip	69.94	0.69	0.69	0.3	0.64	0.63	0.686	0.69	0.39	0.33	0.45
RBF	70.78	0.70	0.70	0.3	0.67	0.62	0.685	0.70	0.41	0.34	0.44
Naïve Bayes	**71.34**	0.77	0.71	0.3	0.70	0.69	0.705	0.71	0.36	0.37	0.5
AdaBoostM1	69.66	0.69	0.69	0.28	0.67	0.67	0.667	0.69	0.43	0.32	0.44
HMM	37.92	0.14	0.37	0	0.5	0.52	0.209	0.37	0.37	0	0.5

Table 3. ILIOU method results

Classifiers	Correctly Classified Instances (%)	Pre	Rec	k	ROC	PRC Area	F-Measure	TP Rate	FP Rate	MCC	Root mean squared error
IB1	98.8	0.988	0.988	0.97	0.990	0.984	0.988	0.988	0.007	0.975	0.10
J48	95.6	0.958	0.956	0.90	0.964	0.950	0.956	0.956	0.035	0.909	0.20
Ran For	97.6	0.977	0.976	0.94	0.997	0.997	0.976	0.976	0.015	0.951	0.13
MLP	**100**	1	1	1	1	1	1	1	0	1	0.01
SMO	78.4	0.78	0.78	0.5	0.75	0.71	0.781	0.78	0.26	0.53	0.46
Jrip	96	0.96	0.96	0.9	0.97	0.95	0.96	0.96	0.02	0.92	0.19
RBF	85.6	0.89	0.85	0.7	0.92	0.91	0.858	0.85	0.08	0.74	0.31
Naïve Bayes	75.2	0.74	0.75	0.4	0.86	0.86	0.745	0.75	0.31	0.45	0.47
AdaBoostM1	92	0.92	0.92	0.8	0.96	0.95	0.921	0.92	0.06	0.83	0.23
HMM	38	0.14	0.38	0	0.5	0.52	0.209	0.38	0.38	0	0.5

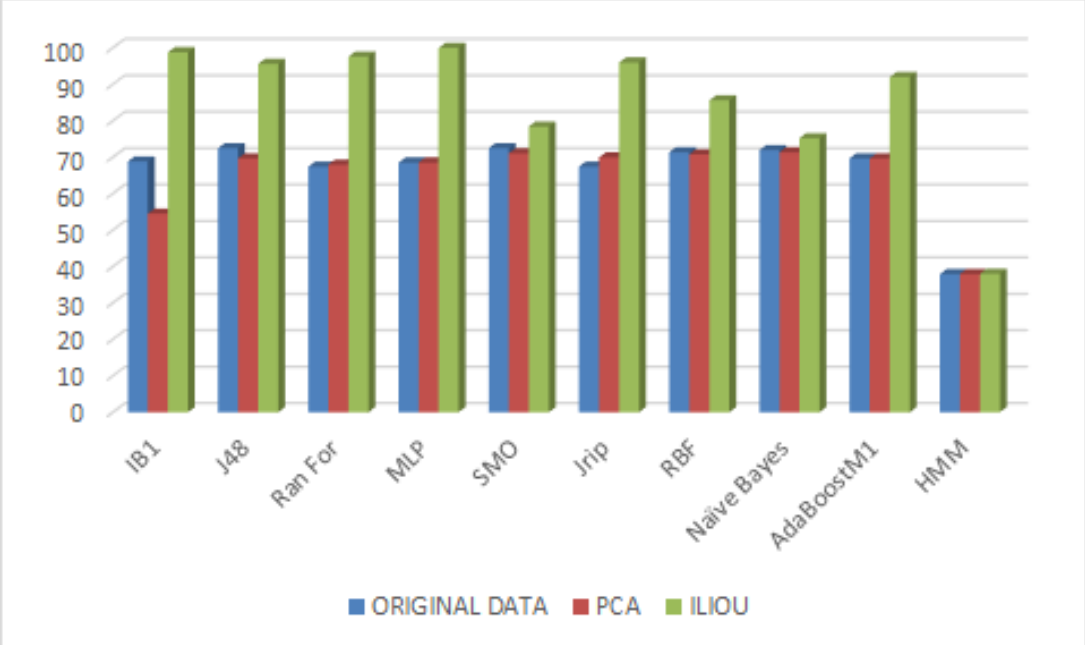

Figure 2. Experimental Results

6. CONCLUSIONS

Data have quality if they satisfy the requirements of the intended use. Factors, such as
accuracy, completeness, consistency, timeliness, believability and interpretability, affect data quality. Given that large real-world databases and data warehouses often consist of "bad" quality data, preparation, cleaning and transformation of data comprises the majority of the work in a data mining procedure. Thus, data preprocessing is a vital procedure for data mining process so that "good" quality data lead us to "good" quality results.

The innovation of our work is that ILIOU preprocessing method not only reduce the variables of the initial dataset as every preprocessing method do but significantly improves the classification results as well, comparing with PCA which is a well-known preprocessing method.

In this paper we used 10 classification schemes. According to the classification results, Iliou data preprocessing method not only outperform PCA but also achieves 100% classification performance.
Concluding, Iliou method can be used to importantly boost classification algorithms performance in similar datasets and seems to be the best preprocessing approach for suicide prediction based on family history.

In the future, a significant challenge for researchers, practitioners and data scientists would be to collaborate in order to create more and bigger databases, and test Iliou and other preprocessing methods with even more classification schemes. Furthermore, Iliou data preprocessing method could be modified so as achieve less classification time and could be ensembled in a classification algorithm.

7. REFERENCES

1. Jacobs EA, Agger-Gupta N, Chen AH, Piotrowski A, Hardt EJ. Language Barriers in Health Care Settings: An Annotated Bibliography of the Research Literature. Woodland Hills, CA: The California Endowment; 2003. pp. 1–80.

2. De Leo, D., Burgis, S., Bertolote, J.M, Kerkhof, A.J.F.M., Bille Brahe, U. (2006). Definitions of Suicidal Behavior. Crisis, 27 (1), 4 -15.
3. O'Carroll, P.W., Berman, A.L., Maris, R.W., Moscicki, E.K.,Tanney, B.L., & Silverman, M.M. (1996). Beyond the Tower of Babel: A nomenclature for suicidology. Suicide and Life Threatening Behavior, 26, 237–252
4. Hawton, K; Saunders, KE; O'Connor, RC (Jun 23, 2012). «Self-harm and suicide in adolescents». *Lancet* 379 (9834): 2373–82.
5. Vijayakumar, L; Kumar, MS; Vijayakumar, V (2011 May). «Substance use and suicide». *Current opinion in psychiatry* 24 (3): 197–202
6. Chang, B; Gitlin, D; Patel, R (2011 Sep). «The depressed patient and suicidal patient in the emergency department: evidence-based management and treatment strategies». *Emergency medicine practice* 13 (9): 1–23
7. Simpson, G; Tate, R (2007 Dec). «Suicidality in people surviving a traumatic brain injury: prevalence, risk factors and implications for clinical management». *Brain injury : [BI]* 21 (13–14): 1335–51
8. Miller, M; Azrael, D; Barber, C (2012 Apr). «Suicide mortality in the United States: the importance of attending to method in understanding population-level disparities in the burden of suicide». *Annual review of public health,* 33: 393–408.
9. Qin P, Agerbo E, Mortensen PB (April 2003). «Suicide risk in relation to socioeconomic, demographic, psychiatric, and familial factors: a national register-based study of all suicides in Denmark, 1981–1997». *Am J Psychiatry* 160 (4): 765–72
10. Gilliland, Richard K. James, Burl E. *Crisis intervention strategies* (7th ed. έκδοση). Belmont, CA: Brooks/Cole, σελ. 215
11. Brent, DA; Melhem, N (2008 Jun). «Familial transmission of suicidal behavior». *The Psychiatric clinics of North America* 31 (2): 157–77
12. Rozanov, V; Carli, V (2012 Jul). «Suicide among war veterans.». *International journal of environmental research and public health* 9 (7): 2504–19.
13. García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. Berlin: Springer; 2015.

14. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques, 3rd ed. Burlington: Morgan Kaufmann Publishers Inc; 2011.
15. Zaki MJ, Meira W. Data Mining and Analysis: Fundamental Concepts and Algorithms. New York: Cambridge University Press; 2014.
16. R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, no. 12, 1995, pp. 1137–1143.
17. Waikato Environment for Knowledge Analysis, Data Mining Software in Java, available online: http://www.cs.waikato.ac.nz/ml/index.html, [Accessed 10/4/2018].
18. T. Iliou, G. Konstantopoulou, M. Ntekouli, D. Lymberopoulos, K. Assimakopoulos, D. Galiatsatos, G. Anastassopoulos, (2016) “Machine Learning Preprocessing Method for Suicide Prediction” Proc in the AIAI 2016, pp 53-60.