ANNALS OF "DUNAREA DE JOS" UNIVERSITY OF GALATI
MATHEMATICS, PHYSICS, THEORETICAL MECHANICS
FASCICLE II, YEAR X (XLI) 2018, No. 1

Article DOI: https://doi.org/10.35219/ann-ugal-math-phys-mec.2018.1.10

EVOLUTIONARY ALGORITHM APPLIED FOR IMPROVING THE ACCURACY OF THE AUTOMATED DETECTION OF PSYCHEDELIC AMPHETAMINES

Catalin Negoita[1*], Mirela Praisler[2]

[1]*Faculty of Sciences and Environment, Department of Chemistry, Physics and Environment, „Dunărea de Jos" University of Galati, Romania*, e-mail: c.negoita@ugal.ro

Abstract
We are presenting a comparative study regarding the improvement of the correct classification rate of an artificial intelligence application designed to recognize the class identity of psychedelic amphetamines based on the similarity of their ATR-FTIR spectra. For this purpose, the most relevant absorptions were first selected by using a metaheuristic, i.e. a genetic algorithm (GA). The latter is a type of evolutionary algorithm (EA) that mimics the natural selection process and which is recommended especially for improving classification models and optimizing searching procedures. Regression models were built with the original spectral dataset, representing the absorptions measured at 1869 wavenumbers, and with the dataset formed by the absorptions measured at the 187 wavenumbers selected as being the most significant by the genetic algorithm. Both models were tested by using the K-Nearest Neighbors and the Random Forest procedures. Several classification figures of merit were determined and compared for these two cases. The results indicate that the GA wavenumber selection leads to a significant improvement of the classification accuracy.

Keywords: Amphetamines, evolutionary algorithm, genetic algorithm.

1. INTRODUCTION

Obtaining a highly correct classification rate of the spectra of illicit drugs of abuse is a real challenge. This is especially true in the case of small molecules such as the synthetic drugs of abuse such as amphetamines and cannabinoids. Previous studies have indicated that good results may be obtained with infrared spectra processed with artificial intelligence techniques such as Principal Component Analysis (PCA) [1, 2], Soft Independent Modeling of Class Analogy (SIMCA) [3, 4] or Artificial Neural Networks [5-7].

In this paper we are presenting the results obtained with a combination of artificial intelligence methods designed to perform the automatic class identity assignment for two of the main classes of designer drugs, i.e. psyhedelic amphetamines and JWH cannabinoids. For this purpose, Partial Least Squares Regression (PLSR) and genetic algorithms (GA) have been used in order to extract the most discriminating features from the ATR-FTIR spectra of the targetted compounds. The selected dataset is then used as input for two classification models, i.e. K-Nearest Neighbors (KNN) and Random Forest.

2. EXPERIMENTAL

The initial database consisted of 60 normalized ATR-FTIR spectra, which had been recorded between 4000 and 400 cm^{-1}, by performing 1868 scans with a 5 cm^{-1} resolution (see Fig. 1). The samples were selected from 3 classes of compounds: psychedelic amphetamines, JWH cannabinoids and negatives (randomly selected substances of forensic interest). This working dataset was splited into two subsets representing a training subset and a model validation subset.

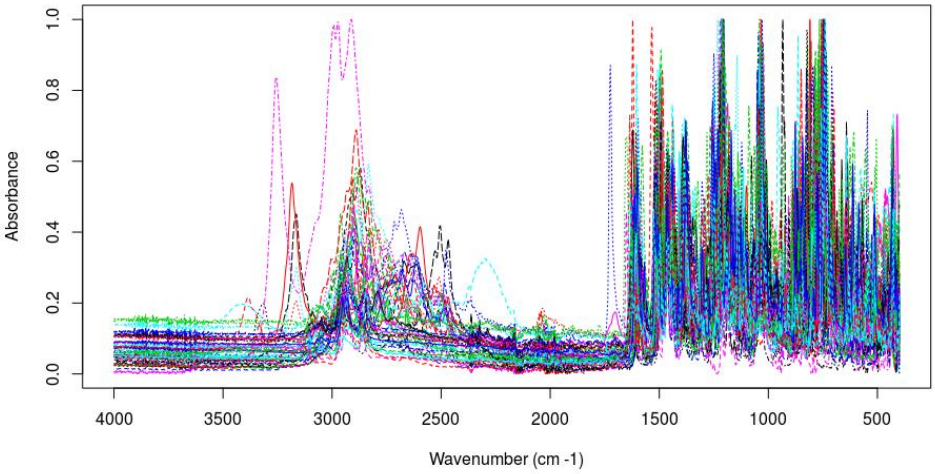


Fig. 1. ATR-FTIR spectra of the 2C-x and DOx hallucinogenic amphetamines and cannabinoids included in the original database.

The model development was done in *R* statistical environment, by using various packages such as *plsVarSel* and *pls*. PLSR and GA have been performed by using the *pls* module and 10 components. The models have been validated by using the leave-one-out cross-validation (LOO) method.

3. RESULTS AND DISCUSSION

The combination of GA and PLSR (GA-PLSR) is inspired from the biological evolution theory and the natural selection process. Applying GA involves the following steps: a) one variable is set randomly as bit '1', while '0' means non-selection; b) a PLSR model is fitted to each variable set and the performance is assessed by using the leave one out cross-validation procedure; c) the variable sets yielding the best results are selected to survive until the next generation emerges; d) crossover and mutation are performed and new variable sets are formed; e) the surviving and modified variable sets are then used in the next iteration.

In order to assess the effect of the variable (wavenumber) selection procedure, *plsr* was applied to the initial database (containing the full ATR-FTIR spectra, i.e. the absorbances measured at the 1868 corresponding wavenumbers). A new database was formed with the selected variables, which contained the absorbances measured at only 186 wavenumbers, which were selected by GA (see Fig. 3).

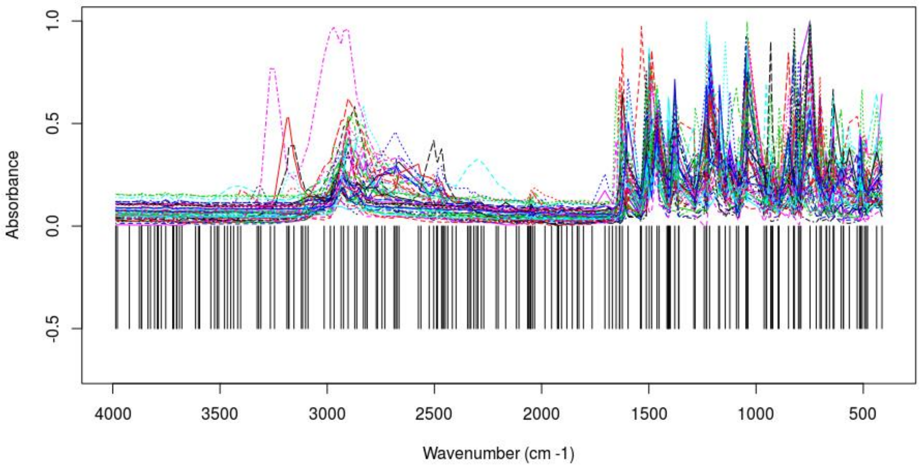
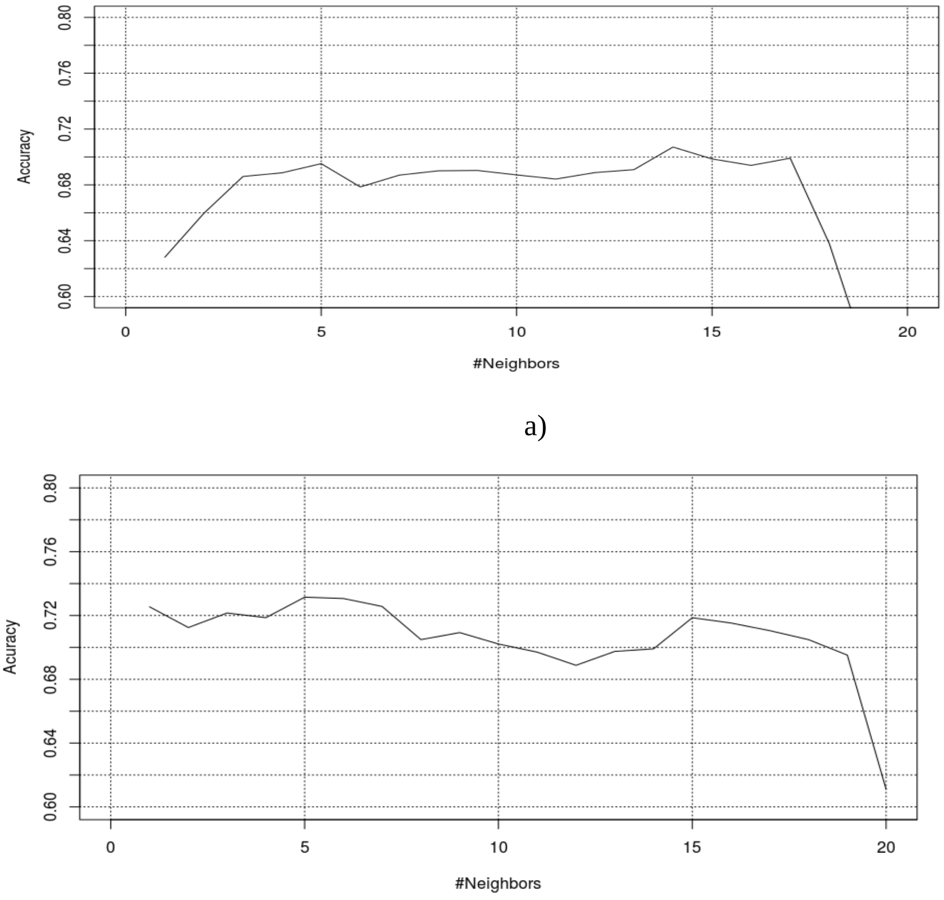


Fig. 2. ATR-FTIR spectra of the 2C-x and DOx hallucinogenic amphetamines and cannabinoids included in the database obtained by GA wavenumber selection. The selected wavenumbers are displayed below the spectra.



a)

b)

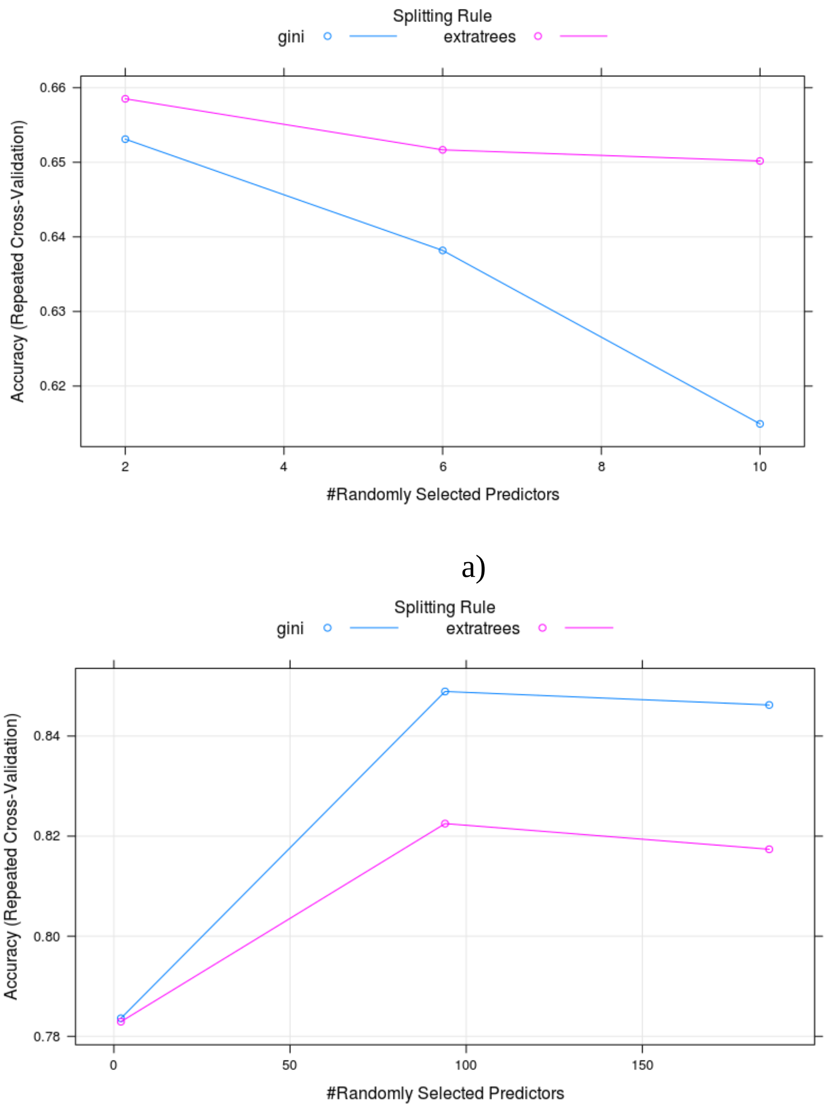
Fig. 3. KNN accuracy for: a)the initial full dataset and b)the GA selected datset

The threshold, number of iterations and the population size have been chosen at the beginning of the procedure. GA-PLSR runs have been performed by varying the GA threshold between 7 and 20, and the number of iterations between 5 and 60. The size of the initial population was 100. Every *plsr* result has been evaluated by calculating the mean squared error of prediction (RMSEP) and the R^2 coefficient. RMSEP has been calculated by using the data obtained for up to 10 components for both datasets. The prediction R^2 was calculated as:

$$R^2 = 1 - SSE/SST \tag{1}$$

where SST is the corrected total sum of squares of the response and SSE is the sum of squared errors for the fitted values.

Figure 3. shows the results obtained by training the KNN model. A comparative overview of the accuracy obtained for the GA generated (selected) dataset and the full (initial) dataset are presented in Fig. 3. It indicates that the selection performed by the genetic algorithm leads to a significantly better accuracy.



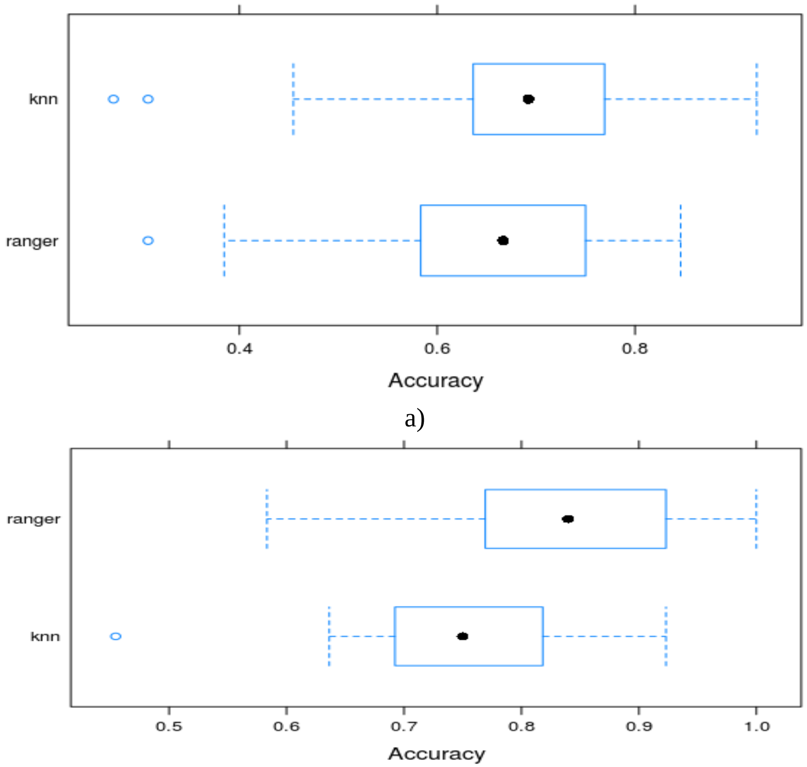
a)

b)

Fig. 4. Random Forest Model Accuracy for: a) the full dataset and b)the GA selected dataset

The Random Forest model was tested with (randomly selected) 10 predictors. Fig. 4. shows that a better accuracy is obtained when the model runs on the dataset generated by the genetic algorithm.

Fig. 5 illustrates the accuracy obtained for both models on both datasets (see Fig. 5). It indicates that both KNN and Random Forrest techniques provide very accurate results in assigning the class identity of the targetted classes of compounds. Secondly, Fig. 5 indicates that when the number of variables is much larger than the number of spectra, as it is the case when the initial full dataset is used as input, a slightly better accuracy is obtained for the KNN model. On the other hand, in the case of the GA selected dataset, a better accurarcy is obtained with the Random Forrest model.



a)

b)

Fig. 5. Comparison of the accuracy obtained for the Random Forrest model and of the KNN method applied for: a) the full dataset; b)the GA selected dataset

4. CONCLUSIONS

Despite the relatively low signal-to-noise ratio, the proposed models successfully identify the class membership of psychedelic amphetamines, cannabinoids and distinguish them from a large variety of negatives. The GA-PLSR model clearly identifies the absorptions that are characterizing the molecular structure similarity of the compounds belonging to each of the modeled classes of compounds. As the selected wavenumbers cover practically the whole spectrum, we may conclude that there is no preferred infrared region.

References

1. M. Praisler, I. Dirinck, J. Van Bocxlaer, A. De Leenheer, D.L. Massart,Exploratory analysis for the automated identification of amphetamines from vapour-phase FTIR spectra, *Analytica Chimica Acta* 404 (2000) 303- 317.

2. M. Praisler, J. Van Bocxlaer, A. De Leenheer, D.L. Massart, Automated recognition of ergogenic aids using Soft Independent Modeling of Class Analogy (SIMCA), *Turkish Journal of Chemistry* 26 (2002) 45-58.
3. M. Praisler, J. Van Bocxlaer, A. De Leenheer, D.L. Massart, Chemometric detection of thermally degraded samples in the analysis of drugs of abuse with GC-FTIR spectroscopy, *Journal of Chromatography A* 962, (2002) 161-173.
4. M. Praisler, I. Dirinck, J. Van Bocxlaer, A. De Leenheer, D. L. Massart, Identification of novel illicit amphetamines from vapor-phase FTIR spectra - a chemometrical solution, *Talanta* 53 (2000) 155-170.
5. S. Gosav, M. Praisler, D. O. Dorohoi, G. Popa, Structure – Activity Correlations for Illicit Amphetamines Using ANN and Constitutional Descriptors, *Talanta* 70 (2006) 922-928.
6. S. Gosav, M. Praisler, D. O. Dorohoi, ANN Expert System Screening for Illicit Amphetamines using Molecular Descriptors, *Journal of Molecular Structure* 834-836 (2007) 188-194.
7. S. Gosav, M. Praisler, Artificial Neural Networks Built for the Recognition of Illicit Amphetamines Using a Concatenated Database, *Romanian Reports of Physics* 54-9/10 (2009) 929–935.