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.2018.1.02

NAIVE BAYESIAN CLASSIFIER DETECTING PHENETHYLAMINES BASED ON THEIR VIBRATIONAL SPECTRA AND ASSOCIATED EIGENVALUES

Stefanut Ciochina[2], Mirela Praisler[1]

[1]*"Dunarea de Jos" University of Galati, Department of Chemistry, Physics and Environment*
[2]*"Dunarea de Jos" University of Galati, Department of Mathematics and Computer Science*
Stefanut.Ciochina@ugal.ro, Mirela.Praisler@ugal.ro

Abstract
In this paper we are presenting an artificial intelligence application designed to screen for controlled phenethylamines. The training set consists of 30 vibrational spectra of stimulant and hallucinogenic amphetamines, as well as of negatives (non-amphetamines representing randomly selected chemicals of forensic interest). The Naive Bayesian Classifier indicates the likelihood that a substance belongs to one of the predefined classes. For this aim, in our case, the Principal Component Analysis (PCA) scores of the targeted compounds are subjected to a Naive Bayesian Classifier. The advantages of combining these two pattern recognition methods over the use of each method independently are discussed from the point of view of the detection efficiency.

Keywords: Amphetamines, ephedrines, pattern recognition.

1. INTRODUCTION

Taking into account the proliferation of clandestine laboratories manufacturing amphetamines, there is an increasing need for portable instruments able to screen *in-situ* for any compounds having similar molecular structures [1, 2]. The challenge is to be able not only to distinguish them from other compounds, but also to refine the class identity assignment according to known structure – activity relationships, i.e. to distinguish the amphetamines that are mainly acting as stimulants of the central nervous system (CNS) from their hallucinogenic counterparts [1, 3].

Expert systems based on a variety of pattern recognition techniques have been developped in order to achieve this goal. In many cases, a combination of data-processing techniques has been necessary to obtain high correct classification rates [4-6].

In this paper we are also presenting a combination of two such chemometrical techniques, which was developped as an application processing the spectra measured with a portable laser infrared spectrometer designed to detect amphetamine analogues according to their pharmacological activity (stimulants or hallucinogens). More specifically, the pre-processed spectra are first processed by using Principal Component Analysis (PCA) [7]. Then, the PCA scores are used as input for the classification process, which is performed with the Naïve Bayes classifier [8].

2. EXPERIMENTAL PART

The results presented in this paper have been obtained with the spectra recorded by using the UT7 quantum cascade laser (QCL), which emits in the 1550 - 1330 cm^{-1} range. The input spectral database consists of the absorptions recorded within this spectral window of 30 illicit amphetamines, i.e. 7 illicit stimulant amphetamines (class code M), 6 hallucinogenic amphetamines (class code T) and 17 non-amphetamines (class code N) of toxicological concern [9-12]. The absorption has been measured 5 cm^{-1} apart.

In order to enhance the most relevant spectral features, a w_{MT} feature weight has been determined by using the Fisher function [9, 10]. For this purpose, the spectra of the stimulant (M) and hallucinogenic amphetamines (T) have been included in class I and the spectra of the negatives (N) in class II. The M compounds include amphetamine and its main analogues, while the T class is formed by 3,4-methylenedioxyamphetamines and its main analogues.

A new database, created by preprocessing the spectra with the $(w_{MT} - 1)^2$ feature weight, was further used as input for an exploratory analysis performed by PCA, by using the MATLAB software. The number of principal components (PCs) that are necessary modeling and discrimination has been established by analyzing the eigenvalues and the explained variance of the first five PCs. These factors indicated that most of the information is found to be represented by the first two PCs.

The scores obtained for these PCs have been used in order to evaluate the potential overlap of the clusters identified in the score plot, based on kernel density estimations. The same scores have also been used for assessing the number of clusters that may be clearly distinguished based on their cohesion, as measured by the individual and mean values of the Silhouette index [13, 14].

A database formed with the PCA scores of the first two PCs has been then subjected to the Naïve Bayes classifier. Its efficiency in assigning the class identity has been evaluated based on structure – activity correlations.

3. RESULTS AND DISCUSSION

The w_{MT} feature weight obtained as mentioned above is presented in Fig. 1. It enhances especially the absorptions found around 1475 cm^{-1}. These absorptions are characteristic to the hallucinogenic amphetamines, these compounds displaying a very strong and stable absorption band in this spectral region. The 1445 cm^{-1} peak is characteristic to the stimulant amphetamines. The difference in intensity between the 1475 and the 1445 cm^{-1} peak may be explained by the fact that the absorptions characteristic to the stimulant amphetamines are weaker, their intensity and position displaying a relatively significant variation (in comparison with the high stability of the bands specific to the hallucinogens) [9-12].

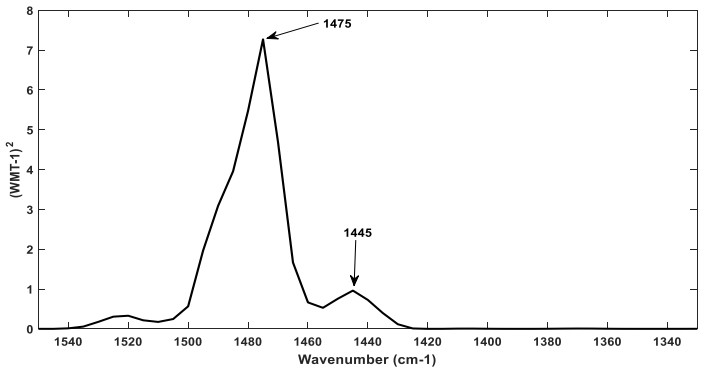


Figure 1. $(w_{MT} -1)^2$ selective amplifier used for preprocessing the infrared spectra.

The number of principal components (PCs) to be taken into account for further modeling has been established by analyzing their eigenvalues and explained variance. As shown in Table 1, the first two PCs are cumulating most of the explained variance (99.73%), the contribution of the following PCs being negligible. This behavior is also indicated by the corresponding eigenvalues. Hence, the database used for classification was formed with the PCA scores of the first two PCs.

Table 1. Eigenvalues and explained variance of the first principal components obtained for the $(w_{MT} - 1)^2$ *preprocessed spectra recorded in the 1550 - 1330* cm^{-1} *spectral domain.*

Eingenvalue	Proportion of variance	Cumulative proportion
6,3932	0,954965	0,954965
0,2835	0,042352	**0,997318**
0,1370	0,002049	0,999367
0,0023	0,000348	0,999715
0,0010	0,000151	0,999866

The resulting score plot is presented in Fig. 2. It indicates that the most distinguishable cluster is found in quadrant IV, being formed by the hallucinogenic amphetamines (T). The cluster formed by the stimulant amphetamines (M) is more dispersed. Most of the points associated to these drugs of abuse are found in quadrant I, but some (M4, M7 and M17) are found in quadrant II, very close to the cluster formed by the negatives at the border between quadrants II and III.

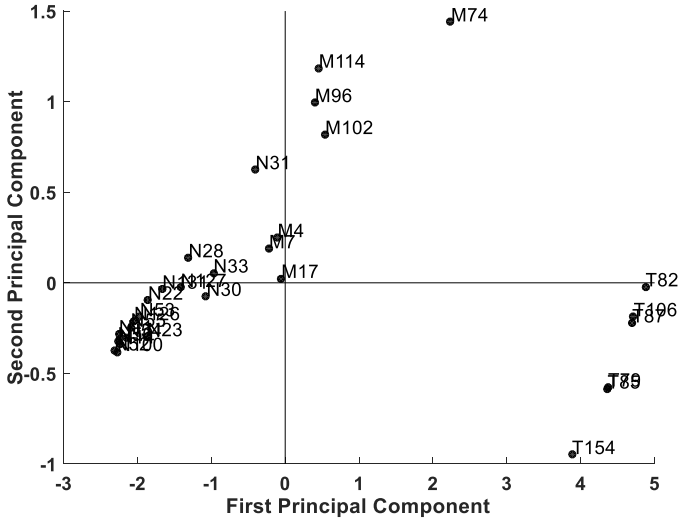
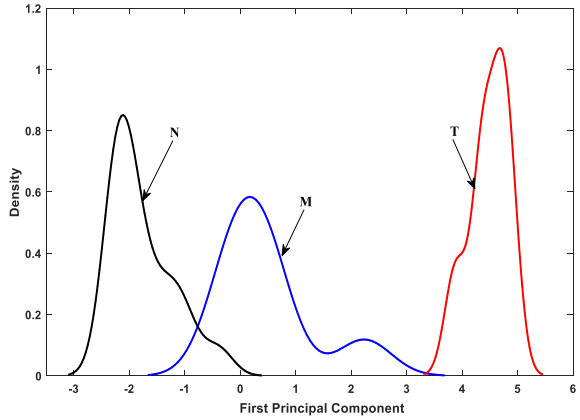
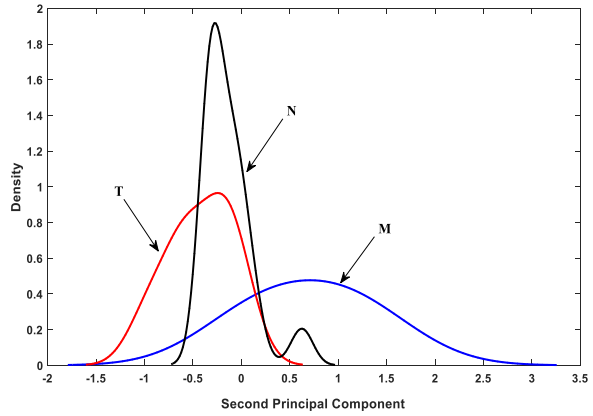


Figure 2. Score plot obtained by using the first two principal components determined for stimulant amphetamines (M), hallucinogenic amphetamines (T) and negatives (N).

Therefore, in order to assess to what extent the clusters may overlap, their distribution has been evaluated by using kernel density estimations. The results indicate (see Fig. 3) that PC1 clearly distinguishes the hallucinogens. On the other hand, some misclassifications may be expected for those stimulant amphetamines and negatives that have small negative PC1 scores. The mean values and the standard deviations of the PC1 and PC2 scores calculated for the samples belonging to each of the modeled classes of compounds are presented in Table 2.



a)



b)

Figure 3. Estimated density associated to the PCA scores of stimulant amphetamines (M), hallucinogenic amphetamines (T) and negatives (N): a) PC1; b)PC2.

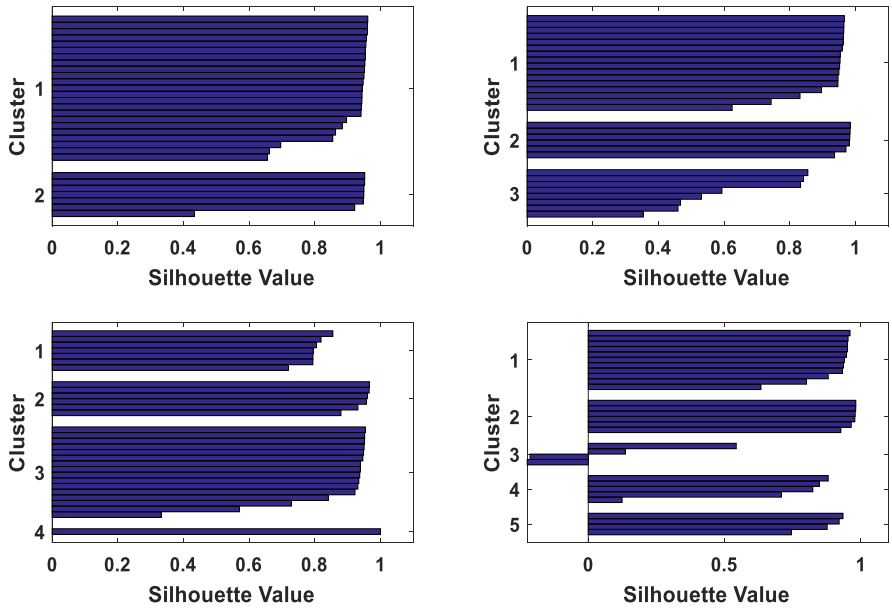
Table 2. Mean values and standard deviations of the PC1 and PC2 scores determined for the compounds forming the modeled classes of compounds.

PC1	**M**	**T**	**N**
Mean	0.4663	4.4825	-1.7741
Standard deviation	0.8374	0.3588	0.5572
PC2	**M**	**T**	**N**
Mean	0.6997	-0.4235	-0.4235
Standard deviation	0.5485	0.3412	0.3412

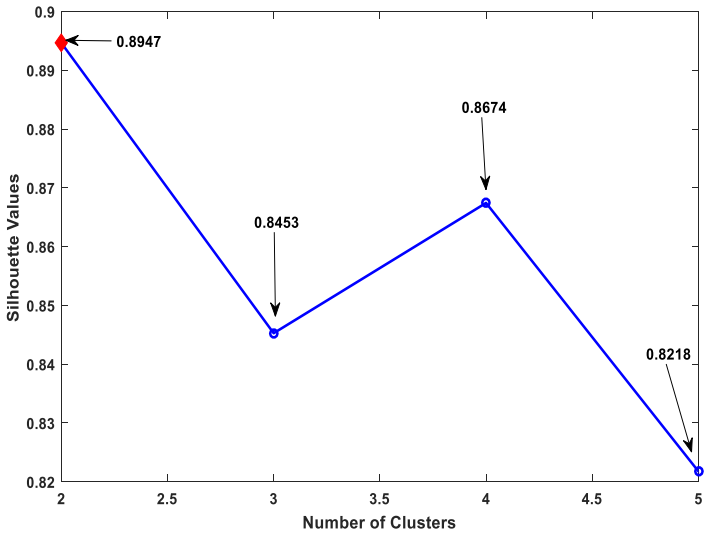
Taking into account the potential overlap between the clusters of M and N compounds, the number of clusters that may be clearly distinguished has been assessed based on their cohesion. For this purpose, they have been characterized by using the Silhouette index (see Fig. 4). It indicates that although the best results may be obtained for two clusters. The second best result is obtained for four clusters. However, the fourth clusters should be formed by only one compound, which is not acceptable. In the case of five clusters, not only some clusters would also be formed by only a few compounds, but the samples included in the third cluster are characterized by negative Silhouette values. This indicates that those samples might have been assigned to the wrong cluster.

On the other hand, Fig. 4a indicates that three clusters may be distinguished well enough. In addition, the mean Silhouette index obtained in this case does not differ very much from the value

determined for four clusters (see Fig. 4b). Therefore, we have concluded that good classification accuracy may be expected for three clusters (M, T and N).



a)



b)

Figure 4. Choosing the number of clusters based on their cohesion: a) individual values; b) mean values.

The database consisting of the first two PCs has been then analyzed by using the Naïve Bayes classifier (see Fig. 5). The system is very efficient in recognizing hallucinogenic amphetamines (T). Very good correct classification rates are also obtained in distinguishing stimulant amphetamines (M) from negatives (N).

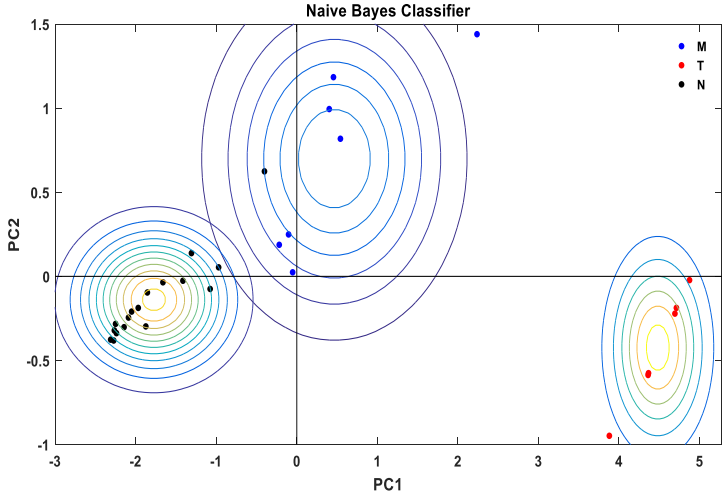


Figure 5. Class identity assignment based on the Naïve Bayes classifier.

Table 3 presents the classification results obtained for some relevant samples. It shows some examples of negatives that are classified as (false) stimulant amphetamines. These is the case of ephedrine (rel-(R,S)-2-(methylamino)-1-phenylpropan-1-ol) and of its analogues or isomers (e.g. pseudoephedrine or norephedrine). These compounds have molecular structures that are extremely similar to amphetamine ((RS)-1-phenylpropan-2-amine) and its analogues, i.e. to the compounds forming the M cluster. The structural similarity yields similar spectra and thus these compounds are classified as stimulant amphetamines.

Table 3. Classification results determined for samples belonging to the modeled classes of compounds.

Tested	**Assigned class identity**	**Posterior**			**Cost**		
		M	T	N	M	T	N
E13	**M**	0.8395	0.0000	0.1605	0.1605	1.0000	0.8395
E39	**M**	1.0000	0.0000	0.0000	0.0000	1.0000	1.0000
E104	**M**	1.0000	0.0000	0.0000	0.0000	1.0000	1.0000
E110	**M**	0.6876	0.0000	0.3124	0.3124	1.0000	0.6876
E111	**M**	0.7362	0.0000	0.2638	0.2638	1.0000	0.7362
E129	**M**	1.0000	0.0000	0.0000	0.0000	1.0000	1.0000
M74	**M**	1.0000	0.0000	0.0000	0.0000	1.0000	1.0000
T154	**T**	0.0001	0.9999	0.0000	0.9999	0.0001	1.0000
N31	**M**	0.9386	0.0000	0.0614	0.0614	1.0000	0.9386
M17	**M**	0.6111	0.0000	0.3889	0.3889	1.0000	0.6111
T82	**T**	0.0000	1.0000	0.0000	1.0000	0.0000	1.0000
N33	**N**	0.0168	0.0000	0.9832	0.9832	1.0000	0.0168

However, we must keep in mind that these compounds are also stimulants of the CNS, although milder than stimulant amphetamines. In adition, ephedrines are the most frequently used precursors used by clandestine laboratories to synthetise stimulant amphetamines, especially methamphetamine. Hence, from these points of view, their classification in the M class is useful.

N31 - caffeine (which is also a CNS stimulant) is also misclassified, as its spectrum is similar to those of M amphetamines in the narrow spectral window of the QCL source of infrared radiation.

On the other hand, caffeine does have a molecular structure quite different from the amphetamines. Once analyzed in laboratory conditions based on its full infrared spectrum (4000 – 600 cm^{-1}), its correct (negative) identity will be revealed without doubt.

4. CONCLUSION

We may conclude that the Naïve Bayes classifier is useful for distinguishing amphetamines from other types of compounds, as well as among themselves. Its accuracy recommends it as an efficient forensic *in-situ* screening tool. The best classification accuracy is obtained for the hallucinogenic amphetamines, i.e. the most toxic compounds among those that have been modeled for this application. In their case, the system is both very sensitive and selective.

Acknowledgements

Part of the research has been funded by EC under the grant agreement n° FP7-SEC-2009-242309 DIRAC. The work of Stefanut Ciochina has been funded by the Romanian Ministry of European Funds within the POSDRU/107/1.5/S/76822 project. The authors are grateful for the financial support.

References

1. S. Karch, *Drug of Abuse Handbook*, 2nd ed. Boca Raton: CRC Press, 2007.
2. J. M. Chalmers, H. G. M. Edwards, M. D. Hargreaves, *Infrared and Raman Spectroscopy in Forensic science*, Chichester: Wiley, 2012.
3. R. Laing (Ed,), *Hallucinogens. A Forensic Drug Handbook*, London: Academic Press, 2003.
4. S. Gosav, M. Praisler, D. O. Dorohoi, G. Popa, Structure – Activity Correlations for Illicit Amphetamines Using ANN and Constitutional Descriptors, *Talanta -The International Journal of Pure and Applied Chemistry* 70 (2006) 922-928.
5. 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.
6. 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.
7. I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York: Springer, 2002.
8. P. Cichosz, *Naïve Bayes classifier*, in P. Cichosz (ed.), *Data Mining Algorithms: Explained Using R*, Chichester: Wiley, 2015.
9. M. Praisler, S. Ciochina, A. Stoica, Artificial intelligence application built for ATS detection with a new portable hollow fiber IRAS spectrometer, *Proceedings of the 18th International Conference on System Theory, Control and Computing ICSTCC 20*14, 17-19 October 2014, Sinaia, Romania, p. 237-242.
10. M. Praisler, S. Ciochina, A. Stoica, L. Dumitriu, Signal selective amplification: a solution for an improved detection of amphetamines with QCL equipped portable GC-IRAS spectrometers, *Proceedings of the 18th International Conference on System Theory, Control and Computing ICSTCC 2014*, 17-19 October 2014, Sinaia, Romania, p. 909-914.
11. M. Praisler, S. Ciochina, PCA Evaluation of Quantum Cascade Lasers as Radiation Sources for Portable IRAS Systems Detecting Amphetamines, *2013 E-Health and Bioengineering Conference* (EHB 2013), Article number 6707370 (21-23 November 2013, Iasi, Romania).

12. S. Ciochina, M. Praisler, Pattern Recognition Techniques Applied for the Detection of Amphetamines Based on Infrared Laser Spectroscopy, *2013 E-Health and Bioengineering Conference* (EHB 2013), Article number 6707369 (21-23 November 2013, Iasi, Romania)
13. A. Starczewski, A. Krzyżak A. *Performance Evaluation of the Silhouette Index*, in L. Rutkowski, M. Korytkowski, R. Scherer, R. Tadeusiewicz, L. Zadeh, J. Zurada J. (Eds.) *Artificial Intelligence and Soft Computing* ICAISC 2015, Lecture Notes in Computer Science, vol. 9120, Springer, Cha
14. L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data. An Introduction to Cluster Analysis*, Willey, 2005.
15. M. Praisler, S. Ciochina, Global clustering quality coefficient assessing the efficiency of PCA class identity assignment, *Journal of Analytical Methods in Chemistry*, volume 2014 (2014), Article ID 342497.