ESANN'2006 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 26-28 April 2006, d-side publi., ISBN 2-930307-06-4.

Pattern analysis in illicit heroin seizures: a novel application of machine learning algorithms

Frédéric Ratle[1], Anne-Laure Terrettaz[2], Mikhaïl Kanevski[1], Pierre Esseiva[2]
and Olivier Ribaux[2] *

1- Institut de Géomatique et d'Analyse du Risque - Faculté des Géosciences et de l'Environnement - Université de Lausanne, Amphipôle, CH-1015 - Switzerland

2- Institut de Police Scientifique et de Criminologie - Ecole des Sciences Criminelles - Université de Lausanne, Batochime, CH-1015 - Switzerland

Abstract. An application of machine learning algorithms to the clustering and classification of chemical data concerning heroin seizures is presented. The data concerns the chemical constituents of heroin as given by a gas chromatography analysis. Following a preprocessing step, where the six initial constituents are reduced to only two significant features, the data are clustered in order to find natural classes which we have supposed to correspond to the country of origin. A classification is then made using a multi-layer perceptron, a probabilistic neural network, a radial basis function network and the k-nearest neighbors method. Results are encouraging and add important information to previous work in the field.

1 Introduction

The chemical profiling of illicit drug seizures has been the focus of recent attention in the crime mapping community. However, methods from machine learning are only beginning to find their use in this promising field of application. Whereas classical crime mapping research has mainly focused on the study of police events themselves (i.e., seizures, arrests, etc.), an extremely valuable information has long been neglected: the chemical signature of the seizures, which can enable crime analysts to establish links between drug distribution networks. Moreover, traditional crime mapping research has for long relied mainly on classical statistics and geographical information science, which provides a collection of tools called Geographical Information Systems (GIS). The latter refers to technology dealing with the processing and visualisation of spatial data. A thorough review of crime mapping applications of GIS can be found in [1]. The introduction of machine learning in the field could potentially reveal structures in crime data (or, on the contrary, reveal the absence of structure or patterns in a particular criminal activity).

As explained in [2], several types of chemical profiles can be studied, and each of these provide information on a different level (producing country, distribution network, etc.):

*Authors would like to thank the Swiss National Science Foundation (grant no.105211-107862) for the financial support.

- The major constituents: the major chemical compounds found in a sample.
- The minor constituents: the minor chemical compounds in a sample.
- The solvent residues: these reveal information about the fabrication process.
- The inorganic compounds; for example, minerals absorbed by the plant of origin.
- The cutting agents; these usually provide information about the street-level distribution network.

Each of these require a particular experimental technique. In practice, these levels are often overlapping. For example, the dilution of the drug can take place either in the producing country or in the distribution country. These variations complicate the interpretation of the data. It is worth noting that cutting agents are most relevant with the study of cocaine, which usually arrives almost pure in the distribution country, whereas heroin is very often diluted prior to its transport. However, this problem will be the focus of future research, the main concern of this investigation being the analysis of the data concerning the major constituents of heroin found in seizures made in Switzerland. Those constituents are listed below:

1. Meconin
2. Acetylcodein
3. Acetylthebaol
4. Monoacetylmorphine
5. Noscapine
6. Papaverine

The concentration of each of these products is measured using a gas chromatography method elaborated by Guéniat in his doctoral thesis and explained in [2].

2 Previous studies in chemical drug profiling

An up-to-date and complete review of the field of chemical drug profiling can be found in Guéniat and Esseiva [2]. In this book, authors have tested several statistical methods for heroin profiling. Among other methods, they have mainly used similarity measures between samples to determine the main data classes. A methodology based on the square cosine function as a intercorrelation measurement is explained in further details in Esseiva et al. [3].

Also, principal component analysis (PCA) and soft independent modelling of class analogies (SIMCA) have been applied for dimensionality reduction. A radial basis function network has been trained on the processed data and showed encouraging results. However, the classes used for classification were based solely on indices of chemical similarities found between data points. This methodology was further developed by the same authors in [4].

In addition, another type of drug-related data was studied by Madden and Ryder [5]: Raman spectroscopy obtained from solid mixtures containing cocaine. The goal was to predict the cocaine concentration in a solid (based on the Raman spectrum) using k-nearest neighbors, neural networks and partial least squares. They have also used a genetic algorithm to perform feature selection. However, their study has been constrained by a very limited number of experimental samples, even though results were good. Also, the experimental method of sample analysis is fundamentally different from the one used in this study (gas phase chromatography).

Similarly, Raman spectroscopy data was studied in [6] using support vector machines with RBF and polynomial kernels, KNN, the C4.5 decision tree and a naive Bayes classifier. The goal of the classification algorithm was to discriminate samples containing acetaminophen (used as a cutting agent) from those that do not. The RBF-kernel SVM outperformed all the other algorithms on a dataset of 217 samples using 22-fold cross-validation.

3 Methodology and results

3.1 Experimental settings

The experiments were all made on the same platform (MS Windows, 3.20 GHz CPU) on Matlab, using the pattern classification toolbox by Stork and Yom-Tov [7], which implements algorithms described in Duda et al. [8]. Regarding the amount of data, 3003 samples were available, which is a very large dataset if we take into account the restricted accessibility of this type of data.

3.2 Reduction of dimensionality

The six major heroin constituents are considered here. Given the simple and deterministic nature of the chemical reactions involved in the heroin fabrication process, it is reasonable to suppose that some concentrations are linear functions of some others. PCA was therefore applied on the six constituents, after centering and normalization of the data. As expected, the two first principal components were responsible for 98.5% of the data variability. Those two first components were therefore kept for the remainder of the study.

3.3 Clustering

Clustering is here a central task, since it is here that structure or patterns within the data can be interpreted. The k-means algorithm was applied, with two, three and four clusters, However, by visual inspection, it can be seen that the data seems to separate in two main shapes.

Figure 1 shows the results obtained by k-means with two clusters. In [2], Guéniat and Esseiva had found 20 chemical classes. However, their work was based on similarity measures between samples; two samples were considered as belonging to two different classes if the similarity between them was less than a certain treshold. However, in this study, we have tried to find the clusters that seem the most “natural” given the shape of the data. From this result, we can emit the hypothesis that heroin found in Switzerland may come from two principal sources or countries. This hypothesis will, however, have to be studied more deeply by experts.

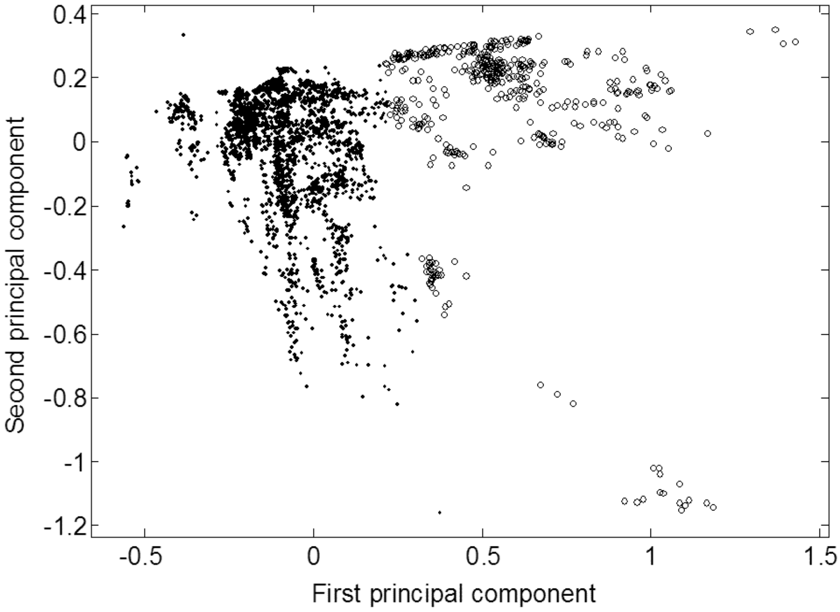

Fig. 1: Results obtained using k-means with two clusters.

3.4 Classification algorithm

Provided the previous clustering results, different classification algorithms have been applied to the clustered data in order assess their predictive potential regarding the two classes that have been found. Four types of methods were used in this investigation: a multi-layer perceptron (MLP), a radial basis function network (RBF), a probabilistic neural network (PNN) and the k-nearest neighbors method (KNN). The parameters to tune (along with the tested values) were the following:

- The number of hidden units for the MLP [1-15];
- The number of hidden units for the RBF network [2-20];
- The variance σ of the Gaussians for the PNN [0.001-0.1];
- The number of neighbors for KNN [1-30].

The quality of the classification was assessed using the test set method. The following evaluation procedure has been used:

1. Randomly separate the data in training, validation and test sets;
2. Train the model as many times as there are parameter values to test;
3. Choose the parameter values which gives the lower validation error. If several parameter values give the same error, choose the one that gives lower capacity;
4. Repeat steps 1 to 3 ten times;
5. Choose the parameter value by a majority vote.

Steps 4 and 5 were performed in order to increase the confidence of the error estimate. Table 1 summarizes the results obtained .

Model	Parameter value	Validation error	Test error
MLP	2 hidden units	0.135	0.153
RBF	6 hidden units	0.129	0.142
PNN	$\sigma = 0.035$	0.000	0.003
KNN	3 neighbors	0.002	0.002

Table 1: Validation and test errors for each model.

As we can see from the results, the PNN and the KNN give both a test error of the same order of magnitude. However, the MLP and the RBF network perform much worse both on the validation and the test set.

It is reasonable to ask whether or not the results would have been the same using all available features, i.e., without performing a PCA prior to clustering. In this particular case, we have observed, as previously stated, an explained variance of 98.5% using the two first principal components, which allows us to suppose the presence of a linear manifold in the space of the initial features. MLP and RBF networks may have proved better using more variables, but this would not have been useful from a practical point of view. Nonetheless, the introduction of other variables than the chemical components, e.g., space and time variables, will be a future topic of research.

Given the obtained results, which show two main clusters which are easily separable by simple classification algorithms, the hypothesis of two main classes in the data corresponding two different production sources seems likely. Further investigation from experts in crime mapping has yet to confirm those observations.

4 Conclusion

The goal of this paper was to apply clustering and classification techniques from the field of machine learning to problem of heroin chemical profiling. Following a principal component analysis, where two features were kept, a clustering algorithm was applied and two main classes were found in the data. Subsequently, four classification algorithms were applied, and the probabilistic neural network and the k-nearest neighbors method outperformed the multi-layer perceptron and the RBF network for this problem.

Future topics of investigation are extremely wide. First, introducing a time component in the whole process could reveal valuable information. To this purpose, recurrent neural networks could be used, which means that MLPs may not have given their last word in this context. Also, the use of support vector machines and other kernel methods if other variables are added could be useful, especially in the context of novelty detection applied to drug profiling.

A central problem which obviously needs to be investigated is the clustering one. Indeed, it has been supposed that the two main clusters corresponded to different sources. Firstly, this hypothesis needs practical reinforcement. Secondly, other clustering methods could lead to different hypotheses, especially multi-objective clustering, which has been popularized in the last years, and spectral clustering, which could help discover more complex structures in the data.

References

[1] S. Chainey, J. Ratcliffe. *GIS and Crime Mapping*, Wiley, Chichester, 2005.

[2] O. Guéniat, P. Esseiva. *Le Profilage de l'Héroïne et de la Cocaïne*, Presses polytechniques et universitaires romandes, Lausanne, 2005.

[3] P. Esseiva, L. Dujourdy, F. Anglada, F. Taroni, P. Margot, A methodology for illicit drug intelligence perspective using large databases, *Forensic Science International*, 132:139-152, 2003.

[4] P. Esseiva, F. Anglada, L. Dujourdy, F. Taroni, P. Margot, E. Du Pasquier, M. Dawson, C. Roux, P. Doble, Chemical profiling and classification of illicit heroin by principal component analysis, calculation of inter sample correlation and artificial neural networks, *Talanta*, 67:360-367, 2005.

[5] M.G. Madden, A.G. Ryder. Machine Learning Methods for Quantitative Analysis of Raman Spectroscopy Data, In Proceedings of the *International Society for Optical Engineering* (SPIE 2002), Vol. 4876, pp 1130-1139, 2002.

[6] M.L. O'Connell, T. Howley, A.G. Ryder, M.G. Madden. Classification of a target analyte in solid mixtures using principal component analysis, support vector machines, and Raman spectroscopy, In Proceedings of the *International Society for Optical Engineering* (SPIE 2005), Vol. 5826, pp 340-350, 2005.

[7] D.G. Stork, E. Yom-Tov. *Computer Manual in MATLAB to accompany Pattern Classification*, Wiley, Hoboken (NJ), 2004.

[8] R.O. Duda, P.E. Hart, D.G. Stork. *Pattern Classification, 2nd Edition*, Wiley, New York, 2001.