Available online at www.sciencedirect.com
ScienceDirect
Chemometrics and Intelligent Laboratory Systems 90 (2008) 123–131
Chemometrics and intelligent laboratory systems
www.elsevier.com/locate/chemolab

Advanced clustering methods for mining chemical databases in forensic science ☆

Frédéric Ratle [a,*], Christian Gagné [b], Anne-Laure Terrettaz-Zufferey [c], Mikhail Kanevski [a], Pierre Esseiva [c], Olivier Ribaux [c]

[a] Institute of Geomatics and Risk Analysis-Faculty of Earth and Environmental Sciences-University of Lausanne, Amphipôle, CH-1015, Switzerland
[b] Information Systems Institute-HEC-University of Lausanne, Internef, CH-1015, Switzerland
[c] School of Criminal Sciences- Faculty of Law- University of Lausanne, Batochime, CH-1015, Switzerland

Received 27 March 2007; received in revised form 3 September 2007; accepted 3 September 2007
Available online 11 September 2007

Abstract

Heroin and cocaine gas chromatography data are analyzed using several clustering techniques. A database with clusters confirmed by police investigation is used to assess the potential of the analysis of the chemical signature of these drugs in the investigation process. Results are compared to standard methods in the field of chemical drug profiling and show that conventional approaches miss the inherent structure in the data, which is highlighted by methods such as spectral clustering and its variants. Also, an approach based on genetic programming is presented in order to tune the affinity matrix of the spectral clustering algorithm. Results indicate that all algorithms show a quite different behavior on the two datasets, but in both cases, the data exhibits a level of clustering, since there is at least one type of clustering algorithm that performs significantly better than chance. This confirms the relevancy of using chemical drugs databases in the process of understanding the illicit drugs market, as information regarding drug trafficking networks can likely be extracted from the chemical composition of drugs.

Keywords: Forensic science; Machine learning; Pattern analysis; Spectral clustering; Kernel methods; Gas chromatography

1. Introduction

While modern spectroscopy and chromatography provide experimental tools that allow collecting large amounts of data related to forensic science, such as illicit drugs samples composition, machine learning and pattern analysis are now a matter of excitement in the forensic science community, in order for experts to analyze and understand the collected data. Indeed, classical data analysis methods often fail in this context given the high number of variables, the noise (coming from both the phenomenon itself and the experimental analysis) that corrupts the data, and the potentially nonlinear relationships between the different variables.

This work places itself at the border of chemometrics, machine learning and forensic science in order to highlight possibly useful patterns in the chemical composition of illicit drug seizures that may guide the investigation process. Also, since a database with drug samples corresponding to known investigations is available, it is possible to determine if geometrical structures that correspond to *real* production or distribution clusters exist in the space of the input variables (i.e., chemical constituents). Finally, since drug profiling is usually done by using samples intercorrelation measurement, this data will allow us to evaluate this method and to compare it with modern clustering techniques.

Preliminary studies were made by the same authors in [1] and [2], where heroin and cocaine data were studied using conventional machine learning approaches. First, Principal Component Analysis (PCA), *k*-means clustering and classification algorithms (MLP, PNN, RBF networks and *k*-nearest

☆ This work is supported by the Swiss National Science Foundation (grant no.105211-107862). The second author gratefully acknowledges postdoctoral fellowships from the ERCIM-SARIT (Europe), the Swiss National Science Foundation, and the FQRNT (Québec).
* Corresponding author.
E-mail address: frederic.ratle@unil.ch (F. Ratle).
URL: http://www.unil.ch/igar (F. Ratle).

0169-7439/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2007.09.001

neighbors) were applied. Also, cocaine data was studied with nonlinear feature extraction techniques such as kernel PCA [3], isomap [4] and locally linear embedding [5]. Kernel PCA shown to be an efficient and robust method for dimensionality reduction in this context.

A comprehensive review of the field of chemical drug profiling can be found in Guéniat and Esseiva [6]. In this book, authors have tested several statistical methods for heroin and cocaine profiling. Among other methods, they have mainly used similarity measures between samples to determine the main data classes. A methodology based on the cosine function as an intercorrelation measurement is explained in further details in Esseiva et al. [7]. Two drug samples are considered as being linked if their correlation is smaller than a given threshold. Also, PCA and Soft Independent Modelling of Class Analogies (SIMCA) have been applied for dimensionality reduction and supervised classification by these authors. A radial basis function neural network has been trained on the processed data and showed good results. The classes used for classification were based solely on statistical correlations in the chemical composition of the different samples. The profiling methodology was further developed in [8] for heroin and [9] for cocaine.

Madden and Ryder [10] have studied similar data: Raman spectra obtained from solid mixtures containing cocaine. The goal was to predict, based on the Raman spectrum, the cocaine concentration in a solid using k-nearest neighbors (KNN), 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 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 chromatography). Similarly, Raman spectroscopy data was studied in [11] using support vector machines with Gaussian 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 Gaussian kernel SVM outperformed all the other algorithms on a dataset of 217 samples using 22-fold cross-validation.

2. Method

Spectral clustering and kernel principal component analysis (kernel PCA) are two classes of machine learning algorithms based on the eigenvalue decomposition of a problem-dependent similarity (or dissimilarity) matrix. These methods can both be cast in the general framework of kernel methods. Readers unfamiliar with kernel methods can find an excellent review of this field in [12]. Kernel methods allow to apply mathematically sound linear methods for data analysis to nonlinear datasets, by implicitly projecting the input data in a high-dimensional Hilbert space, called the *feature space*, induced by some distance measure between data points.

2.1. Spectral clustering

Classical clustering algorithms, such as k-means, usually search for ball-shaped clusters by minimizing criteria such as intra-cluster variance. K-means can be summarized as follows, where i is an index over the whole dataset:

1. Initialize randomly K cluster centers m_k.
2. Compute the cluster assignment vector $C(i) = \text{argmin}_{1\leq k\leq K} ||x_i - m_k||$.
3. Compute the new cluster centers $m_k^{'} = \frac{1}{N_k}\Sigma_{x_i\in c_k} x_i$, where N_k is the number of points included in c_k, the k^{th} cluster.
4. Repeat 2 and 3 until the cluster assignment vector does not change.

These methods cannot perform well on arbitrary-shaped clusters. Spectral clustering aims at finding clusters that exhibit a specific geometry, albeit not easily found by maximizing cluster compactness. Very often, the mathematical objects formed by the data points mostly lie in a space of inferior dimensionality than that of the input space. For instance, Fig. 1 illustrates clearly the usefulness of such a method. The spirals dataset is two-dimensional, but has an intrinsic dimensionality of one. Rather than working directly with data points, spectral clustering uses an *affinity* matrix, which is a possibly nonlinear measure of similarity between points.

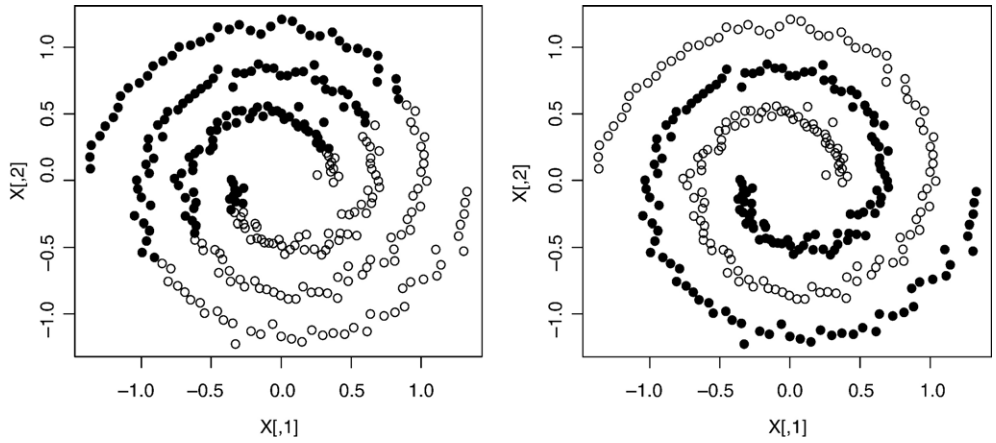

Fig. 1. Clustering the spirals dataset with k-means (left) and spectral clustering (right). The example is taken from [13].