A Comparison of One-Class Classifiers for Novelty Detection in Forensic Case Data*

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

[1] Institute of Geomatics and Risk Analysis, Faculty of Earth and Environmental Sciences, University of Lausanne, CH-1015, Switzerland
frederic.ratle@unil.ch

[2] School of Criminal Sciences, Faculty of Law, University of Lausanne, CH-1015, Switzerland

Abstract. This paper investigates the application of novelty detection techniques to the problem of drug profiling in forensic science. Numerous one-class classifiers are tried out, from the simple k-means to the more elaborate Support Vector Data Description algorithm. The target application is the classification of illicit drugs samples as part of an existing trafficking network or as a new cluster. A unique chemical database of heroin and cocaine seizures is available and allows assessing the methods. Evaluation is done using the area under the ROC curve of the classifiers. Gaussian mixture models and the SVDD method are trained both with and without outlier examples, and it is found that providing outliers during training improves in some cases the classification performance. Finally, combination schemes of classifiers are also tried out. Results highlight methods that may guide the profiling methodology used in forensic analysis.

1 Introduction

Analytical techniques such as gas chromatography are becoming widespread in forensic science in order to find underlying patterns in crime-related data, especially in the analysis of illicit drugs composition. Indeed, it has become largely accepted that the chemical signature of drug samples can provide information about the origin or the distribution network of the products and producers. An important issue that arises in this application is, given a set of chemical samples which can be related to known criminal investigations, how can one characterize this dataset in order to determine if a new sample can be linked to a known data cluster. If it cannot, it could be part of a "new" cluster. To this end, one-class classification is a novel and efficient way of approaching this problem.

In this paper, we perform a comparison of several popular one-class classifiers to the problem of drug profiling. The aim is to determine the most promising

* This work is supported by the Swiss National Science Foundation (grant no.105211-107862).

methods for this application, and to find potential strengths and weaknesses of the novelty detectors. A remarkable characteristic of the datasets is that class labels corresponding to links confirmed by investigators are available and allow a real evaluation of the performance of the methods and of the relevance of the chemical composition of drugs in order to classify samples.

2 Related work

Introductory work on chemical drug profiling in forensic science can be found in [1] and [2]. In these papers, no "true" class labeling is available; only chemical similarities are used as class membership criteria. Nonetheless, a profiling method based on samples correlation is devised. Several distance measures are used, and results are good when considering only chemical links as class criteria.

The datasets used here have been previously studied by the present authors in [3] and [4] using nonlinear dimensionality reduction techniques and various classification algorithms. In [5], authors apply the SVDD algorithm to novelty detection in mass spectra data. However, since no class labels are available, the performance of SVDD is assessed using a comparison with a clustering method.

3 Novelty detection

Novelty detection, also called one-class classification, is usually defined as the task of detecting a signal or pattern that a learning system is not aware of during training. Broad reviews of the subject can be found in [6], [7] and [8]. Even though the problem of outlier detection is a classical one in statistics, one-class classifiers have only been popularized recently. Most statistical approaches, such as Mahalanobis distance or extreme value theory, rely on strong assumptions, which are not always respected when dealing with small and noisy datasets. Many machine learning approaches, apart from density-based methods, go round these assumptions by trying to model the support of the data rather than its whole distribution. As suggested by Tax [9], one-class classifiers usually fall into one of these categories: density estimation methods, boundary methods and reconstruction methods.

3.1 Density estimation methods

Density estimation methods aim at estimating the whole distribution of the target data. A rejection threshold is then fixed so that points located in the far tails of the distribution are rejected. We shortly describe here the three density-based methods used in this study.

Gaussian distribution Here, a single Gaussian distribution is fitted to the target data. The mean and covariance matrix is estimated from the data, and the points comprised in the two tails are considered outliers. The rejection threshold is set such that 5% of the target data is rejected.

Parzen density estimation The Parzen density estimator is a mixture of kernels - typically Gaussian - with each of them centered on one training point. It can be expressed simply as

$$p(\mathbf{x}) = \sum_{i=1}^{N} K\left(\frac{\mathbf{x} - \mathbf{x}_i}{h}\right) \tag{1}$$

N is the size of the target training set and K is the kernel. One parameter, the width h (smoothing parameter), has to be tuned. Again, the rejection threshold is set such that 5% of the target data is rejected.

Gaussian mixture models Gaussian mixture models (GMM) are used to characterize the target data distribution by using a linear combinations of Gaussian distributions. Generally speaking, the likelihood of a mixture of Gaussians can be expressed as

$$p(\mathbf{x}) = \sum_{i=1}^{K} \pi_k N(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \tag{2}$$

The π_k's are scalar weights. Unlike Parzen density estimation, the number of Gaussians is specified and usually much smaller than the size of the training set. The means μ_k and covariances $\mathbf{\Sigma}_k$ are estimated by maximum likelihood using the expectation-maximization (EM) algorithm.

Two variants of GMM are tested: Gaussians with a diagonal covariance matrix and with a full covariance matrix. In the former case, elliptic clusters are assumed, while the latter case can take into account arbitrary-shaped clusters.

3.2 Boundary methods

Boundary methods, rather than estimating the distribution, aim at constructing a boundary - such as a sphere - around the target data. Points that fall outside the limits of the boundary are rejected. Here, k-nearest neighbors and SVDD are used.

K-nearest neighbors KNN first calculates the distances of the test point to its k neighbors, and averages these distances in order to have a single measure. It then computes the distances from these k neighbors to *their* k-nearest neighbors. Based on these distances, the local density of each point is computed, and the new point is rejected if its local density is inferior to that of its neighbors in the training set.

Support vector data description Support vector data description (SVDD), introduced in [10], is a method for characterizing the target data distribution without explicitely estimating the distribution parameters. It works by fitting

the smallest possible hypersphere around the target data in the feature space induced by a specified kernel, typically a Gaussian kernel. Data points that fall outside the hypersphere when projected in the feature space are rejected. This method has many similarities with the support vector method for novelty detection presented in [11]. However, in the latter work, an optimal hyperplane is built between target and outlier data, while SVDD builds a hypersphere.

3.3 Reconstruction methods

The goal of reconstruction methods is to develop a simplified representation of the data via clusters or principal components. These methods are numerous: k-means, principal components analysis, self-organizing maps, etc. Only k-means has been chosen among the reconstruction methods.

K-means In order to perform k-means clustering, the number of clusters in the target data has to be specified. Following this, boundaries are constructed around each cluster such that a certain fraction (5% here) of the target data is rejected. Again, points that fall outside the boundaries are considered outliers. It can be supposed that this type of method will work best for clusters that are well-separated.

3.4 Combination of classifiers

Ensemble methods have become increasingly popular when dealing with noisy real-world problems. This is also true for the problem of one-class classification, for which combination schemes have been proposed [12]. In this paper, we test two approaches: average and product of the posterior probabilities of the classifiers. These probabilities are either directly obtained when using a density-based method, or estimated when using reconstruction or boundary methods.

3.5 ROC analysis

A very useful assessment tool in classification and novelty detection tasks is the well-known receiver operating characteristic (ROC) curve. This curve represents the true positives (targets accepted as such) plotted against the false positives (outliers accepted as target), when varying the acceptation threshold. The area under the ROC curve (AUC) is thus a good measure of the classification performance. A random guess classifier is expected to have an AUC of 0.5, if the number of samples is large enough, while a classifier achieving a perfect separation will have an AUC of 1. Consequently, the AUC criterion must be maximized in order to obtain a good separation between targets and outliers.

4 Datasets and methodology

4.1 The data

Many types of substances can be found in a drug sample, and each of these can possibly provide information about a certain stage of drug processing. The interested reader may find a thorough description of this processing in [13]. This study focusses on the major chemical constituents, measured using GC/FID (gas chromatography and flame ionization detector). Details regarding the experimental procedure can be found in [1]. Each sample is characterized by features corresponding to the proportion of each chemical it contains. The first dataset (heroin) has 7 features, while the second (cocaine) has 13 features. The proportions of the chemical constituents have been estimated for each sample by using the area under the peaks in its chromatogram, after removal of the background noise. Figure 1 shows a typical chromatogram of a heroin sample.

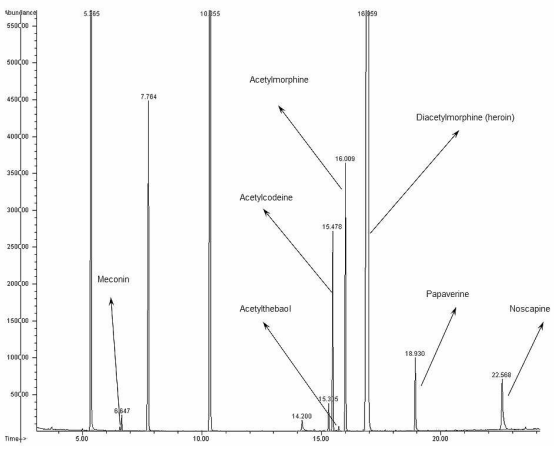

Fig. 1. An example of a chromatogram for a heroin sample. Each feature corresponds to the proportion of a constituent, estimated by the area under its corresponding peak.

Fig. 2 shows the labeled datasets projected on their two first principal components, in order to give an indication of the type of clusters that might be encountered. These figures show that the classes vary in shape and exhibit different scales. This could be expected, since the class labeling corresponds to networks of people involved in trafficking, while the input data corresponds to chemical constituents. It is thus of no surprise that the correlation between chemical profiles may not always match the links found by investigation, since two persons linked within a network do not necessarily share identical products from a chemical perspective.

The data consist of 323 heroin samples (with originally 3 classes) and 310 cocaine samples (10 classes). Each class corresponding to a distinct case (regardless of the chemical content of the samples), we have drawn out and tagged as

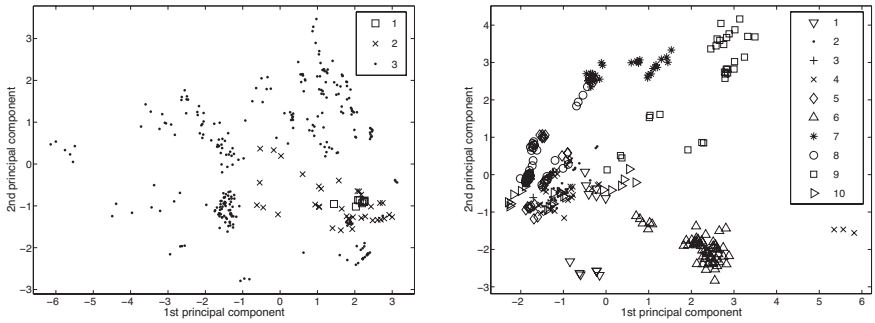

Fig. 2. Projection on the two first principal components of heroin (left) and cocaine (right).

outliers one case from the heroin dataset and two cases from the cocaine dataset, the number of classes being superior for the latter. Each dataset thus contains 1 target class (containing the remaining original classes) and 1 outlier class (the drawn out cases). The rationale for this is that we want to classify a sample as being linked to a known network or not.

4.2 Experimental setup

All experiments have been performed in Matlab. The Data Description toolbox (DDTools) [14] has been used. When necessary, parameter values (σ or k) were assigned using line search and a k-fold cross-validation scheme. The main difference with normal k-fold cross-validation is that, outliers being available, the outlier set was also split into k folds, but not used for training. The procedure can be summarized as follows:

1: **for** $i = 1$ to k **do**
2: Remove partition p_i^T of target dataset T to obtain T'.
3: Remove partition p_i^O of outlier dataset O.
4: Train the one-class classifier on T'.
5: Compute the AUC a_i for dataset $p_i^T \cup p_i^O$.
6: **end for**
7: Compute the cross-validation error $e = \frac{1}{K} \sum_{j=1}^{K} a_i$.

As the datasets are rather small given the number of variables, using only half of the training targets to test the classifiers may not allow a good characterization of the target data. Two measures are thus given:

1. AUC on training target data and independent test outliers (called training AUC or simply AUC below).
2. AUC by the k-fold cross-validation method previously described (using 5 folds, and called AUC-CV).

The first method obviously overestimates the AUC, while the second might be both pessimistic or optimistic. However, our prime goal here is to compare the methods one against another.

5 Results and discussion

Tables 1 and 2 show the obtained results, which are averaged over 10 runs.

	Heroin			Cocaine		
	AUC	AUC-CV	param.	AUC	AUC-CV	param.
K-means	64.6 ± 0.0	51.5 ± 4.5	$k = 2$	95.4 ± 5.3	84.9 ± 4.1	$k = 8$
KNN	97.7 ± 0.0	62.2 ± 0.0	$k = 2$	96.6 ± 0.0	87.7 ± 0.0	$k = 9$
Gauss	64.0 ± 0.0	45.7 ± 0.0	-	98.4 ± 0.0	90.8 ± 0	-
Parzen	92.1 ± 0.0	58.7 ± 0.0	$h = 0.5$	94.8 ± 0.0	89.0 ± 0.0	$h = 1.5$
GMM I	55.3 ± 0.2	41.5 ± 8.7	2 clusters	87.7 ± 0.0	86.4 ± 2.2	8 clusters
GMM II	84.7 ± 3.9	62.5 ± 7.1	2 clusters	98.1 ± 1.5	83.2 ± 2.1	8 clusters
SVDD	59.6 ± 12.0	66.4 ± 7.8	$\sigma = 2$	88.6 ± 3.6	90.2 ± 1.1	$\sigma = 3$
all-mean	91.0 ± 0.8	60.0 ± 1.2	-	98.2 ± 1.2	87.2 ± 2.2	-
all-product	91.1 ± 0.7	60.1 ± 1.3	-	99.1 ± 0.8	87.2 ± 2.2	-

Table 1. AUC for the heroin and cocaine dataset, without outliers in the training process. GMM I designates the Gaussian mixture model with a diagonal covariance matrix, and GMM II the model with the full matrix.

	Heroin			Cocaine		
	AUC	AUC-CV	param.	AUC	AUC-CV	param.
GMM I	70.9 ± 9.4	42.7 ± 8.3	2 clusters	98.3 ± 0.8	86.0 ± 3.0	8 clusters
GMM II	84.8 ± 3.0	64.8 ± 7.1	2 clusters	99.8 ± 0.1	84.7 ± 2.0	8 clusters
SVDD	78.4 ± 2.4	40.5 ± 4.1	$\sigma = 2$	96.2 ± 0.3	86.7 ± 1.5	$\sigma = 3$
all-mean	83.3 ± 4.6	60.2 ± 4.1	-	98.6 ± 0.4	86.0 ± 1.9	-
all-product	82.2 ± 2.1	57.2 ± 7.7	-	99.7 ± 0.2	85.4 ± 1.9	-

Table 2. AUC for the heroin and cocaine dataset, with outliers in the training process.

Figure 3 shows ROC curves on test outliers and training target data for the best and the worst classifier. Since the test AUC was estimated with cross-validation, the corresponding curves cannot be illustrated. It can be seen that the performances for the second dataset are located within a smaller interval.

Results show a surprising difference between the two datasets regarding the general performance of the methods. First, for the heroin dataset, the gap between AUC and AUC-CV is considerably larger than that of the cocaine dataset. KNN and Parzen perform best at achieving a good separation between target and outlier data, but they both provide an average performance on cross-validation.

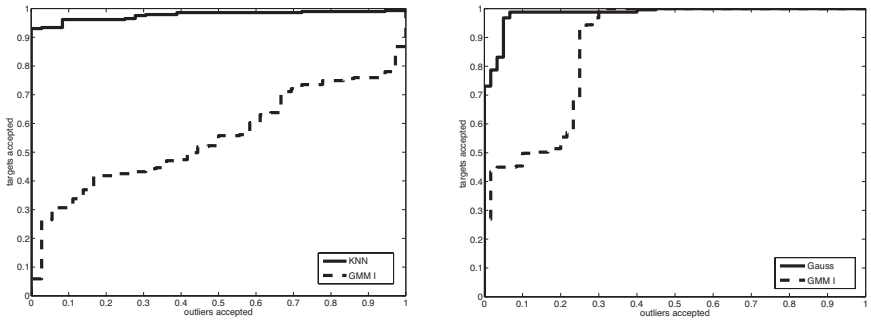

Fig. 3. ROC curves for the best and the worst classifiers obtained on training target data and independent outliers, for both datasets (heroin on the left and cocaine on the right). Since the test AUC was estimated using cross-validation, the corresponding ROC curves cannot be obtained. For the second dataset, the classification performance is comprised within a smaller interval.

The GMM with a full covariance matrix has produced above average results for both AUC and AUC-CV, while the GMM with a diagonal matrix has performed poorly. The SVDD method, even though providing a poor training AUC, seems to be by far the most robust. The AUC has not decreased at all between the training AUC and AUC-CV. In fact, the performance improved, but this can probably be explained by an "optimistic" partitioning during cross-validation. K-means and the Gaussian distribution performed below average. The latter even performed worse than random guessing on AUC-CV, as did GMM I.

On the cocaine dataset, most methods are more consistent. Indeed, the AUC-CV is much more closer to the training AUC for all the methods. All algorithms performed reasonably well, although SVDD has again appeared slightly more robust. Some methods have improved significantly when applied to this dataset. While the Gaussian distribution was among the worst classifiers for heroin data, it outperforms all the other methods on cocaine data. Some of these observations are summarized in Table 3.

	Heroin dataset		Cocaine dataset	
	best	worst	best	worst
AUC	KNN	GMM I	Gauss	GMM I
AUC-CV	SVDD	GMM I	Gauss	GMM II
Computational cost	Gauss	SVDD	Gauss	SVDD
Robustness	SVDD	KNN	SVDD	GMM II
Easiness	Gauss	GMM I-II	Gauss	GMM I-II

Table 3. Comparison of the methods with respect to AUC, AUC-CV, computational cost, robustness and easiness. Easiness is defined as the number of parameters to tune (the smaller the better).

Results of classifier combinations are somewhat mitigated. The ensemble classifier performs well above the average of the base classifiers for both training and cross-validation AUC. However, the result reaches at best the performance obtained with the best one-class classifier. The product combination rule gives slightly superior results, but the difference is not significant. Considering the additional computational cost induced by using more than one classifier, combinations are not extremely interesting on these datasets if the base classifiers are already good. However, combinations might still be interesting when no knowledge of the methods' performance is known (i.e., with unlabeled data).

The addition of outliers in the training process significantly increased the training AUC of SVDD and GMM I and II on both datasets. However, when looking at the AUC-CV, there is no significant change in the performance of the one-class classifiers, given the standard deviations of the results. Oddly enough, the CV performance of SVDD decreases when outliers are presented during training. This, however, is likely a particularity of this specific dataset. The same remarks can be made regarding combination of classifiers. At best, the AUC reaches that of the best classifier.

From these results, it can be inferred that the structure of cocaine data is close to well-separated Gaussian-like clusters. Indeed, the simple Gaussian distribution performed very well, and the prediction performance of all the methods is in general very high. The class separation in heroin data seems to be quite more complicated. All methods, whilst sometimes achieving a good separation (Parzen, KNN, GMM II), have a poor prediction performance. In both cases, the SVDD method has shown to be the most robust. Most importantly, given these results, it is reasonable - at least for cocaine data - to suppose that information regarding the network from which comes a sample might be extracted on the basis of its chemical composition.

6 Conclusion

Several one-class classifiers have been applied and assessed using the AUC criterion for novelty detection in chemical databases of illicit drug seizures. The two datasets have proven very different: far better prediction performance has been obtained with the cocaine dataset, as it could be seen with cross-validation. In most cases, the SVDD method has appeared more robust, even though other methods have outperformed it in some cases. No significant difference was noted between general types of outlier detectors, i.e., density-based, boundary or reconstruction methods. Combinations of classifiers provided better than average results, but at best a similar performance as the best classifier. In addition, providing outliers during training improved the training AUC, but did not change significantly the cross-validation AUC.

In general, results suggest, especially for cocaine, that information regarding the origin of a sample (more precisely, the distribution network) might be extracted from its chemical constituents. This is a very interesting result, since nothing would indicate *a priori* that this is the case. Indeed, products circu-

lating in the same network could come from different producers. Overall, these results have highlighted one-class classification methods that could contribute to the profiling methodology in forensic analysis. Future research topics include considering the time variable. Chemical compositions might exhibit seasonality, and integrating time would likely provide different results.

7 Acknowledgements

Authors thank S. Ioset for the preparation of the database and B. Petreska for improving the manuscript.

References

1. P. Esseiva, L. Dujourdy, F. Anglada, F. Taroni, P. Margot, A methodology for illicit heroin seizures comparison in a drug intelligence perspective using large databases, *Forensic Science International*, **132**:139-152, 2003.
2. 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.
3. F. Ratle, A.L. Terrettaz-Zufferey, M. Kanevski, P. Esseiva, O. Ribaux, Pattern analysis in illicit heroin seizures: a novel application of machine learning algorithms, *Proc. of the 14th European Symposium on Artificial Neural Networks*, d-side publi., 2006.
4. F. Ratle, A.L. Terrettaz-Zufferey, M. Kanevski, P. Esseiva, O. Ribaux, Learning manifolds in forensic data, *Proc. of the 16th Int. Conf. on Artificial Neural Networks*, Springer, 2006.
5. C. Tong and V. Svetnik, Novelty detection in mass spectral data using a support vector machine method, Proc. of *Interface 2002*.
6. M. Markou and S. Singh, Novelty detection: a review - part 1: statistical approaches, *Signal Processing*, **83**: 2481-2497, 2003.
7. M. Markou and S. Singh, Novelty detection: a review - part 2: neural network based approaches, *Signal Processing*, **83**: 2499-2521, 2003.
8. S. Marsland, Novelty detection in learning systems, *Neural Computing Surveys*, **3**: 157-195, 2003.
9. D.M.J. Tax, One-class classification, Ph.D. thesis, University of Amsterdam, 2001.
10. D.M.J. Tax and R.P.W. Duin, Support Vector Data Description, *Machine Learning*, **54**: 45-66, 2004.
11. B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, J. Platt, Support vector method for novelty detection, *Advances in Neural Information Processing Systems 12*, 2000.
12. D.M.J. Tax, Combining one-class classifiers, Proc. of *Multiple Classifier Systems*, LNCS 2096, Springer, 2001.
13. O. Guéniat, P. Esseiva, *Le Profilage de l'Héroïne et de la Cocaïne*, Presses polytechniques et universitaires romandes, Lausanne, 2005.
14. D.M.J. Tax, DDtools, the Data Description Toolbox for Matlab, version 1.5.7, 2007.