On the Robustness of Kernel-Based Clustering

Fabio A. González[1], David Bermeo[1], Laura Ramos[1], and Olfa Nasraoui[2]

[1] BioIngenium Research Group, Universidad Nacional de Colombia, Bogotá
[2] Knowledge Discovery & Web Mining Lab, The University of Louisville
{fagonzalezo,jdbermeol,lmramoss}@unal.edu.co,
olfa.nasraoui@louisville.edu

Abstract. This paper evaluates the robustness of two types of unsupervised learning methods, which work in feature spaces induced by a kernel function, kernel *k*-means and kernel symmetric non-negative matrix factorization. The main hypothesis is that the use of non-linear kernels makes these clustering algorithms more robust to noise and outliers. The hypothesis is corroborated by applying kernel and non-kernel versions of the algorithms to data with different degrees of contamination with noisy data. The results show that the kernel versions of the clustering algorithms are indeed more robust, i.e. producing estimates with lower bias in the presence of noise.

1 Introduction

The presence of noise and outliers are a main source of data quality problems in statistics, data mining and machine learning. Many factors can cause noise and outliers, including errors in measurements, data entry and communication, as well as simple deviation of some data samples. These problems along with assumptions made about data are known to lead to incorrect and biased results. There has been a good amount of work devoted to dealing with noise and outliers. In some approaches, outliers are eliminated from the data as part of the data preprocessing stage. In other approaches, learning and inference algorithms are designed in such a way that they can resist noise and outliers; these algorithms are described as *robust*. *Robust statistics* is a field of statistics that deals with the development of techniques and theories for estimating the model parameters while dealing with deviations from idealized assumptions [7,10].

Data clustering is one of the most important data analysis tools with many applications and a great variety of algorithms. As with other learning algorithms, outliers in the data can result in bad parameter estimation, thus generating bad clusterings. The objective of this paper is to evaluate the robustness of an important type of clustering algorithms, namely those that are based on kernel methods. In particular, we evaluate how clustering algorithms behave when the input data is contaminated with increasing amounts of noisy data. The evaluated algorithms are the conventional versions and kernelized versions of *k*-means and Symmetric Non-negative Matrix Factorization . Our main hypothesis is that the use of a non-linear kernel may improve the robustness of the algorithm. The preliminary experimental results in this paper confirm this hypothesis, suggesting that kernel-based methods are a viable alternative for performing robust clustering.

L. Alvarez et al. (Eds.): CIARP 2012, LNCS 7441, pp. 122–129, 2012.

The rest of the paper is organized as follows: Section 2 discusses the problem of clustering and how it can be modeled as a matrix factorization problem; Section 3 introduces robust statistics and robust clustering; Section 4 presents the experimental evaluations; and finally, Sections 5 presents our conclusions and future work.

2 Clustering and Matrix Factorization

2.1 Clustering

Clustering is the most important problem in unsupervised learning. In general, the goal of a clustering algorithm is to find groups (called *clusters*) in a set of data samples., such that the clusters are homogeneous, i.e. contain similar data samples, while data samples from different clusters are different. Depending on the type of clustering algorithm, this goal could be accomplished in different ways. In this work we focus on a particular type of clustering algorithms which are based on the optimization of an objective function.

One of the most popular clustering algorithms is k-means., which represents the clusters by a set of centroids $M = \{m_1, \ldots m_k\}$ that minimize the following objective function:

$$SSE = \min_{M} \sum_{i=1}^{k} \sum_{x \in C_i} (x - m_i)^2 \tag{1}$$

where $\{C_1, \ldots, C_k\}$ is a disjunct partition of the input data set X, such that $X = \bigcup_{i=1}^{k} C_i$. The minimization is accomplished by an optimization process that iteratively reassigns data points to clusters, thus refining the centroid estimations.

2.2 Non-negative Matrix Factorization

The general problem of matrix factorization is to decompose a matrix X into two factor matrices A and B :

$$X_{n \times l} = A_{n \times k} B_{k \times l} \tag{2}$$

This could be accomplished by different methods including: Singular Value Decomposition (SVD), Non-negative Matrix Factorization (NMF), and Probabilistic Latent Semantic Analysis, among others. The factorization problem can be formulated as an optimization problem:

$$\min \quad _{A,B} d(X, AB) \tag{3}$$

where $d(,)$ is a distance or divergence function and the problem could have different types of restrictions. For instance, if $d(,)$ is the Euclidean Distance and there are no restrictions, the problem is solved by finding the SVD; if X, A and B are restricted to be positive, then the problem is solved by NMF.

This type of factorization may be used to perform clustering. The input data points are the columns of X (l n-dimensional data points). The columns of A correspond to the coordinates of the centroids. The columns of B indicate to which cluster each sample belongs, specifically if x_j belongs to C_i, then $B_{i,j} = 1$, otherwise $B_{i,j} = 0$. With this interpretation, the objective function in (1) is equivalent to the objective function in (3)

using the Euclidean distance. An important advantage of this approach is that values in the matrix B are not required to be binary, in fact, they can take continuous values. These values can be interpreted as soft membership values of data samples to clusters, i.e., NMF can produce a soft clustering of the input data [3].

NMF has been shown to be equivalent to other unsupervised learning methods such as probabilistic latent semantic analysis and kernel k-means. Also, there are different versions of NMF which impose new restrictions or weaken some of its restrictions. For instance, Semi-NMF allows negative values in matrices X and A in (3), Convex-NMF imposes that A must be a combination of the data input, $A = XW$ [3]. In this work, we use a particular version of NMF, Symmetric-NMF (SNMF) which produces the following factorization:

$$(X^T_{lxn} X_{nxl}) = H_{lxk} H^T_{kxl} \tag{4}$$

An important characteristic of this version of NMF is that it is amenable to be used as a kernel method. This is discussed in the next subsection.

2.3 Kernel Methods

In contrast with traditional learning techniques, kernel methods do not need a vectorial representation of the data. Instead, they use a kernel function that allows kernel methods to be naturally applied to unstructured or complex structured data such as text, strings, trees and images [12].

Informally, a kernel function measures the similarity of two objects. Formally, a kernel function, $k : X \times X \to \mathbb{R}$, maps pairs (x, y) of objects in a set X, the problem space, to the space of real numbers. A kernel function implicitly generates a map, $\Phi : X \to F$, where F corresponds to a Hilbert space called the feature space. The dot product in F is calculated by k, specifically $k(x, y) =< \Phi(x), \Phi(y) >_F$. Given an appropriate kernel function, complex patterns in the problem space may correspond to simpler patterns in the feature space. For instance, non-linear boundaries in the problem space may be transformed to linear boundaries in the feature space.

Both k-means and SNMF have kernelized versions, which receive as input a kernel matrix instead of a set of data samples represented by feature vectors. The kernel version of k-means is called, unsurprisingly, kernel k-means (KKM). In the case of SNMF, the kernelized version works as follows [3].

SNMF starts with an initial estimation of the factor matrix H and iteratively updates it using the updating equation:

$$H_{i,k} = H_{i,k}(1 - \beta + \beta \frac{((X^T X)H)_{i,k}}{(HH^T H)_{i,k}})$$

The kernel version of the algorithm is obtained by using a kernel matrix K instead of the expression $(X^T X)$, where K is an $l \times l$ matrix with $K_{i,j} = k(x_i, x_j)$. There are different types of kernels, some of them general and some of them specifically defined for different types of data. The most popular general kernels are the linear kernel

$$k(x, y) =< x, y >, \tag{5}$$

the polynomial kernel

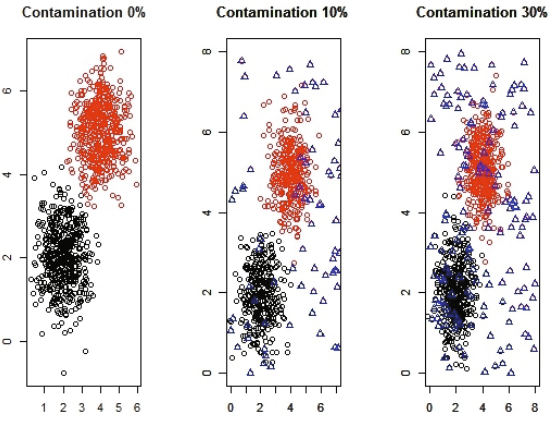

Fig. 1. Example of a data set with two clusters with different degrees of contamination

$$k(x,y) = p(< x,y >),$$

where $p(\,)$ is a polynomial with positive coefficients, and the Gaussian (or RBF) kernel

$$k(x,y) = e^{\frac{\|x-y\|^2}{2\sigma^2}}. \tag{6}$$

The cluster centroids estimated by the kernel versions of both algorithms are in the feature space, resulting from transforming the data via the mapping function Φ, and correspond to the points $C_j = \frac{1}{n}\sum_{x_i \in C_j} \Phi(x_i)$. However, we are interested in the pre-image in the original space of the centroids, i.e., points $\hat{C}_j$ such that $\Phi(\hat{C}_j) = C_j$. However, it is possible that an exact pre-image may not even exist, so we look for the $\hat{C}_j$ that minimizes the following objective function: $\min_{\hat{C}_j} \left\|\hat{C}_j - C_j\right\|^2$. According to Kwok et al. [9], the optimum C_j can be found by iterating the following fixed-point formula:

$$\hat{C}_j^{t+1} = \frac{\sum_{x_i \in C_j} \exp(\frac{-||\hat{C}_j^t - x_i||)}{s})x_i}{\sum_{x_i \in C_j} \exp(\frac{-||\hat{C}_j^t - x_i||}{s})} \tag{7}$$

3 Robust Clustering

Robust statistics is an area of statistics that deals with the problem of estimating the parameters of a parametric model while dealing with deviations from idealized assumptions [7,10]. These deviations could be caused by contamination of data by outliers, rounding and grouping errors, and departure from an assumed sample distribution. Outliers can cause classical estimators to heavily depart from the actual values of the parameters making them useless. Robust statistics provides theory and techniques to study and develop robust estimators.

In the particular case of clustering, outliers and noise are an important issue that could cause the incorrect detection of clusters, the detection of spurious clusters and a biased estimation of cluster parameters [5]. Figure 1 shows an example of a data set with two clusters and different degrees of contamination. It is important to notice that, in general, the source of contamination is not known, so it is difficult to make assumptions about the type of distribution or distribution parameters. A robust clustering algorithm is one that is able to deal with contaminated data in such a way that its estimations are not affected by the presence of contaminated data. There are different types clustering strategies, but in this work, we concentrate on partitional, prototype based methods such as k-means. In this case, clusters are characterized by centroids, which are in fact location estimators. A robust estimation of cluster centroids requires that the value of the estimation be affected as little as possible by the presence of contaminated data. Robust statistics provides different methods to measure the robustness of an estimator [8], such as the breakdown point, influence function and sensitivity curve. The goal is to measure how an estimator behaves with different proportions and values of contamination. In many cases, this analysis is performed assuming the worst conditions, such as using asymptotic analysis when the value of the contamination tends to infinity. But these extreme conditions are neither acceptable nor possible in most engineering applications [2]. In this work, and following [2], we assume a bounded contamination and perform a non-asymptotic analysis. The contamination model is given by

$$F = (1-\varepsilon)G + \varepsilon E, \tag{8}$$

where G is the distribution of real, uncontaminated samples, E is the distribution of contaminated data and F is the overall sample distribution. The effect of contamination in the data is measured by the bias of the estimation:

$$\text{bias}(\hat{\Theta}, X) = \left\| \Theta - \hat{\Theta}(X) \right\|, \tag{9}$$

where Θ are the real parameters of G, X is a sample generated using the model in Eq. (8), $\hat{\Theta}$ is a estimator function and $\|\cdot\|$ is an appropriate norm. In this work, $\Theta = (C_1, \ldots, C_k)$ with $C_i \in \mathbb{R}^n$, the centroid of the i-th cluster, and $\|\cdot\|$ is the Euclidean norm.

Different robust estimators, for both location and scale, have been proposed. In the case of clustering, the development of robust methods has been studied since the nineties. Some of the representative works include: trimmed k-means [1], robust fuzzy clustering methods [11,2] and minimum covariance determinant estimator [6]. Interestingly, there has not been, to our knowledge, any work that explored the robustness of kernel-based clustering methods. While some works proposed kernel-based clustering methods that could be considered robust, their focus was not to study the robustness gained by the use of particular kernels but rather the implementation of certain robust strategies, e.g. minimum volume covering ellipsoid estimation[4], using kernel methods. The present work studies how the use of appropriate kernels provides a conventional clustering method, such a k-means, with robust properties.

4 Experimental Evaluation

The goal of the experimental evaluation is to study the influence of a popular non-linear kernel, the Gaussian kernel, in two kernel-based clustering algorithms. Both algorithms

are applied to two data sets (one synthetic and one real) with different degrees of contamination. Two kernels are used: an identity kernel and a Gaussian kernel. The main hypothesis is that the Gaussian kernel makes the algorithms more resilient to contamination. The performance of the algorithms is evaluated by measuring the bias (Eq. (9)) and evaluating how it behaves when the amount of contamination is increased.

4.1 Datasets

Two datasets were used: a synthetic dataset and the real *iris flower* dataset. The synthetic dataset has 150 samples from a mixture of three multivariate Gaussians in $\mathbb{R}^4$, each Gaussian having a different mean and covariance matrix. The iris flower data set is a multivariate dataset that has been widely used in statistics and data mining. Each sample in the dataset corresponds to a flower characterized by four numerical features and a categorical attribute that indicates the species of the flower.

4.2 Experimental Setup

Both data sets are contaminated by adding a set of samples generated from a multivariate uniform distribution with a support that includes the range of the corresponding data set features. Different percentages of contamination are used: 0%, 5%, 10%, 15%, 20%, 25% and 30%. For instance, a data set with a ~30% of contamination is a dataset with 214 samples, 150 corresponding to the original samples and 64 corresponding to contamination. For each algorithm, two types of kernels are used: the identity kernel (Eq. (5)) and the Gaussian kernel (Eq. (6)). For the Gaussian kernel, different values of σ were tested following a logarithmic scale, $\sigma = 2^i$ for $i \in [-5, \ldots, 5]$. Each algorithm configuration was run 10 times and the average bias is reported. The bias is calculated using Eq. (9), where $\hat{\Theta}(X)$ correspond to the cluster centroids estimated by each algorithm and, in the case of the configurations with the Gaussian kernel, are back-projected to the original space using Eq. (7).

4.3 Results and Discussion

Figures 2 and 3 show the evolution of the bias when the percentage of contamination increases, for both methods and for different kernels. Figures 2(a) and 3(a) show the results for KKM with three different kernels: linear kernel (equivalent to conventional *k*-means), Gaussian kernel with $\sigma = 2^{-2}$, and Gaussian kernel with $\sigma = 2^4$. In the case of the Gaussian kernel, a systematic evaluation of different values for the parameter σ was performed and the best performing value ($\sigma = 2^{-2}$) and an average performing value ($\sigma = 2^4$) are reported. Figures 2(b) and 3(b) show the corresponding results for SNMF, with the linear kernel corresponding to the conventional (non-kernelized) version.

For both algorithms, the worst performance is exhibited by the linear kernel, and the best performance is accomplished by the Gaussian kernel with $\sigma = 2^{-2}$. These results confirm the hypothesis that the use of a Gaussian kernel makes both algorithms more robust and resilient to noise and outliers. Also, it is clear that the robustness that is induced by the Gaussian kernel depends on the parameter σ. This parameter is related to the scale of the data and the results indicate that an appropriate identification of the scale has an important effect on the method's robustness.

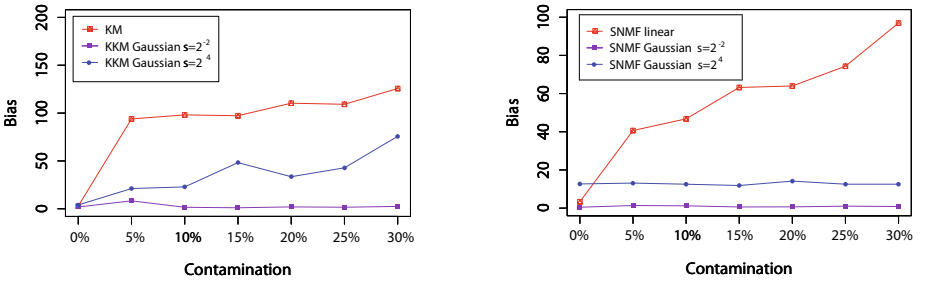

Fig. 2. Bias vs contamination in the Iris dataset using (left) K-Means (KM), Kernel K-Means (KKM) and (right) kernel Symmetric Non-negative Matrix Factorization (SNMF)

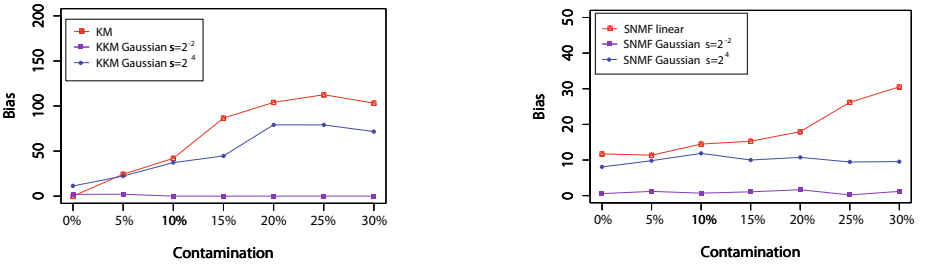

Fig. 3. Bias vs contamination in the synthetic dataset using (left) K-Means (KM), Kernel K-Means (KKM) and (right) kernel Symmetric Non-negative Matrix Factorization (SNMF)

The results also show that SNMF has a better performance, in terms of robustness, than KKM. A possible explanation is the fact that SNMF produces a soft clustering in contrast with the hard clustering produced by KKM. In fact, Davé et al. [2] found a connection between fuzzy membership functions and weight functions used in robust statistics.

5 Conclusions and Future Work

The main hypothesis of this work is that the robustness of kernel-based clustering methods is increased by the use of Gaussian kernels. The results of the exploratory experiments performed in this paper provide evidence to support this claim. Kernel methods are popular and well regarded machine learning methods thanks to their ability to learn complex non-linear models. This ability is due, in part, to the kernel trick, which allows finding non-linear patterns in the original problem space by learning linear patterns in a kernel-induced higher-dimensional space. The results of this study suggest an additional correspondence with useful applications: non-robust estimation in a Gaussian kernel-induced space corresponds to robust estimation in the original problem space.

The main question posed by our findings is why the use of the Gaussian kernel makes the corresponding algorithms more robust to noise and outliers. Our conjecture is that back-projection (Eq. 7) of a mean value from a Gaussian kernel-induced feature space can be considered as a robust W-estimator [10]. A deeper study of this conjecture is part of our future work.

Acknowledgements. This work was partially funded by the project "Sistema para la Recuperación de Imágenes Médicas utilizando Indexación Multimodal" number 110152128767. Part of this work was carried out while the first author was visiting the KD&WM Lab at the U of Louisville supported by the Fulbright Visiting Scholar Fellowship.

References

1. Cuesta-Albertos, J.A., Gordaliza, A., Matran, C.: Trimmed k-Means: An Attempt to Robustify Quantizers. The Annals of Statistics 25(2), 553–576 (1997)
2. Davé, R.N., Krishnapuram, R.: Robust clustering methods: a unified view. IEEE Transactions on Fuzzy Systems 5(2), 270–293 (1997)
3. Ding, C., Li, T., Jordan, M.I.: Convex and Semi-Nonnegative Matrix Factorizations. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(1), 45–55 (2010)
4. Dolia, A., Harris, C., Shawetaylor, J., Titterington, D.: Kernel ellipsoidal trimming. Computational Statistics & Data Analysis 52(1), 309–324 (2007)
5. García-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A.: A review of robust clustering methods. Advances in Data Analysis and Classification 4(2-3), 89–109 (2010)
6. Hardin, J., Rocke, D.M.: Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Computational Statistics & Data Analysis 44(4), 625–638 (2004)
7. Huber, P.J.: Robust Statistics. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken (1981)
8. Hubert, M., Rousseeuw, P.J., Van Aelst, S.: High-Breakdown Robust Multivariate Methods. Statistical Science 23(1), 92–119 (2008)
9. Kwok, J.T.Y., Tsang, I.W.H.: The pre-image problem in kernel methods, vol. 15, pp. 1517–1525. IEEE (2004)
10. Maronna, R.A., Martin, R.D., Yohai, V.J.: Robust statistics. Wiley (2006)
11. Nasraoui, O., Krishnapuram, R.: A robust estimator based on density and scale optimization and its application to clustering. In: Proceedings of the Fifth IEEE International Conference on Fuzzy Systems, vol. 2, pp. 1031–1035. IEEE (1996)
12. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)