Robust Estimation in Reproducing Kernel Hilbert

Joseph A. Gallego-Mejia ,[*]
Department of Computer Science
Universidad Nacional de Colombia
jagallegom@unal.edu.co

Fabio. Gonzalez. O
Department of Computer Science
National University Of Colombia
fagonzalezo@unal.edu.co

Abstract

Our work shows that estimating the mean in a feature space induced by certain kinds of kernels is the same as doing a robust mean estimation using an M-estimator in the original problem space. In particular, we show that calculating the average on a feature space induced by a Gaussian kernel is equivalent to perform robust mean estimation with the Welsch M-estimator. Besides, a new framework is proposed that was used to build four new robust kernels: Tukey's, Andrews', Huber's and Cauchy's robust kernels. The new robust kernels, combined with kernel matrix factorization clustering algorithm, were compared to state-of-the-art algorithms in clustering tasks. The result shows that some of the new robust kernels perform in a par with state-of-the-art algorithms.

1 Research Problem

The principal problem addressed by this work is to explore the connection between robust statistics and kernel methods. In particular, the robustness properties of location estimators calculated through certain kinds of kernels, which are shown to be equivalent to robust M-estimators. Also, we explore the impact that using this kind of kernels has on the performance of some kernel-based clustering techniques.

In this work, the first author found the new robust kernel and did the experimentation. The second author proposes the general framework and the relationship between robust Welsch M-estimator and Gaussian kernel.

2 Motivation

Robust statistics is a branch of statistics that deals with outliers and deviation from the assumptions. Several methods have been developed to mitigate the bias generated by those deviations. Some of them rely on a generalization of maximum likelihood estimation called M-estimatoion. The principal idea is to build functions that mitigate the influence of outliers without explicitly dropping them (6; 7; 8; 9).

A kernel function $K(x, y)$ may implicitly define an intrinsic high dimensional feature space without ever compute each coordinate in that space, this is popular known as the kernel trick (2). This ability is used by some methods, such as SVM, to learn linear models in the feature space that corresponds to non-linear models in the original space. We discover an analogous correspondence that connects conventional non-robust location estimation in the feature space with robust location estimation in the original space. This connection has important consequences on the performance of clustering algorithms which are based on certain types of kernels.

[*] {jagallegom, fagonzalezo}@unal.edu.co, MindLab Group

LatinX in 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

(5) investigated empirically that in clustering tasks, the Gaussian kernel has robustness against increased contamination when compared to the linear kernel. Motivated by these results, we found a formal proof that when a particular kernel is used, the mean estimation in the feature space is analogous to perform the mean estimation with a robust M-estimator in the data space.

This result allowed us to find the connection between robust M-estimators and particular properties of kernels. Based on this, we designed several new robust kernels.

There are four key parts to understand robust location estimation when a particular kernel is used: first, we have an initial data; second, a feature space is induced by a kernel function; third, the mean of the data is calculated in the feature space; finally, the projection of the mean in the feature space is back to the data space with an approximation function. We showed that using a certain kernels, the mean estimation in the feature space is equivalent to doing mean estimation with a robust m-estimator in the data space.

3 Technical Contribution

Our principal contributions in this work are the following:

- We proved that performing non-robust mean estimation in a a feature space induced by a Gaussian kernel is equivalent to doing robust mean estimation in the original space with the robust Welsch location M-estimator.

  Proposition 1. *Given a set of points $\{d_1, \ldots, d_n\} \subseteq \mathcal{X}$, the approximate pre-image of its centroid in a feature space, $\mathcal{F}$, induced by a Gaussian kernel, k, corresponds to the Welsch location M-estimator. In other words:*

  $$P_\Phi(\mu) = P_\Phi\left(\frac{1}{n}\sum_{i=1}^{n}\Phi(d_i)\right) = \arg\min_{y\in\mathcal{X}}\sum_{x_i\in S}\rho_{\text{welsch}}(\|x_i - y\|)$$

  where $P_\Phi(\mu)$ is the approximate pre-image.

- We generalized the previous result and found new robust Kernels based on Tukey's, Andrews', Cauchy's and Huber's robust M-estimators.

4 Experimental Results

Two kernel clustering methods were used: Convex nonnegative matrix factorization (CNMF) (4) and Kernel K-Means (KKM) (3). Besides, ten state-of-the-art algorithms were used including Nonnegative matrix factorization random walks (10), Left-stochastic matrix factorization (LSD) (1) and Robust Manifold NMF (RMNMF) (11). Thirteen data sets were used in the experiment. Linear kernel, Gaussian kernel, Tukey's kernel and Andrews' kernel were used in combination with kernel algorithms. The hypothesis in our work was that using robust kernel improve accuracy in clustering tasks.

The summary of our experimental results can be found in figure 1. It was found that Tukey's kernel improves the results obtained by the Gaussian kernel; this would imply that this kernel could be used in other domains where the Gaussian kernel has been successful.

References

[1] Raman Arora, Maya Gupta, Amol Kapila, and Maryam Fazel, *Clustering by left-stochastic matrix factorization*, Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 761–768.

[2] N Cristianini and J Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*, 2000.

[3] Inderjit S Dhillon, Yuqiang Guan, and Brian Kulis, *Kernel k-means: spectral clustering and normalized cuts*, (2004), 551–556.

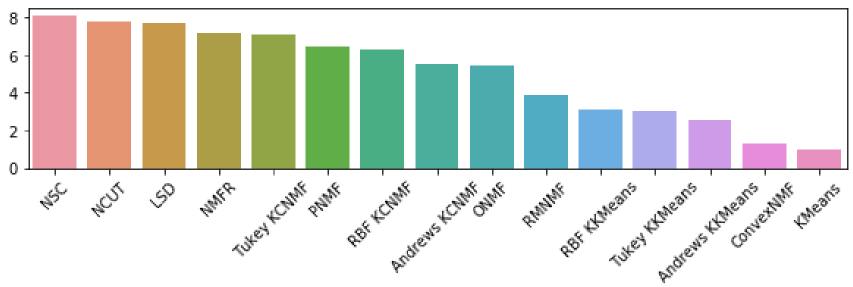

Figure 1: The average rank of each clustering method on the different datasets. The height of each bar corresponds to the average rank, a higher value is better

[4] Chris Ding, Tao Li, and Michael I Jordan, *Convex and semi-nonnegative matrix factorizations*, Pattern Analysis and Machine Intelligence, IEEE Transactions on **32** (2010), no. 1, 45–55.

[5] Fabio A González, David Bermeo, Laura Ramos, and Olfa Nasraoui, *On the robustness of kernel-based clustering*, Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, Springer, 2012, pp. 122–129.

[6] Frank R Hampel, Elvezio M Ronchetti, Peter J Rousseeuw, and Werner A Stahel, *Robust statistics: the approach based on influence functions*, vol. 114, John Wiley & Sons, 2011.

[7] PJ Huber, *Robust statistics*, 2011.

[8] Frank R.Hampel, *Robust statistics*, 1985.

[9] H Rieder and PJ Huber, *Robust statistics, data analysis and computer intensive methods*, 1996.

[10] Zhirong Yang, Tele Hao, Onur Dikmen, Xi Chen, and Erkki Oja, *Clustering by nonnegative matrix factorization using graph random walk*, Advances in Neural Information Processing Systems, 2012, pp. 1079–1087.

[11] Lijun Zhang, Zhengguang Chen, Miao Zheng, and Xiaofei He, *Robust non-negative matrix factorization*, Frontiers of Electrical and Electronic Engineering in China **6** (2011), no. 2, 192–200.