HOLDING STARS - COMPARATIVE ANALYSIS OF UNSUPERVISED LEARNING METHODS ON DATASETS OF PULSARS AND 3D HAND POSTURES

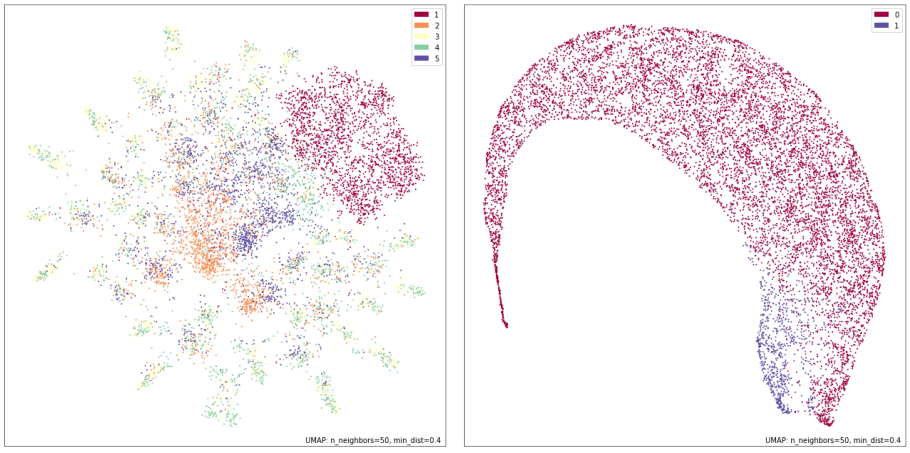
Elron Bandel
Department of Computer Science
Bar-Ilan University
elronbandel@gmail.com

ABSTRACT

In this work we conduct comparative analysis of different unsupervised learning methods for Anomaly Detection, Density Estimation and Clustering. We applied the methods on the MoCap Hand Postures Data Set and the pulsar radio emission HTRU2 Data Set . The data sets are much different, still we find common trends, UMAP for dimension reduction and Non-linear clustering methods are significantly better in finding high quality clusters. In both datasets GMM learned cluster that highly correlated with the real classes of the datasets. Our conclusion is that both data sets contain some non-linear differences between classes but within the same class features are distributed normally.

1 INTRODUCTION

In the past decades growing amount of information is being stored digitally. The accessibility of large data sets raised the possibility to learn significant insights from the available data. Naturally, most of the data available does not contain human made analysis that can guide learning systems it is required to gather insights without any annotated supervision. Clustering algorithms can indicate on patterns that characterize behavioural different groups in data, Density Estimators can obtain some knowledge about the distribution of the data and Anomaly detection methods reveal what instances of the data are rare and incidental. We used many methods to extract insights from the data sets, then compare their results and discuss each result thoroughly.



(a) MoCap Hand Postures Data Set (b) HTRU2 Data Set

Figure 1: UMAP 2D projections of the Data sets colored by the class

2 METHODS

To reduce the dimensions of the data instances, for the ease of dealing with low dimension inputs, we use Principal Component Analysis (PCA) [Pearson (1901)] or Uniform Manifold Approximation and Projection (UMAP) [McInnes et al. (2020)]. In order to project the data to 2D or 3D, which are easy to plot, we use UMAP or t-distributed Stochastic Neighbor Embedding (t-SNE) [van der Maaten & Hinton (2008)]. To estimate the density function of the distribution of the data we use Kernel Density Estimation [Rosenblatt (1956), Parzen (1962)] and Gaussian Mixture Model (GMM) [Hartigan (1985)]. To define anomalies in the data and detect outlires we use clustering based methods such as clustering with Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [Ester et al. (1996)] and than removing samples that did not fit to any cluster, we use also methods for outlires detection such as Local Outlire Factor (LOF) [Breunig et al. (2000)]. To find meaningful clusters in the data we use classic clustering algorithms, such as, K-Means [MacQueen (1967)] and Gaussian Mixture Mode (GMM) [Hartigan (1985)], as well as modern non-linear clustering algorithms, such as, Louvain Algorithm, [Blondel et al. (2008)], Prim Based Clustering [VanderPlas (2016)] and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [Ester et al. (1996)]. To evaluate the quality of the clusters we use Silhouette score [Rousseeuw (1987)] for internal evaluation and V-measure [Rosenberg & Hirschberg (2007)] to evaluate correlation with predefined target clusters. To ensure the differences in our evaluation results are statistically significant we use t-test [Gosset (1908)] for pair comparison and One Way Analysis of variance (ANOVA) [Kirk (1968)] for comparing many experiments. To further understand the statistical significance of the causes of differences we use Scheffe [Scheffé (1961)] as post-hoc statistical test.

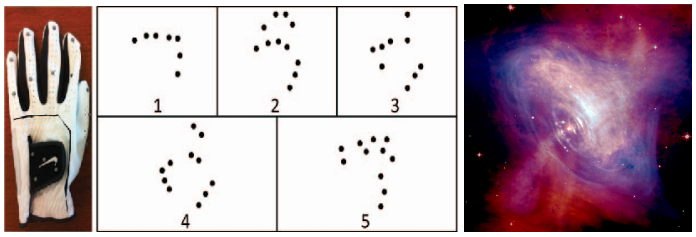
3 DATASETS

3.1 MOCAP HAND POSTURES DATA SET

MoCap Hand Postures Data set (MoCap) [Gardner et al. (2014b), Gardner et al. (2014a)] contains instances which are unorderd cloud of 3D points representing one of five hand postures. The data was gathered with few candidates that held their hand in one of the target postures wearing special glove that with markers at different points along the glove (Figure 2a). Special camera took the picture of the posture and with simple computation the markers normalized to 3D point representing relative position to the other points. To order features we filled the missing parts with the median and sorted the features by distance from regression line representing the center axis of the hand.

3.2 HTRU2 DATA SET

HTRU2 [Lyon et al. (2019)] is a data set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey [Lyon et al. (2016)]. Pulsars (Figure 2b) are a rare type of Neutron star that produce radio emission detectable here on Earth. The data set contain some statistics that describe characteristics of the radio emissions of the candidates. In the absence of additional info, each candidate could potentially describe a real pulsar. However in practice almost all detections are caused by radio frequency interference (RFI) and noise, making legitimate signals hard to find.



(a) Hand Postures Taken With Special Glove (b) Pulsar illustration

Figure 2: Illustrations of the instances measured in the Data sets

Average Silhouette Score of different Clusetring Trained with different Dimension Reduction Method			
Clustering	**PCA**	**UMAP**	**T-Test P Value**
K-Means	0.282	0.425	< 0.01
GMM	0.05	0.382	< 0.01
Louvain	0.16	0.35	< 0.01
Prim	0.1	0.6	< 0.01
Dbscan	0.322	0.55	< 0.01

Table 1: The affect of dimension reduction method on the average Silhouette score of clusters obtained with different clustering algorithms applied on MoCap data set

Log Likelihood of GMM Density Estimation Models				
Data set	**UMAP**	**PCA**	**None**	**p value ANOVA**
MoCap	-4.065	-26.12	-193.63	< 0.01
HTRU2	-3.358	-16.24	-16.131	< 0.01

Table 2: The affect of Dimension reduction method on the average per-sample average log-likelihood given a test set in different splits.

Average Log Likelihood of Kernel Density Estimation Models			
Data set	**Gaussian**	**Exponential**	**t-test p value**
MoCap	-2.874×10^{9}	-3.131×10^{6}	< 0.001
HTRU2	-4.166×10^{6}	-1.562×10^{5}	< 0.001

Table 3: The affect of kernel type on the average per-sample average log-likelihood given a test set in different splits.

4 RESULTS

In order to get statistically signficant insights we conduct every experiment 30 times with different randomization seed. The results we discuss are always the average of the results over the 30 experiments. When possible we splited the data for train and test with ratio of 0.25,0.75 every experiment with randomly different split.

4.1 DIMENSION REDUCTION

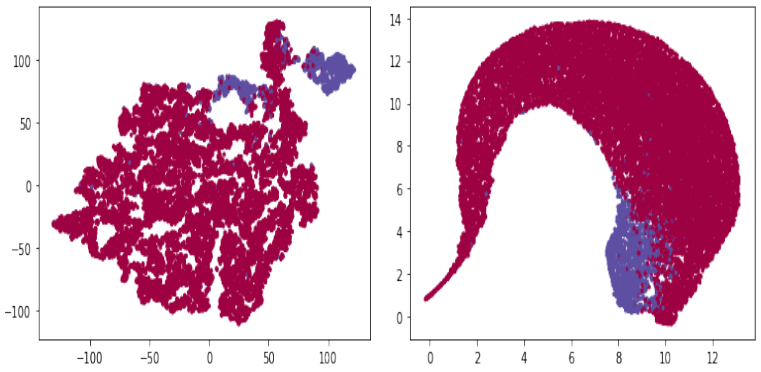
In both datasets we found UMAP signficantly better then the other methods for both representation and 2D plots. When we reduce the dimensions to only 5 dimensions we can see that UMAP obtain significantly better clusters then PCA when applied on the MoCap data set (Table 1). The same results with same level of significance hold in the HTRU2 dataset as well.

The same trend can be seen in the Density Estimation algorithms, UMAP achieved better Estimations then PCA across several methods. For example in GMM when trained for density estimation it is significantly better to use UMAP then PCA or no dimension reduction at all (Table 2) (supported as well by the post-hoc Scheffe test with p value < 0.01 between UMAP and the other pairs) .

Visualizations with UMAP (Figure 3b) seems to be slightly better then t-SNE (Figure 3a), UMAP seems to keep the clusters together even tough slightly less separated.

4.2 DENSITY ESTIMATION AND ANOMALY DETECTION

As we describe in the clustering section we found that in both data sets the GMM clusters are highly correlated with the target classes suggesting that the data has some Gaussian structure. When comparing between different Kernels for Kernel Density Estimator we found that the Gaussian kernel achieved significantly better score than exponential kernel 3.



(a) t-SNE (b) UMAP

Figure 3: HTRU2 2d projectiosn with different dimension reduction methods colored in purple if they are Pulsars otherwise in red.

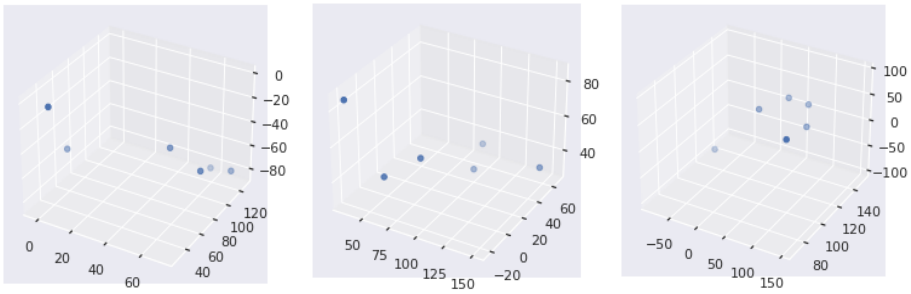
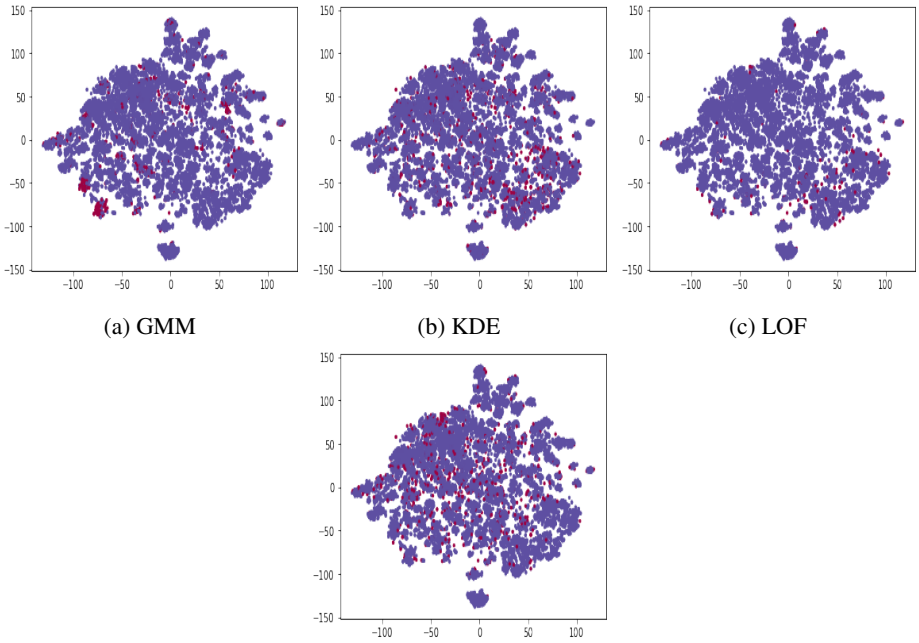


Figure 4: 3D plot of Hand Posture point clouds of anomalous examples detected by Local Outlire Factor.

Average Silehoutte score W/O LOF Outlires Removal			
Method	**With LOF**	**Without**	**t-test p value**
K-Means	0.345	0.34	< 0.01
GMM	0.34	0.319	< 0.001
Louvaine	0.213	0.145	< 0.01
Prim	0.198	0.1982	0.88

Table 4: Average Silhouette score of clusters found in HTRU2 data set with different clustering methods W/O LOF Outlires Removal

We used the density estimator to find the 5% most improbable examples and found that both KDE and GMM detected objects that seems to be in the edges of the distribution (Figure **??** and 8b). We used clustering base method for outlires detection, we classify as outlier that doesn't belong to any DBSCAN cluster. DBSCAN seem to classify many samples as out of clusters (Figure 5d). Finally Local Outlire Factor seemed to detect samples which does look like anomalies (Figure 5c). Our experiments show as well that using LOF for outlire removal significantly contribute to the quality of the clusters obtained by different algorithms methods (Table 4). This trend can be explained by the fact that in both datasets the objects of the same class have many instances that are similar to each other, thus, locally they have many neighbours, therefore LOF classify them as outlires only if they have no neighbours that are similar to them (see in Figure 4 sample of Hand Posture outlires that look nothing like any of the classes in Figure 2a). This anomaly detection and removal help the clustering algorithms to detect the distinct cluster without being disturbed by undefined instances that are much different than their cluster.



(a) GMM (b) KDE (c) LOF

(d) DBSCAN

Figure 5: t-SNE 2D Projections of Anomalies in the MoCap Dataset detected by different algorithm colored in red.

4.3 CLUSTERING

We applied 5 different clustering algorithms on both data sets. We found that for MoCap the best clusters in quality, as measured by Silhouette score, are the non linear clusters, obtained by clustering methods that take into account local density, such as DBSCAN and Prim based clustering (Figure 6).

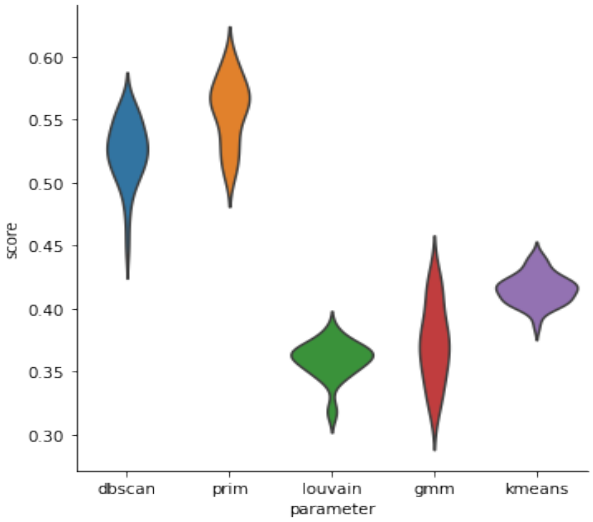


Figure 6: The distributions of Silhouette scores over 30 random runs on the MoCap data set with every clustering algorithm.

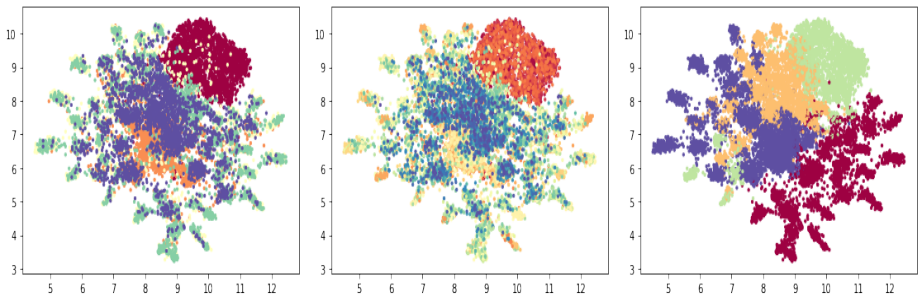
Average V-measure score of clusters correlation with target classes		
Method	**MoCap**	**HTRU2**
K-Means	0.261	0.266
GMM	0.318	0.3449
Louvaine	0.21	0.108
Prim	0.262	< 0.001
DBSCAN	0.2615	< 0.001
ANNOVA p value	< 0.001	< 0.001

Table 5: The average V-measure score of the correlation between clusters obtained by different methods and the target classes of the data sets.

In the other hand for both datasets, the best clustering algorithms, in terms of correlation with the target class (as measured by the V-measure), was Gaussian Mixture Model (Table 5). The GMM was significantly better with < 0.001 p value in the one way ANNOVA test and this result was significantly better then any other algorithm in the post-hoc Scheffe test.

looking on the distributions of the different classes projected into two dimensions (Figure 1) we can see some normal shapes that can explain how mixture of gaussians could learn the distribution of those classes.

In the other hand, judging by the visual projections of the clusters (Figure 6), the clusters that achieved the highest Silhouette scores (Figure 7a) look more like the target classes (Figure 7b) then the cluster that achieved the best V-measure correlation with the target clusters (Figure 7c). Our conclusion both the data sets contain mixture of features the are roughly distribute normally but with many non-linear changes that make the data much more complex. The non-linear algorithms succeeded in learning the complexity of the clusters (Table 4) in the data but missed their Gaussian character (Table 5), where the simple GMM succeeded greatly.

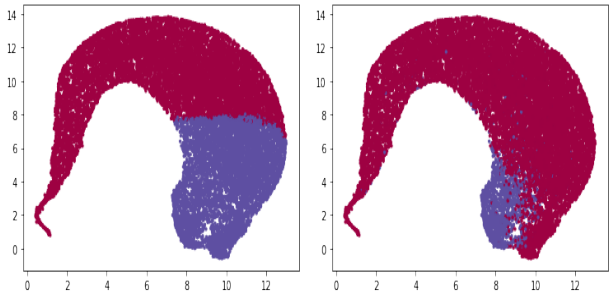


(a) Best Clusters (Prim) (b) Target Classes (c) Best Correlation (GMM)

Figure 7: UMAP 2D Projections of the MoCap dataset colored by the Clusters obtained by the best methods we found and the target classes.

5 Conclusions

The Hand Posture data set and the HTRU2 are completely different data set but surprisingly they have some similarities. Both data sets have some Gaussian like characteristics within clusters but at the same time contain some more complex patterns separating between clusters. The Gaussianity of the data sets is reflected in the relatively high V-measure score of the GMM cluster in both datasets. The complexity and non-linear characters of the datasets are reflected in the success of the non-linear clustering algorithms, as well as, the contribution of non-linear anomaly removal methods to the process of obtaining insights from the data. Our conclusion is that both data sets contain some non-linear differences between classes but within the same class features might be distributed normally.



(a) Best Clusters (GMM) (b) Target Classes

Figure 8: UMAP 2D Projections of the HTRU2 dataset colored by the Clusters obtained by the best method we found and the target classes.

REFERENCES

Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, Oct 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/10/p10008. URL http://dx.doi.org/10.1088/1742-5468/2008/10/P10008.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

M. Ester, H. Kriegel, J. Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, 1996.

Andrew Gardner, Christian A Duncan, Jinko Kanno, and Rastko Selmic. 3d hand posture recognition from small unlabeled point sets. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 164–169. IEEE, 2014a.

Andrew Gardner, Jinko Kanno, Christian A Duncan, and Rastko Selmic. Measuring distance between unordered sets of different sizes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 137–143, 2014b.

John A Hartigan. Statistical theory in clustering. *Journal of classification*, 2(1):63–76, 1985.

Roger E Kirk. Experimental design: Procedures for the behavioral. *Sciences*, 93:237–244, 1968.

R. J. Lyon, B. W. Stappers, S. Cooper, J. M. Brooke, and J. D. Knowles. Fifty years of pulsar candidate selection: from simple filters to a new principled real-time classification approach. *Monthly Notices of the Royal Astronomical Society*, 459(1):1104–1123, 04 2016. ISSN 0035-8711. doi: 10.1093/mnras/stw656. URL https://doi.org/10.1093/mnras/stw656.

RJ Lyon, BW Stappers, Lina Levin, MB Mickaliger, and Anna Scaife. A processing pipeline for high volume pulsar candidate data streams. *Astronomy and Computing*, 28:100291, 2019.

James B. MacQueen. Some methods for classification and analysis of multivariate observations. 1967.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

Emanuel Parzen. On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076, 1962. doi: 10.1214/aoms/1177704472. URL https://doi.org/10.1214/aoms/1177704472.

Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 6*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. URL https://www.tandfonline.com/doi/abs/10.1080/14786440109462720.

Andrew Rosenberg and Julia Hirschberg. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pp. 410–420, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D07-1043.

Murray Rosenblatt. Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3):832 – 837, 1956. doi: 10.1214/aoms/1177728190. URL https://doi.org/10.1214/aoms/1177728190.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL http://www.sciencedirect.com/science/article/pii/0377042787901257.

H. Scheffé. The analysis of variance. wiley, new york 1959, 477 seiten, $ 14,00. *Biometrische Zeitschrift*, 3(2):143–144, 1961. doi: https://doi.org/10.1002/bimj.19610030206. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.19610030206.

L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. 2008.

Jake VanderPlas. mst$_c lustering : Clusteringviaeuclideanminimumspanningtrees.$*Journal of Open Source Software*, 1 (1) : 12, 2016.*doi* : . URL https://doi.org/10.21105/joss.00012.