LAAT: Locally Aligned Ant Technique for discovering multiple faint low dimensional structures of varying density

Abolfazl Taghribi, Kerstin Bunte, Rory Smith, Jihye Shin, Michele Mastropietro, Reynier F. Peletier, and Peter Tiňo

---

Our method is neither a traditional clustering, nor a manifold learning method. However, we performed additional experiments that demonstrate how the application of LAAT as a preprocessing tool can improve the results of subsequent clustering. Although LAAT is an independent approach for the detection of noisy structures, it can be applied as a preprocessing step to enhance clustering results, when clusters form elongated manifolds corrupted by noise and/or are embedded in a noisy background. In the following we compare denoising with LAAT to the denoising step of two recent techniques, namely LLPD[1] and ADBSCAN[2], to cluster two real-world data sets from the UCI repository [3], namely the Skin and Banknotes datasets, that contain 245057 and 1372 samples in two classes, respectively. Both contain noise and the points are distributed on elongated clusters or manifolds, which adds to the complexity of the clustering problem and makes them suitable for demonstrating LAAT.

We first cluster each dataset with these techniques using the publicly available code and repeat each experiment using the parameter ranges recommended for their use. We present the best result under the constraint that the correct number of clusters is determined beforehand. Next, the data is denoised by LAAT and subsequently clustered using LLPD and ADBSCAN. For ADBSCAN the noise percentage can be set to zero, such that it does not remove any further points during the clustering step, which could otherwise remove important parts of the clusters. For LLPD the cut-off value needs to be selected again, as suggested in [1]. Panels a-c in Figure 1 illustrate the 44%, 18.1% and 18.1% points selected as noise from Banknotes by LLPD, LAAT, and ADBSCAN, respectively. For the purpose of visualization we projected the data into 3-dimensions using parametric linear tSNE [4]. LAAT identifies more noise points between the two clusters resulting in easier separation by subsequently applied LLPD and ADBSCAN clustering.

LLPD identifies these noise points only by removing a larger fraction of points (44%). The bottom row visualizes the result of removing 12% of noise points in the Skin data set using LLPD and LAAT. Note that ADBSCAN could not determine the correct number of clusters. While LAAT mostly removes outliers and noise between the clusters, LLPD removes many points on the major manifolds themselves, which visually appears to remove fewer points since points detected as noise are hidden within the structure in the 3D plot. Table 1 summarizes the Normalized Mutual Information (NMI) index using the true cluster memberships to

Table 1: Evaluation of clustering using the Normalized Mutual Information (NMI) index for LLPD and ADBSCAN, on their own and subsequently applied after noise removal with LAAT.

Dataset	Method	Noise %	NMI
Banknotes	LLPD	44	0.704
Banknotes	LAAT+LLPD	22	0.932
Banknotes	ADBSCAN(n=0.35)	18.1	0.621
Banknotes	LAAT+ADBSCAN	18.1	0.901
Skin	LLPD	12	0.957
Skin	LAAT+LLPD	12	0.962

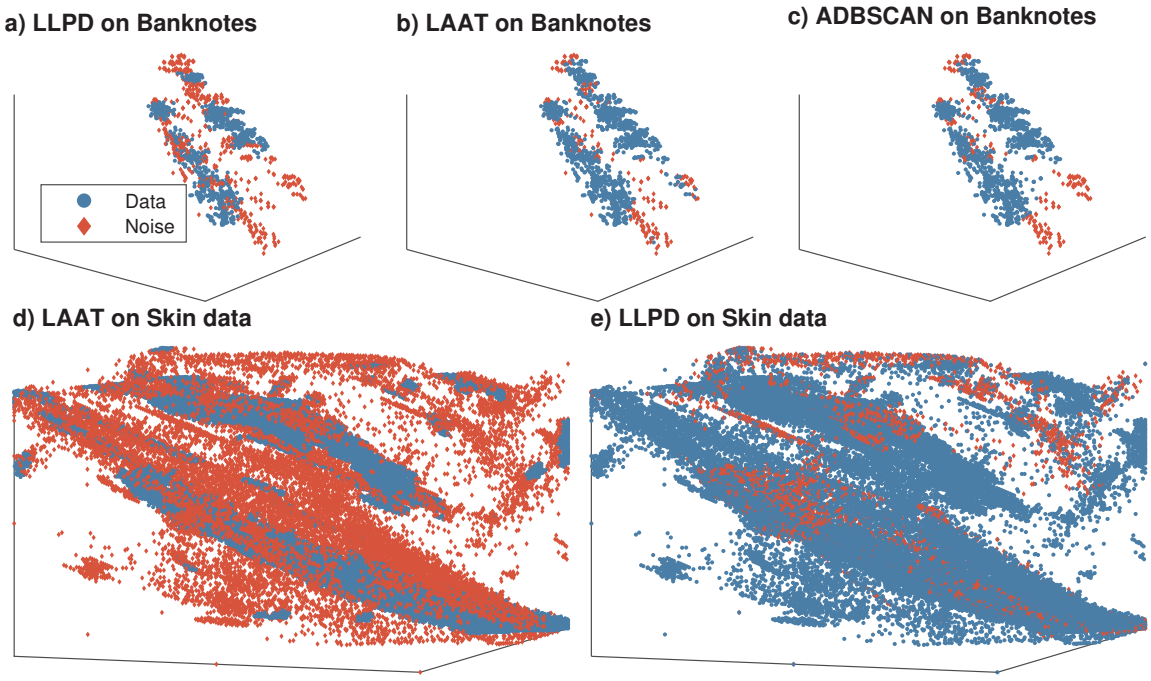

Figure 1: Plots of denoised data by LLPD, LAAT and ADBSCAN for Banknotes (top) and Skin data (bottom).

provide a quantitative comparison of the different strategies. We observe that LAAT improves the result for these elongated clusters. Note, that LLPD needs to remove double the amount of points in the Banknotes data set until it recognizes the points between the manifolds as noise. In contrast, removing only half the amount of data with LAAT the remaining points separate the clusters very well, increasing the agreement of the cluster membership with the ground truth by over 20%. This demonstrates the ability of LAAT to separate structures, such as clusters, meaningfully in the clustering context. Note that the general aim of noise removal for structure detection does not require a measure on separation. In fact, interesting astronomical structures, such as streams, filaments, walls, and nodes are typically connected.

References

[1] A. Little, M. Maggioni, and J. M. Murphy, "Path-Based Spectral Clustering: Guarantees, Robustness to Outliers, and Fast Algorithms," *Journal of Machine Learning Research*, vol. 21, no. 6, pp. 1–66, 2020.

[2] H. Li, X. Liu, T. Li, and R. Gan, "A novel density-based clustering algorithm using nearest neighbor graph," *Pattern Recognition*, vol. 102, p. 107206, June 2020.

[3] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[4] K. Bunte, S. Haase, M. Biehl, and T. Villmann, "Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences," *Neurocomputing*, vol. 90, pp. 23–45, Aug. 2012.