A Feature Selection method based on Topological aspect of Feature Manifolds: Local Maximum Selection

Jung Ho PARK, Jeamin PARK, Hyeon-Woo JEONG, Hyun-Young CHOI

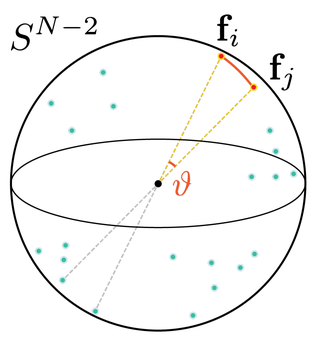
Feature manifold $\mathbb{R}P^{N-2}$

If a feature of the data set is normalized by $f \mapsto \frac{f-\mu_f}{\sqrt{N}\sigma_f}$, where N is the number of samples, they are projected on $(N-2)$-dimensional sphere S^{N-2} by a map

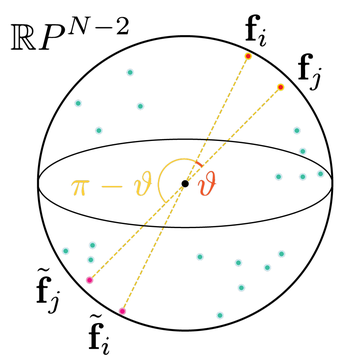
$$\mathbb{R}^N \to \{f = (f_1, , f_N) : f_1 + \cdots + f_N = 0\} \to S^{N-2}.$$

Furthermore, we identify features that situated directly opposite sides because they are the same factor in the process of ML so that the feature manifold can be regarded as $(N-2)$**-dimensional projective space** $\mathbb{R}P^{N-2} = S^{N-2}/(x \sim -x)$.

1. For $f_i, f_j \in S^{N-2}$,
$$d(f_i, f_j) = \angle(f_i, f_j) = \arccos\left(corr(f_i, f_j)\right)$$
is a standard geodesic distance on S^{N-2}.
2. By identifying $\mathbb{R}P^{N-2} = S^{N-2}/(x \sim -x)$ in set-wise sense, we give a metric on $\mathbb{R}P^{N-2}$ by the **Hausdorff distance**, which is formulated by
$$\tilde{d}(f_i, f_j) = \min\{\angle(f_i, f_j), \pi - \angle(f_i, f_j)\} = \arccos|corr(f_i, f_j)|.$$
3. Process:



$\arccos(\text{correlation}(\mathbf{f}_i, \mathbf{f}_j))$

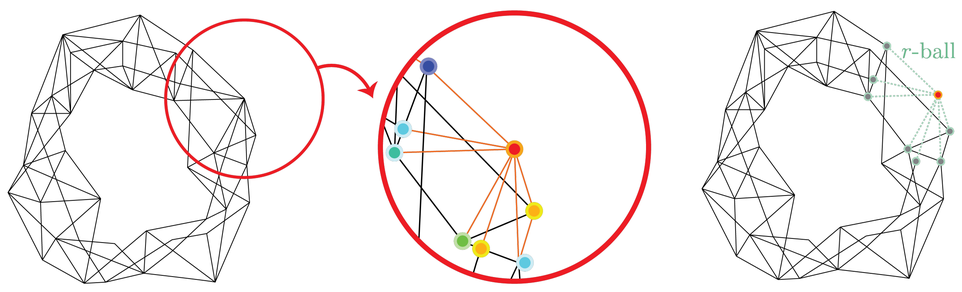


$\arccos(|\text{correlation}(\mathbf{f}_i, \mathbf{f}_j)|)$

Local Maximum Selection

For a good performance of ML, we chose wide-spread high-score features. The selection is carried out in sequence by choosing the feature attaining local maximum and removing the its neighbor.

1. **k-Nearest Neighbor Network(kNNN)** is obtained from feature manifold by connecting edges to k nearest features for each feature.
2. **Normalized Morse Index(NMI)** is a function on the vertex set of kNNN $\text{NMI} : V(\text{kNNN}) \to [0, 1]$ given by
$$\text{NMI}(f) = \frac{\#(\text{neighbor whose score is lower than } f)}{\text{degree of } f}.$$
3. Steps:
   1. Choose the feature whose NMI is maximal along feature family.
   2. Remove r-ball in kNNN centered at chosen feature.
   3. Repeat 1,2.
4. **Process:**



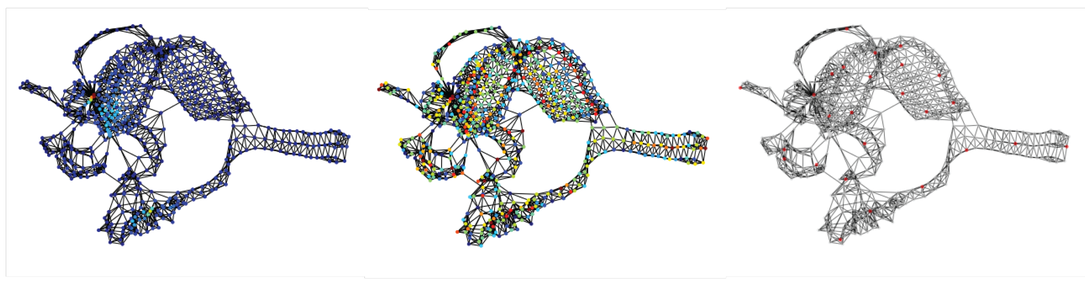
Numerical experiment

1. Data description

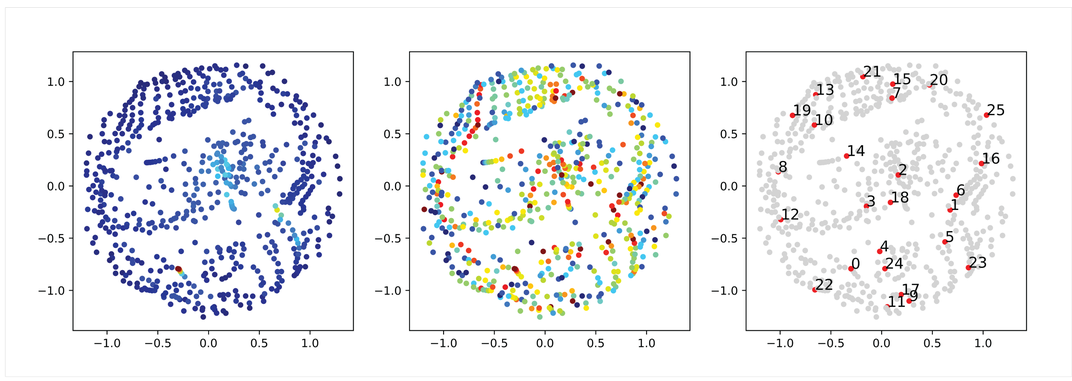
Dataset name	Problem description and reference	Number of classes	Number of features	Number of instances
isolet	Classification of Isolated Letter Speech Recognation	2	617	6238(training) 1559(test)

2. Machine Learning Models:
   - Logistic Regression (LR)
   - Random Forest RF
   - Gaussian Naive Bayes (GNB)
   - Decision Tree Classifier (DTC)
   - Gradient Boosting Classifier (GBC)
   - XG Boost Classifier (XG)
3. We observed that the tendency of accuracy of ML according to the number of features selected by our method (LMS). And we compared it to those of random selection.

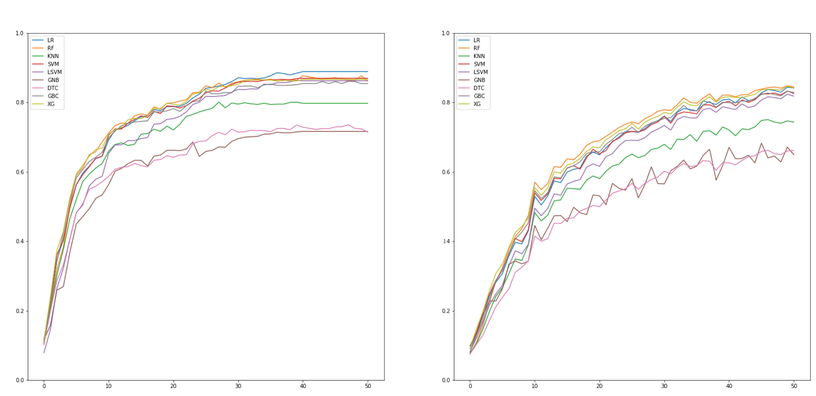
Result



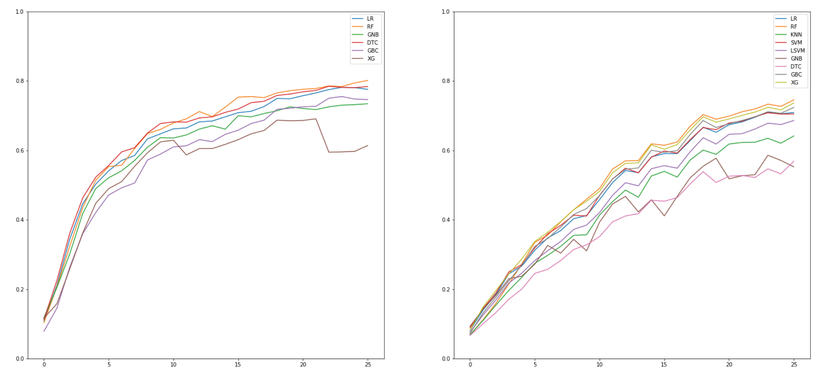
kNNN of features in data. *Left*: Fisher score. *Middle*: NMI. *Right*: Selected features by LMS with $r = 3$.



Visualization of the feature manifold by Multidimensional Scaling(MDS). *Left*: Fisher score. *Middle*: NMI. *Right*: Selected features by LMS with $r = 3$.



Left: Result of LSM with r=2. *Right*: Result of random selection.



Left: Result of LSM with r=3. *Right*: Result of random selection.

Conclusion

In this study, we suggest a new feature selection method applying Morse theory and topological manifold which are geometric notions. High accuracy in numerical experiment proves that our theoretical approach is reasonable. Moreover, when taking proper dominant features sequentially, our method converges faster than Random Select Method.

References

[1].Vin de Silva and Gunnar Carlsson "Topological estimation using witness complexes," Eurographics Symposium on Point-Based Graphics (2004) [2].Gurjeet Singh, Facundo Mémoli, "Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition," Eurographics Symposium on Point- Based Graphics (2007) [3].Quanquan Gu, Zhenhui Li, Jiawei Han, "Generalized Fisher Score for Feature Selection," In Proc. of the 27th Conference on Uncertainty in Artificial Intelligence (UAI), Barcelona, Spain, 2011