An Intelligent Clustering Algorithm for High Dimensional and Highly Overlapped Photo-Thermal Infrared Imaging Data

Nian Zhang and Lara Thompson

Department of Electrical and Computer Engineering, University of the District of Columbia, 4200 Connecticut Ave NW, Washington, DC, 20008/Department of Mechanical Engineering, University of the District of Columbia, 4200 Connecticut Ave NW, Washington, DC, 20008

Abstract

This paper analyzes the noise-free but highly overlapped photo-thermal infrared imaging data set involving four analytes and two substrates. We developed an effective

performance due to the large search space known as “the curse of dimensionality”. Principal component analysis (PCA) is a quantitatively rigorous method for removing irrelevant and redundant features [6]. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. The method generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information. Several top ranking principal components will be selected to form a new feature space. The original samples will be transformed to this new feature space in the directions of the principal components. Although the PCA can effectively reduce the number of dimensions by selecting the top ranking principal components, PCA method is not able to select a subset of features which are important to distinguish the classes. It only guarantees that when you project each observation on an axis (along a principal component) in a new space, the variance of the new variable is the maximum among all possible choices of that axis. This means that each feature is considered separately, thereby ignoring feature dependencies, which may lead to worse classification performance.

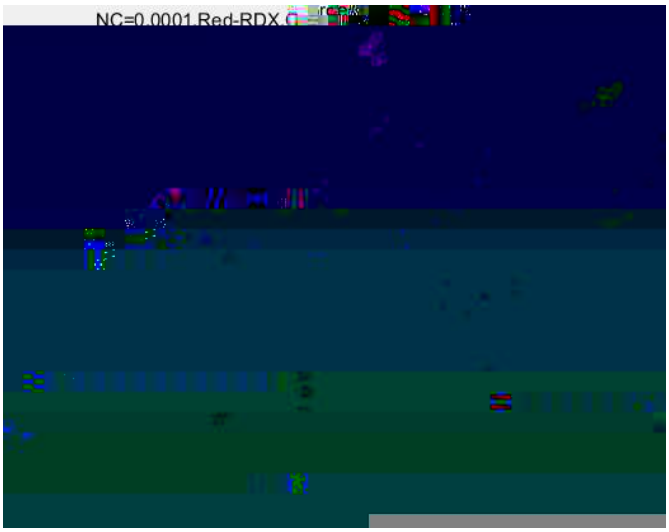
The feature selection problem is essentially a combinatorial optimization problem which is computationally expensive. We consider the feature selection problem in unsupervised learning scenario, which is particularly difficult due to the absence of class labels that would guide the search for relevant information. The existing and most powerful unsupervised feature selection technique is principal component analysis (PCA) [7]. It is often useful to measure data in terms of its principal components rather than on a normal x-y axis. They’re the underlying structure in the data. They are the directions where there is the most variance, the directions where the data is most spread out. The PCA technique were applied to the data set to reveal the patterns in data, as well as reduce the dimension of feature vectors (i.e. vectors containing the principal components). First we deconstruct the set into eigenvectors and eigenvalues. An eigenvector is a direction, and an eigenvalue is a number, telling you how much variance there is in the data in that direction. The amount of eigenvectors/values that exist equals the number of dimensions the data set has. The k-means clustering algorithm is demonstrated in Table I.

Table I. K-means Clustering Algorithm

Steps	Activities
1	k initial "means" (k is a estimated value) are randomly generated within the data domain.
2	k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram [8] generated by the means.
3	The centroid of each of the k clusters becomes the new mean.
4	Steps 2 and 3 are repeated until convergence has been reached.

4. Analysis & Results

We presented the principal component analysis (PCA) results using a combination of top PCs, i.e. PC1 and PC2. We displayed the data in PC1-PC2 axes, as shown in Fig. 2. Then we used the K-mean clustering algorithm to classify them into 6 classes. The classes are demonstrated in PC1-PC2 axes, as shown in Fig. 3. Red represents RDX, Green represents TNT, Blue represents AN, Magenta represents Sucrose, Black represents



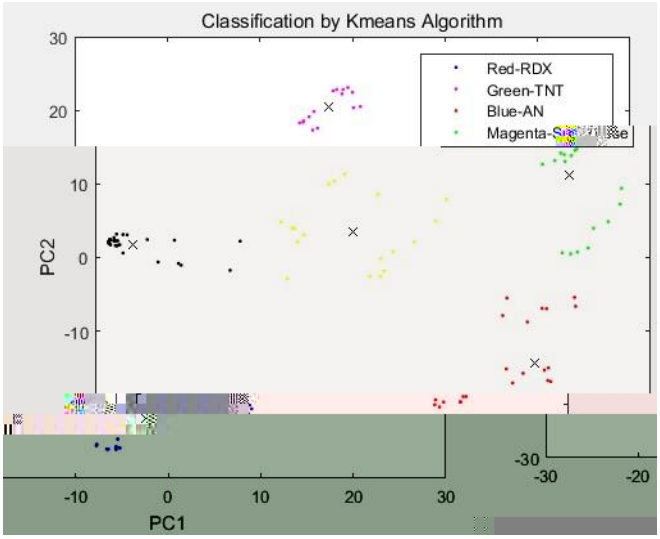


TABLE IX. K-MEANS CLUSTERING RESULTS FOR DATA ON PC1 AND PC2

5. Conclusions

This paper analyzes the noise-free but highly overlapped photo-thermal infrared imaging data set. The principal component analysis (PCA) was used to reduce the dimension of data space to the top principal components feature (PC1-PC2) space, and thus the most prominent features or patterns were revealed. Then we used the K-mean clustering algorithm to classify them into four analytes and two substrates. We used the performance evaluation matrices

5. M. Kubat, S. Matwin, "Addressing the Curse of Imbalanced Training Sets: One-Sided Selection," Proceedings of the 14th Annual International Conference on Machine Learning, 1997.
6. Principal Component Analysis (PCA), http://www.mathworks.com/help/stats/principal-component-analysis-pca.html?requestedDomain=www.mathworks.com.
7. D. Cai, C. Zhang, X. He, "Unsupervised Feature Selection for Multi-Cluster Data," *The 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (*KDD'10), 2010, pp. 333-342.
8. F. Aurenhammer, "Voronoi Diagrams –