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 algorithm which combines the principal component analysis and k-mean clustering to classify the materials into six classes. We conducted a dimension reduction by applying the principal component analysis (PCA) on the data to transform the original data to the principal component 1 (PC1) and principal component 2 (PC2) feature space. The data were revealed in PC1-PC2 space and formed into clusters. Then we used the K-mean clustering algorithm to classify them into six classes including RDX, TNT, AN, Sucrose, Paint, and Glass. The prototype of the predominant analytes in each cluster will determine the class. After that, we conducted the performance evaluation by calculating the probability of detection (POD), false alarm rate (FAR), accuracy, precision, and recall based on true positive (TP), false negative (FN), false positive (FP), and true negative (TN). The results demonstrated that the proposed algorithm effectively reduced dimension and accurately determined the classes of those analytes and substrates.

Keywords

Classification, clustering, principal component analysis, k-mean clustering, big data, high dimensional, overlapped data.

1. Introduction

Headings Recent advances in modern technologies, such as infrared remote sensing technology, 4D CT-scans technology, and DNA microarrays have led to the proliferation of massive and imbalanced data [1]. Classification of these data becomes very difficult because they are usually overlapped and minority classes surrounded by majority classes [2][3]. The massive data analysis problem presents a significant new challenge to the machine learning and data mining community because in most instances real-world data is imbalanced, and therefore it has attracted significant research recent years [4]. In many cases, identifying rare objects is of crucial importance, because such skewed distributions (i.e. if sample from one class is in higher number than other) normally carry the most interesting and critical information. This critical information is necessary to support the decision-making process in battlefield scenarios, such as anomaly or intrusion detection. The fundamental issue with imbalanced learning is the ability of imbalanced data to compromise the performance of standard learning algorithms, which assume balanced class distributions or equal misclassification penalty costs [5]. Therefore, when presented with complex imbalanced data sets these algorithms may not be able to properly represent the distributive characteristics of the data.

The overlapped data problem becomes even severe when the dimensionality of data is high. In this situation, feature selection usually becomes essential to the learning algorithm because high-dimensional data tends to decrease the efficiency of most learning algorithms. This is also widely known as the curse of dimensionality. Feature selection techniques are designed to find an optimal subset of relevant feature subset of the original features which are the most distinct features that can be used to differentiate samples into different classes. Therefore, this paper will analyze the noise-free but highly overlapped and imbalanced data set. The objective is to develop algorithms to discover the underlying mechanism that affect the clustering performance on different combination of principal components and different number of features.

The remaining of the paper is organized as follows. In Section 2, the high dimensional and overlapped data set is described. In Section 3, the proposed methodology is presented. Principal component analysis (PCA) and k-means clustering algorithm are described. In Section 4, the analysis and results are presented. In Section 5, the conclusions are given.

2. Data Set

The data set is a synthetic photo-thermal infrared imaging spectroscopy dataset. The dimension is 1254x123, where its rows are the features and the columns are the instances. It contains six different kinds of analytes. Each column contains 1254 features. The color of the data point is proportional to signal strength, i.e. red represents high, and blue represents low. We may demonstrate the signal matrix for all the 123 samples. This can be seen by display the data set in false color plot which will show visible or non-visible parts of the electromagnetic spectrum. The false color plot of the data set is shown in Fig. 1. In this data set, some complicated and overlapping samples can happen. The mixing IR absorption/emission features reflects the primary challenge to a useful detection technique in the real-world application.

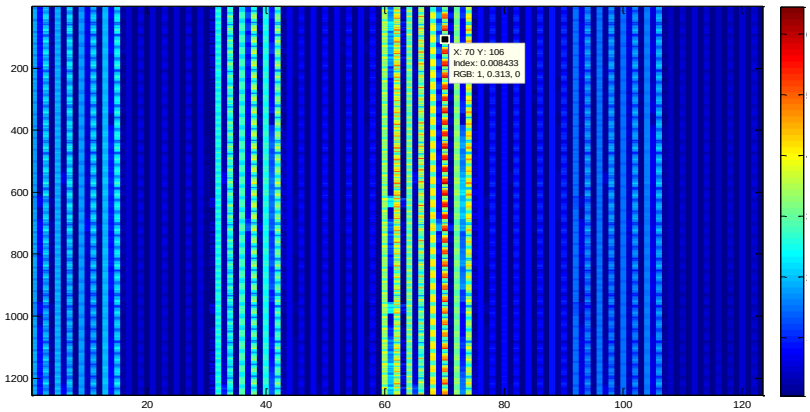

Fig. 1 False color plot of data set. The data set has 123 samples and 1254 features.

3. Methodology

A data set contains relevant, irrelevant, and redundant features. However, irrelevant and redundant features are not useful for classification, and they may even reduce the classification

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 Paint, and Yellow represents Glass. The centroids of the classes were indicated by black crosses.

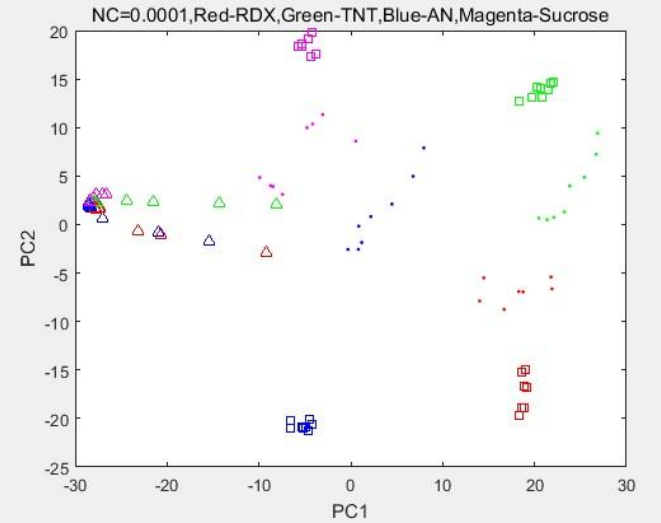

Fig. 2 All the data projected on PC1-PC2 space after applying PCA.

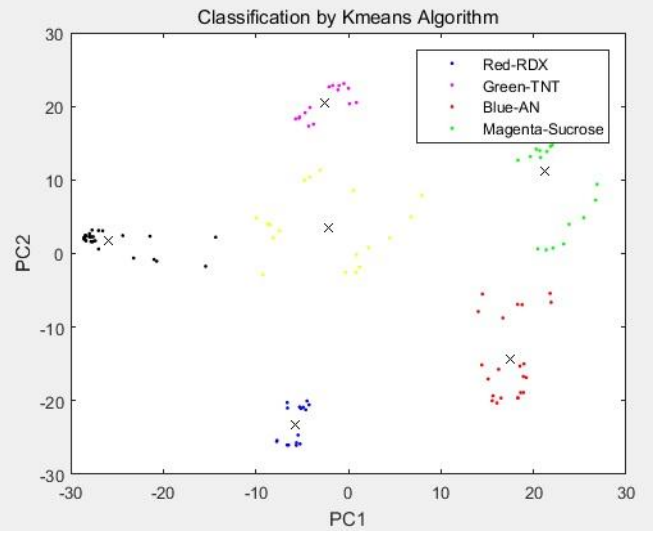

Fig. 3 Six clusters consisting of the four analytes and two substrates were formed using the K-means clustering method.

The number of analytes and substrates in each of the six classes are shown in Table II. The rows represent RDX, TNT, AN, Sucrose, Paint, and Glass, respectively. The columns represent Class 1, Class 2, Class 3, Class 4, Class 5, and Class 6.

Table II. Classification Using the K-means Clustering Algorithm

	Class1	Class2	Class3	Class4	Class5	Class6
RDX	0	0	22	0	1	2
TNT	0	0	0	24	1	3
AN	16	0	0	0	2	4
Sucrose	0	15	0	0	2	3
Paint	0	0	0	0	8	7
Glass	0	0	0	0	4	9

By counting the largest number of analytes/substrate in each class, we can find the analyte/substrate dominating that class. For each cluster, the majority analytes will determine the class that this cluster belongs to. Therefore, we find Class 1 represents AN, Class 2 represents Sucrose, Class 3 represents RDX, Class 4 represents TNT, Class 5 represents Paint, and Class 6 represents Glass. These labels are shown in the tables in the following Clustering Analysis section.

Clustering Analysis

We labeled the samples by indicating not only TP, FN, FP, and TN, but also different color. Six tables are generated for the six analytes in Table III-VIII.

Legend:

TABLE III. LABELING FOR RDX

Analytes	Class 1 (AN)	Class 2 (Sucrose)	Class 3 (RDX)	Class 4 (TNT)	Class 5 (Paint)	Class 6 (Glass)
RDX	0 FN	0 FN	22 TP	0 FN	1 FN	2 FN
TNT	0 TN	0 TN	0 FP	24 TN	1 TN	3 TN
AN	16 TN	0 TN	0 FP	0 TN	2 TN	4 TN
Sucrose	0 TN	15 TN	0 FP	0 TN	2 TN	3 TN
Paint	0 TN	0 TN	0 FP	0 TN	8 TN	7 TN
Glass	0 TN	0 TN	0 FP	0 TN	4 TN	9 TN

TABLE IV. LABELING FOR TNT

Analytes	Class 1 (AN)	Class 2 (Sucrose)	Class 3 (RDX)	Class 4 (TNT)	Class 5 (Paint)	Class 6 (Glass)
RDX	0 TN	0 TN	22 TN	0 FP	1 TN	2 TN
TNT	0 FN	0 FN	0 FN	24 TP	1 FN	3 FN
AN	16 TN	0 TN	0 TN	0 FP	2 TN	4 TN
Sucrose	0 TN	15 TN	0 TN	0 FP	2 TN	3 TN
Paint	0 TN	0 TN	0 TN	0 FP	8 TN	7 TN
Glass	0 TN	0 TN	0 TN	0 FP	4 TN	9 TN

TABLE V. LABELING FOR AN

Analytes	Class 1 (AN)	Class 2 (Sucrose)	Class 3 (RDX)	Class 4 (TNT)	Class 5 (Paint)	Class 6 (Glass)
RDX	0 FP	0 TN	22 TN	0 TN	1 TN	2 TN
TNT	0 FP	0 TN	0 TN	24 TN	1 TN	3 TN
AN	16 TP	0 FN	0 FN	0 FN	2 FN	4 FN
Sucrose	0 FP	15 TN	0 TN	0 TN	2 TN	3 TN
Paint	0 FP	0 TN	0 TN	0 TN	8 TN	7 TN
Glass	0 FP	0 TN	0 TN	0 TN	4 TN	9 TN

TABLE VI. LABELING FOR SUCROSE

Analytes	Class 1 (AN)	Class 2 (Sucrose)	Class 3 (RDX)	Class 4 (TNT)	Class 5 (Paint)	Class 6 (Glass)
RDX	0 TN	0 FP	22 TN	0 TN	1 TN	2 TN
TNT	0 TN	0 FP	0 TN	24 TN	1 TN	3 TN
AN	16 TN	0 FP	0 TN	0 TN	2 TN	4 TN
Sucrose	0 FN	15 TP	0 FN	0 FN	2 FN	3 FN
Paint	0 TN	0 FP	0 TN	0 TN	8 TN	7 TN
Glass	0 TN	0 FP	0 TN	0 TN	4 TN	9 TN

TABLE VII. LABELING FOR PAINT

Analytes	Class 1 (AN)	Class 2 (Sucrose)	Class 3 (RDX)	Class 4 (TNT)	Class 5 (Paint)	Class 6 (Glass)
RDX	0 TN	0 TN	22 TN	0 TN	1 FP	2 TN
TNT	0 TN	0 TN	0 TN	24 TN	1 FP	3 TN
AN	16 TN	0 TN	0 TN	0 TN	2 FP	4 TN
Sucrose	0 TN	15 TN	0 TN	0 TN	2 FP	3 TN
Paint	0 FN	0 FN	0 FN	0 FN	8 TP	7 FN
Glass	0 TN	0 TN	0 TN	0 TN	4 FP	9 TN

TABLE VIII. LABELING FOR GLASS

Analytes	Class 1 (AN)	Class 2 (Sucrose)	Class 3 (RDX)	Class 4 (TNT)	Class 5 (Paint)	Class 6 (Glass)
RDX	0 TN	0 TN	22 TN	0 TN	1 TN	2 FP
TNT	0 TN	0 TN	0 TN	24 TN	1 TN	3 FP
AN	16 TN	0 TN	0 TN	0 TN	2 TN	4 FP
Sucrose	0 TN	15 TN	0 TN	0 TN	2 TN	3 FP
Paint	0 TN	0 TN	0 TN	0 TN	8 TN	7 FP
Glass	0 FN	0 FN	0 FN	0 FN	4 FN	9 TP

We conducted the clustering performance evaluation by calculating the evaluation metrics, including the probability of detection (POD), false alarm rate (FAR), accuracy, precision, and recall. This process was facilitated by developing an automatic algorithm to determine the true positive (TP), false negative (FN), false positive (FP), and true negative (TN), which are components of the above performance evaluation matrices. The above evaluation metrics are defined as follows [9]:

Probability of Detection: POD = TP/(TP+FN)

False Alarm Rate: FAR = FP/(FP+TN)

Precision: P = TP/(TP+FP)

Recall: R = TP/(TP+FN)

Accuracy: Accuracy = (TP+TN)/(TP+FP+FN+TN)

From the equations, we see that precision measures how often an instance was predicted as positive that is actually positive, while recall measures how often a positive class instance in the data set was predicted as a positive class instance by the classifier. In imbalanced data set, the goal is to improve recall without hurting precision. These goals, however, are often conflicting, since in order to increase the TP for the minority class, the number of FP is also often increased, resulting in reduced precision.

The k-means clustering algorithm on the data on PC1 and PC2 was performed. The clustering performance including the probability of detection (POD), false alarm rate (FAR), accuracy, precision, and recall was conducted. The performance results are shown in Table I. The six clusters were determined by the majority analyte type in each cluster.

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

	Class 1 (AN)	Class 2 (Sucrose)	Class 3 (RDX)	Class 4 (TNT)	Class 5 (Paint)	Class 6 (Glass)
POD	73%	75%	88%	86%	44%	69%
FAR	0	0	0	0	9%	17%
Accuracy	87%	88%	98%	97%	86%	81%
Precision	100%	100%	100%	100%	44%	32%
Recall	73%	75%	88%	86%	53%	69%
F1 Score	84%	86%	94%	92%	48%	44%

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 to measure the accuracy of classification. The experimental results demonstrated that the combination of the principal component analysis and K-means clustering algorithm are efficient for achieving dimensional reduction and clustering on highly overlapped photo-thermal infrared imaging data. The accuracy of the classification of AN, Sucrose, RDX, TNT, Paint, and Glass is 87%, 88%, 98%, 97%, 86%, and 81%, respectively.

Acknowledgement

This work was supported in part by the National Science Foundation under Grants HRD #1505509 and HRD #1533479.

References

1. H. He and E. A. Garcia, “Learning from Imbalanced Data,” *IEEE Trans. Knowledge and Data Engineering*, vol. 21, no. 9, 2009, pp. 1263-1284.
2. R. Longadge, S. Dongre, and L. Malik, “Class Imbalance Problem in Data Mining: Review,” *International Journal of Computer Science and Network (IJCSN)*, vol. 2, no. 1, 2013.
3. N. Chawla, N. Japkowicz, and Aleksander Kolcz “Special Issue on Learning from Imbalanced Data Sets,” *SIGKDD Explorations*, vol. 6, no. 1, 2004, pp 1-6.
4. N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research,* vol. 16, 2003, pp. 321-357.

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 – A Survey of a Fundamental Geometric Data Structure," *ACM Computing Surveys*, vol. 23, no.3, 1991, pp. 345–405.
9. H. He and Y. Ma, Editors, Imbalanced Learning: Foundations, Algorithms, and Applications, Wiley-IEEE, ISBN: 978-1-118-07462-6, Hardcover, 216 pages, Wiley-IEEE, July 2013.