Comparison of Clustering Pipelines for the Analysis of Mass Spectrometry Imaging Data

Sanaiya Sarkari, Chanchala D. Kaddi, Rachel V. Bennett, Facundo M. Fernández and May D. Wang- *IEEE Senior Member*

***Abstract*— Mass spectrometry imaging (MSI) is valuable for biomedical applications because it links molecular and morphological information. However, MSI datasets can be very large, and analyzing them to identify important biological patterns is a challenging computational problem. Many types of unsupervised analysis have been applied to MSI data, and in particular, clustering has recently gained attention for this application. In this paper, we present an exploratory study of the performance of different analysis pipelines using k-means and fuzzy k-means clustering. The results indicate the effects of different pre-processing and parameter selections on identifying biologically relevant patterns in MSI data.**

I. Introduction

Mass spectrometry imaging (MSI) is a technique used to measure the molecular composition of a sample across its surface. MSI data is three-dimensional, with spatial (*x,y*) dimensions corresponding to the sample dimensions, and a spectral (*z*) dimension corresponding to the m/z (mass-to-charge ratio) values measured by the mass spectrometer. A key advantage of MSI is its ability to simultaneously measure the spatial distribution of thousands of m/z values in a single dataset [1]. This is much greater than multiplex techniques in immunohistochemistry, which also require prior knowledge of targets and the use of specific antibodies. Thus, MSI is an '-omic' technology that enables the discovery of meaningful spatial molecular patterns. However, MSI data size presents an analytical challenge.

Many different unsupervised techniques have been applied to identify patterns in MSI data. Principal component analysis (PCA), a dimensionality reduction technique, is a common method [2]. Related methods, such as non-negative matrix factorization and independent component analysis, have also been tested [3, 4]. Clustering is an important approach for MSI data analysis. Both hierarchical and k-means clustering have been implemented for MSI data, and have been shown to yield highly relevant clusters corresponding to biological structures. For example, Alexandrov and colleagues recently showed that k-means can differentiate between tumor and non-tumor regions in larynx carcinoma MSI data [5]. Other studies by the same group applied clustering techniques to matrix assisted laser desorption/ionization (MALDI)-MSI data to perform functionally and anatomically meaningful spatial segmentation [6, 7]. Hierarchical clustering has also been investigated in MSI; McCombie and colleagues used hierarchical clustering, together with PCA and discriminant analysis [8]. Another recent study also performed hierarchical clustering following PCA in order to compare MALDI-MSI data to histological data for cancer, and found that the results are not always congruent [9].

While these recent studies have demonstrated the efficacy of clustering for MSI analysis, typically a fixed clustering protocol – i.e., data pre-processing steps and clustering with a chosen algorithm and parameters – is followed. However, the end clustering result is sensitive to the choices of data pre-processing methods, clustering algorithm, and its parameters. For k-means clustering in MSI, parameters of particular interest are the value *k,* the distance metric, and the dimensionality reduction method, if any. Systematic comparison of alternative choices for each of these pipeline components could lead to improved clustering results.

In this paper, we present an exploratory study of the effects of these alternatives on clustering MSI data using k-means and fuzzy k-means. The results of this comparative analysis can assist in the design of clustering pipelines for MSI data analysis, and may thereby improve detection of meaningful biological patterns.

II. Methods

Fig. 1 provides an overview of the workflow in this study.

A. Data

Two MALDI-MSI datasets of mouse brain tissue from different spatial perspectives were analyzed. The first is from a coronal perspective (spatial dimensions: 103x169, spectral dimension: 8,000 m/z values); the second is from a sagittal perspective (spatial dimensions: 104x168, spectral dimension: 8,000 m/z values).

B. Dimensionality Reduction

As previously noted, PCA is a popular technique for assessment of MSI data, and can be used as a precursor to

This research has been supported by grants from The Parker H. Petit Institute for Bioengineering and Bioscience (IBB), Johnson & Johnson, Bio Imaging Mass Spectrometry Initiative at Georgia Tech, National Institutes of Health (Bioengineering Research Partnership R01CA108468, Center for Cancer Nanotechnology Excellence U54CA119338), Georgia Cancer Coalition (Distinguished Cancer Scholar Award to MDW), Microsoft Research, the National Science Foundation (GRFP to CDK), P.E.O. international (Scholar Award to CDK) and the Georgia Institute of Technology Undergraduate Research Opportunities Program (PURA travel award to SS).

S. Sarkari is with the Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta GA 30332.

C. D. Kaddi is with the Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta GA 30332.

R. V. Bennett is with the School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta GA 30332.

F. M. Fernández is with the School of Chemistry and Biochemistry, Georgia Institute of Technology, Atlanta GA 30332.

M. D. Wang is with the Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta GA 30332 (e-mail: maywang@bme.gatech.edu)

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

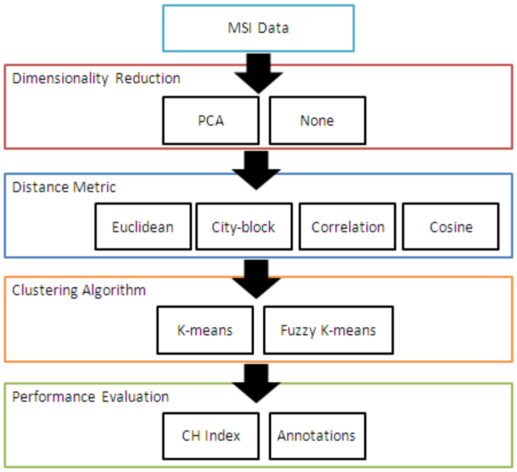

Figure 1. Flow chart showing workflow: the effects of dimensionality reduction, distance metrics, and clustering algorithm were examined.

performed clustering analysis. In this study, clustering analyses were performed both with and without a preceding PCA step. When PCA was performed, the number of principal components retained was chosen to explain 90% of the variance in the datasets.

C. *Clustering Algorithms and Distance Metrics*

K-means and fuzzy k-means clustering are examined in this study. K-means works by partitioning the dataset into a pre-specified number (*k*) of clusters. Each observation (data point) is assigned to the cluster closest to it, as measured by a specified distance metric. In this study, four distance metrics were investigated with k-means: Euclidean distance, city-block (Manhattan) distance, correlation distance, and cosine distance. In fuzzy k-means, the probability that an observation belongs to every possible cluster is determined. Euclidean distance was used with the fuzzy k-means algorithm. For both approaches, nine values of *k* ranging between two and 10 were examined.

D. *Performance Evaluation*

Considering the two different clustering algorithms, four distance metrics, and application of dimensionality reduction, 10 different analysis pipelines were considered. First, we examined how these differences influence cluster quality, as measured by the Calinski-Harabasz (CH) index [10]. This is an intrinsic evaluation that measures cluster quality. The index is defined as the normalized ratio of the between- and within-group sums of squares. Higher values of the index indicate more well-defined clusters.

Second, we evaluated the performances of different clustering analysis pipelines in terms of identifying meaningful biological patterns. Manual annotation of the two MSI datasets yielded a set of 21 m/z images representing the predominant spatial patterns in the data. These were used as references to compare the effects of the clustering algorithms, distance metrics, and dimensionality reduction. For a given value of *k*, the clustering results were represented as *k* binary images. For each set of clustering results, the spatial correspondence of each cluster image to the set of 21 m/z images was evaluated by calculating the Pearson correlation between the binary cluster image and the binary m/z image. Otsu's method was applied to generate binary m/z images. Then, each of these 21 m/z images was associated with the maximum value from among the *k* correlation values calculated for it. High correlation values indicate that the output of a given pipeline corresponds to biologically meaningful spatial patterns.

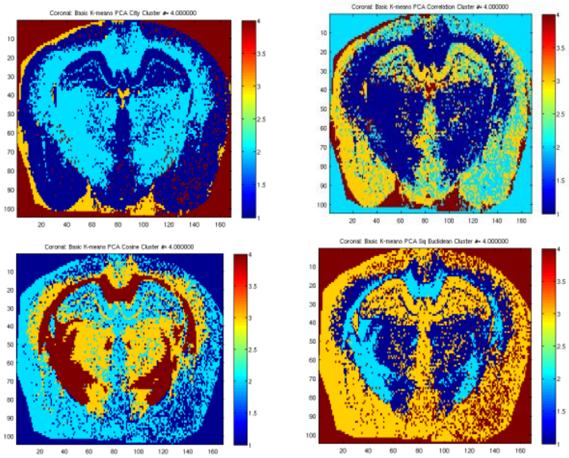
Coronal: Basic K-means PCA City Cluster #= 4.000000
Coronal: Basic K-means PCA Correlation Cluster #= 4.000000
Coronal: Basic K-means PCA Cosine Cluster #= 4.000000
Coronal: Basic K-means PCA Sq Euclidean Cluster #= 4.000000
10
20
30
40
50
60
70
80
90
100
20
40
60
80
100
120
140
160
1
1.5
2
2.5
3
3.5
4

Figure 2. K-means clustering results (*k* = 4) using the coronal dataset. Top: city-block (l), correlation (r), Bottom: cosine (l), and Euclidean (r).

III. RESULTS

Fig. 2 shows an example clustering results from k-means, for *k* = 4, from the four different distance metrics, following PCA. Although the number of clusters is held constant, different spatial structures are highlighted in these results, indicating the effect of the distance metric. In this example, Euclidean distance and correlation appear to reveal finer structural details than city-block and cosine distance.

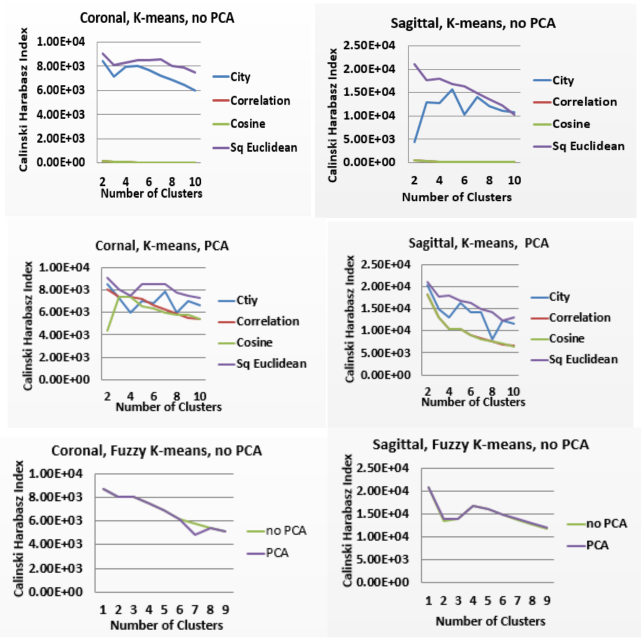

Figure 3. Variations in the Calinski-Harabasz index across different clustering pipelines in the coronal and sagittal datasets.

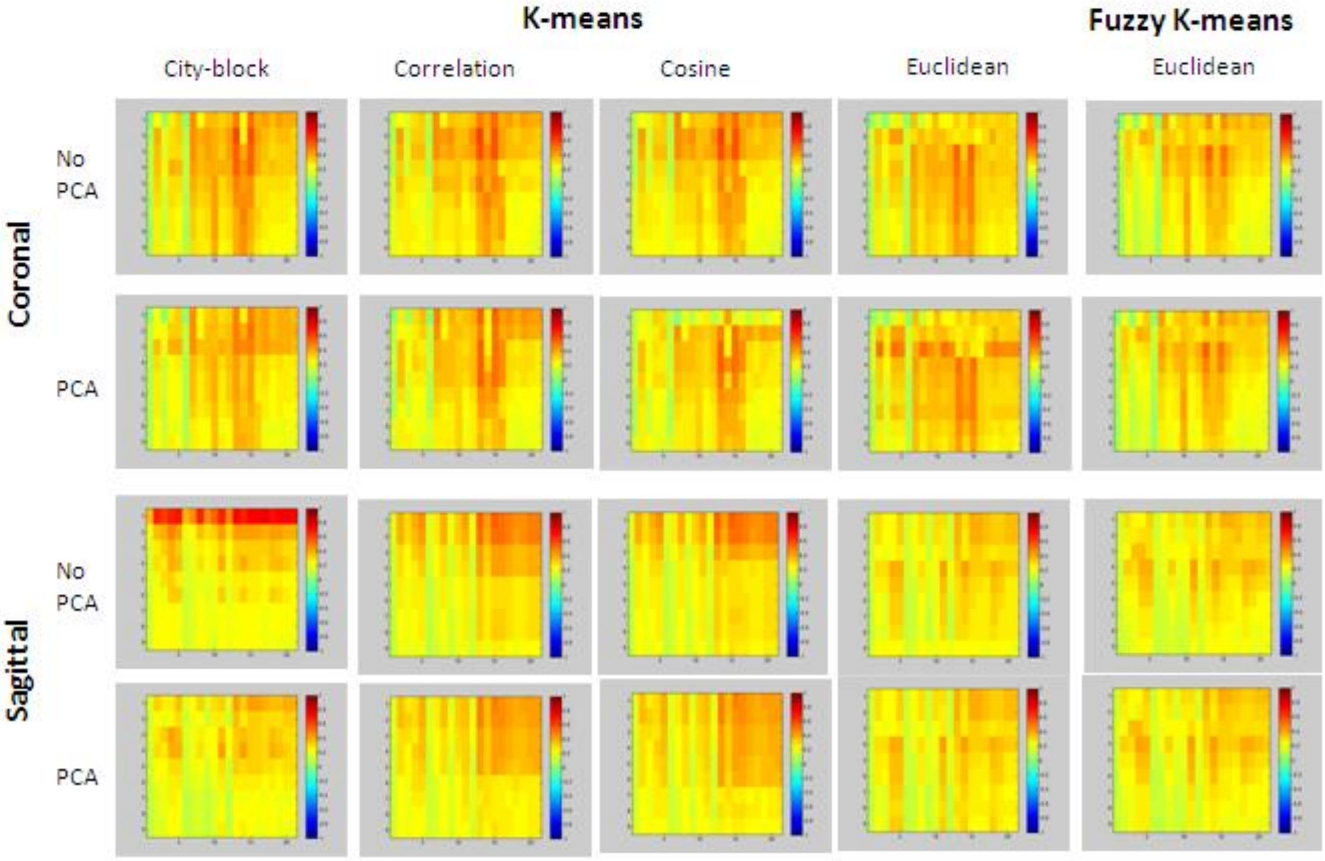

Figure 4. Mean correlation of the binary cluster images from 10 pipelines with reference m/z images selected via manual annotation. In each panel, the horizontal axis represents the 21 reference m/z images, and the vertical axis represents different values of *k* (top: *k* = 2; bottom: *k* = 10).

Fig. 3 demonstrates how the CH index differs among the clustering pipelines for different values of *k* in the two datasets. As expected, for almost all of the pipelines, *k* = 2 yields the highest CH index, since this typically identifies the major variation: tissue vs. non-tissue regions. Euclidean distance was consistently associated with the highest CH index values, while those for the cosine and correlation distances were lower. However, as shown from the previous clustering result, this does not necessarily indicate that cosine and correlation distances do not yield meaningful spatial patterns. For k-means, dimensionality reduction via PCA generally increased the CH index for higher numbers of clusters. This is important because higher numbers of clusters could identify more detailed structural patterns. However, for fuzzy k-means with Euclidean distance, PCA did not have a notable effect.

Next, clustering results from each pipeline were evaluated by comparing them with m/z values known to be representative of the dataset. These results are shown in Fig. 4. As previously noted, nine possible values of *k* were tested, ranging from two to 10. Each of these *k* images was correlated with each of the 21 representative m/z images chosen by manual annotation. There are 20 panels of size 9x21 shown in Fig 4., each describing a particular clustering pipeline on one of the two datasets. Thus, for each row (*k* value) in a single 9x21 panel, the entries represent the maximum correlation of any of the *k* images with the 21 m/z images. We are interested in seeing whether certain clustering pipelines yield clusters which are highly correlated to the set of reference m/z images; this would be indicated by more reddish regions in the panels. Overall, smaller values of *k* (located near the top of the panels) were associated with higher correlation values. This means that when fewer clusters were generated, they tended to be more highly correlated to one or more of the m/z images. Additionally, the correlation values tended to be higher with some m/z images in particular, as shown by the vertical lines observed in some of the reddish regions. In both datasets, PCA increased the correlation values for Euclidean distance. The results for the cosine and correlation distances were mixed. In the sagittal dataset, implementation of PCA appeared to increase the correlation values for larger values of *k* (located further down on the panels). In the coronal dataset, PCA decreased correlation values for smaller *k*. For city-block distance, in both datasets PCA appeared to slightly reduce the correlation values. In fuzzy k-means, again, PCA did not appear to have a notable effect.

IV. Discussion

In this paper, we present a comparative analysis of the effects of dimensionality reduction and distance metric choice on the results of k-means clustering for MSI datasets. Results on experimental MALDI-MSI data showed that these clustering pipeline parameters have a notable effect on the broad and fine physiological structures highlighted by the clusters. Additionally, they affected cluster quality as measured by the Calinski-Harabasz index. Finally, clusters from alternative pipelines were compared to a set of 21 m/z images that represent the predominant spatial patterns in the datasets, and several patterns were observed.

The results of this study so far provide several observations on the effects of different components in the analysis pipeline. First, the effects of PCA are mixed. Applying PCA prior to clustering improved cluster quality as measured by the Calinski-Harabasz index. However, for some distance metrics, it actually decreased the correlation of the cluster images to the set of reference m/z images. Second, the Euclidean distance consistently led to the highest cluster quality in terms the Calinski-Harabasz index. However, since k-means clustering minimizes the within-cluster sum of square distances when using the Euclidean distance metric, repeating this analysis with other cluster quality metrics will be useful. The Euclidean distance also yielded some high correlation values with the reference images, particularly for the coronal dataset. However, the other three distance metrics also yielded high correlation values. Notably, several of these high correlation values were for different reference m/z images (columns in the panels) than those highlighted by Euclidean distance. This observation indicates that it could be valuable to identify distance metrics that tend to yield complementary results.

We identify several routes for improvement and further research on this topic. First, the observations thus far do not lead to conclusive recommendations for a clustering analysis pipeline. These observations have been made based on analysis of two MALDI-MSI datasets of the same biological subject. Stronger conclusions could be obtained after further testing on a range of MSI datasets, including both other MALDI-MSI datasets and those from other ionization modalities, such as DESI. Additionally, testing on MSI datasets of synthetic (non-biological) samples which contain very specific spatial patterns for reference could further improve assessment. Third, the effects of the variables of interest could be tested using other cluster quality measures.

Only limited results were obtained in this study for fuzzy k-means. Comprehensive testing of fuzzy k-means with additional distance metrics is a future target. The investigation of other variants of k-means for MSI clustering, such as k-mediods and harmonic k-means, is of interest, as well as hierarchical clustering.

In conclusion, this type of comparative study can provide understanding into the strengths and weaknesses of different pipelines for clustering in MSI. Identifying pipelines which tend to perform better can then assist researchers in identifying biologically meaningful patterns through MSI.

REFERENCES

[1] K. Schwamborn and R. M. Caprioli, "Molecular imaging by mass spectrometry - looking beyond classical histology," *Nature Reviews Cancer,* vol. 10, pp. 639-646, Sep 2010.

[2] T. Alexandrov, "MALDI imaging mass spectrometry: statistical data analysis and current computational challenges," *BMC Bioinformatics,* vol. 13, Nov 2012.

[3] M. Hanselmann, M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. A. Heeren, and F. A. Hamprecht, "Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis," *Analytical Chemistry,* vol. 80, pp. 9649-9658, Dec 2008.

[4] R. M. Parry, A. S. Galhena, C. M. Gamage, R. V. Bennett, M. D. Wang, and F. M. Fernandez, "OmniSpect: An Open MATLAB-Based Tool for Visualization and Analysis of Matrix-Assisted Laser Desorption/Ionization and Desorption Electrospray Ionization Mass Spectrometry Images," *Journal of the American Society for Mass Spectrometry,* vol. 24, pp. 646-649, Apr 2013.

[5] T. Alexandrov, M. Becker, O. Guntinas-Lichius, G. Ernst, and F. von Eggeling, "MALDI-imaging segmentation is a powerful tool for spatial functional proteomic analysis of human larynx carcinoma," *Journal of Cancer Research and Clinical Oncology,* vol. 139, pp. 85-95, Jan 2013.

[6] T. Alexandrov, M. Becker, S. O. Deininger, G. Ernst, L. Wehder, M. Grasmair, F. von Eggeling, H. Thiele, and P. Maass, "Spatial Segmentation of Imaging Mass Spectrometry Data with Edge-Preserving Image Denoising and Clustering," *Journal of Proteome Research,* vol. 9, pp. 6535-6546, Dec 2010.

[7] T. Alexandrov and J. H. Kobarg, "Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering," *Bioinformatics,* vol. 27, pp. I230-I238, Jul 2011.

[8] G. McCombie, D. Staab, M. Stoeckli, and R. Knochenmuss, "Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis," *Analytical Chemistry,* vol. 77, pp. 6118-6124, Oct 2005.

[9] S. O. Deininger, M. P. Ebert, A. Futterer, M. Gerhard, and C. Rocken, "MALDI Imaging Combined with Hierarchical Clustering as a New Tool for the Interpretation of Complex Human Cancers," *Journal of Proteome Research,* vol. 7, pp. 5230-5236, Dec 2008.

[10] B. Desgraupes, "Clustering Indices". Retrieved from http://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf