Knowledge Discovery from the Hyperspectral Sky

Erzsébet Merényi
Research Professor

Department of Statistics, and
Department of Electrical and Computer Engineering
Rice University, Houston, Texas

erzsebet@rice.edu
Focus on Complexity and Discovery

- **Complexity** is a challenge in the analytics of Big Data, new algorithms are needed.

- Big Data have various complexities, not only “more” or “less”, but “different”
  - Even among different **hyperspectral data**

- **Discovery** is finding what we do not know ... can’t characterize in advance (no models) -> more / unknown complexity makes it more difficult

- **Neural maps** as tools: may be the closest analog to how the brain makes sense of big / complex data

*Hyperspectral data: fused “wide data” — in this talk all channels are used together.*
Hyperspectral imaging of terrestrial (planetary) and astronomical objects

Astronomy example

Sample image planes from ALMA Band 7, HD 142527

Ch 1 50 51 120 121 170

170 channels: C^{18}O, ^{13}CO, CS lines stacked
Spectral resolution: 0.122 MHz

Sample emission spectra

ALMA spectra from combined C^{18}O, ^{13}CO, CS lines, showing differences in composition, Doppler shift, temperature (Data credit: JVO, project 2011.0.00318.5)

Evolution of terrestrial imaging

Broad-band: Landsat TM 6 channels
MODIS 28 channels
Hyper-spectral 100-500 channels

Variations of absorptions in spectra of melting snow in response to temperature changes (speclab.cr.usgs.gov)
Knowledge Discovery from the Hyperspectral Sky

Hyperspectral imaging of terrestrial (planetary) and astronomical objects

Astronomy example

Image planes from ALMA Band 7, HD 142527

Astronomical images can have thousands of channels!

ALMA has receiver Bands 1 – 10. This sample is from one.

All channels will be used together (as n-dim pattern vectors) for analysis

Sample emission spectra

ALMA spectra from combined C^{18}O, ^{13}CO, CS lines, showing differences in composition, Doppler shift, temperature (Data credit: JVO, project 2011.0.00318.5)

Variations of absorptions in spectra of melting snow in response to temperature changes (speclab.cr.usgs.gov)
Complex (complicated) data space

- High-dimensional
- Large (number of data points)
- Multi-modal (has clusters)
- Highly structured
  - Not linearly separable
  - Widely varying shapes and sizes
  - ... densities (vary within and across clusters)
  - ... proximities
  - ... local dimensionalities

No statistical models

Hyperspectral data have many clusters with widely varying shapes, sizes, densities, proximities, local dimensionalities ...

Small hyperspectral data can also be complex and resist discovery with many methods.

Highly structured data space

Imagine in 100 dimensions!

Motivation - The First Case, from Tucson 😊
Finding olivine and pyroxene subgroups of S asteroids with Self-Organizing Maps

8-color survey (Tholen, 1984)
589 objects, in 8 spectral bands
0.3 µm 0.8 µm 1.1 µm

52-color survey (Bell et al. 1988)
117 objects, in 52 bands
2.5 µm

Previous work

- Tholen taxonomy of asteroid compositions established based on spectral shapes in 8-color survey
- Techniques used for clustering: PCA, Minimum Spanning Tree, band ratios, G-mode analysis
- Bell’s 52-color survey: extended spectral range and (hyperspectral) resolution (albeit less objects)
- Discovery of more structure was expected – but not found
  - Specifically, end groups of silicate (S) asteroids

Our SOM portrait of 8-color objects:
Matches Tholen’s taxonomy

Colors: Tholen labels
Fences: Our SOM clusters
Mislabeling
Doubtful classification

Howell, Merényi, Lebofsky, JGR 99, 1994

E. Merényi, Rice U erzsebet@rice.edu
Knowledge Discovery from the Hyperspectral Sky
Motivation - The First Case, from Tucson 😊
Finding olivine and pyroxene subgroups of S asteroids with Self-Organizing Maps

- **8-color survey (Tholen, 1984)**
  - 589 objects, in 8 spectral bands
  - 0.3 µm to 0.8 µm

- **52-color survey (Bell et al. 1988)**
  - 117 objects, in 52 bands
  - 0.8 µm to 2.5 µm

SOM portrait of **60-color objects**: Does not match Tholen’s taxonomy

Our SOM portrait of **8-color objects**: Matches Tholen’s taxonomy

Colors: Tholen labels
Fences: Our SOM clusters
Mislabeling
Doubtful classification

Pink & Purple: S types
So: Olivine rich
Sp: Pyroxene rich

Howell, Merényi, Lebofsky, JGR 99, 1994

Knowledge Discovery from the Hyperspectral Sky
None of the 76 pair-wise PC plots resolve all 16 known clusters in this data set. Colors and symbols indicate the known labels.

All clusters were found from an SOM.

Prototype-based Learning With Self-Organizing Map
Most widely used machine learning model of biological neural maps

Formation of basic SOM (Kohonen, early 80’s)

Simultaneous
- Adaptive Vector Quantization (VQ), and
- Ordering (indexing) of the prototypes in the SOM grid according to their similarities

SOM learns the structure of the data and represents it on a low-dimensional lattice, in a topology preserving fashion.

(If learning goes correctly … )

The SOM learns very well. Extraction of the prototype groups from the learned SOM is the challenge.
Structure discovery in complex data with Self-Organizing Maps

- Effective post-processing of the SOM is key to the extraction of clusters
  - The information learned by SOMs is generally underutilized for interpretation of data structure (cluster extraction)
  - Advanced / information theoretical variants are underutilized, metrics untapped.

- Exploitation of the SOM makes a difference for complex data

Side note on SOM efficiency:

Prototype-based learning produces sparse representation of data, reduces volume during learning – advantage over graphical methods for Big Data

\[ N = 10^6 \text{ data need } \sim 10^{12} \text{ graph edges; } N \rightarrow N^2 \]
\[ N = 10^6 \text{ data can be expressed by } \sim 10^3 \text{ SOM prototypes; } N \rightarrow \sqrt{N} \]
Structure / complexity of data as expressed by Voronoi tessellation and Delaunay graph

Artificial (noiseless) 2-d data, with learned SOM prototypes shown in the data space

The V-cell and D-graph structure increases from left to right.

Cannot show the V-cells / D-graph of higher-d data in data space!
D-graph and masked D-graph of the Clown
wrt 17 x 17 SOM prototypes

- The prototypes, learned by an SOM, nicely follow the data distribution
  - The prototypes are at the vertices of the D-graph

Emerging: Curve of mouth
Mass/shape of nose, eyes
Gaps between eyes and nose

Body: low density comp. to nose
The importance of SOM learning: builds masked D-graph

- **Martinetz and Schulten, Topology Representing Networks, IEEE TNN 1994:** Competitive Hebbian learning – as in neural maps - guarantees the construction of the masked D-graph of the (learned) prototypes (under one condition).

- **Easy to do:** For each data point \( v \in M \subseteq \mathbb{R}^n \) record the BMU and 2\(^{nd}\) BMU pairs (in the learned SOM)
  - these will be the connected edges of the masked D-graph (V-neighbors in data space);
  - pairs of prototypes that are not chosen together as BMU and 2\(^{nd}\) BMU by any data point, will not be connected in the D-graph.

- The generated masked D-graph can be stored as an *Adjacency matrix* \( A \) that has a 1 at \( A(i,j) \) if prototypes \( i \) and \( j \) are connected (selected together) by at least one data point.
Connectivity (CONN) graph representation

(Taşdemir & Merényi, IEEE TNN 2009)

Masked Delaunay graph - binary

“Clown” 2-d data

(Vesanto & Alhoniemi IEEE TNN 2000)

Weighted masked Delaunay graph

Adjacency

Connectivity
Connectivity (CONN) graph representation

(Taşdemir & Merényi, IEEE TNN 2009)

Masked Delaunay graph - binary

“Clown” 2-d data

(Vesanto & Alhoniemi IEEE TNN 2000)
Connectivity (CONN) graph representation & visualization in data space vs on the SOM lattice (*Taşdemir & Merényi, IEEE TNN 2009*)

A classic representation: U-matrix ($\Sigma || w_i - w_j ||$) overlain the SOM grid

CONNnectivity matrix draped over SOM grid: The SOM / CONN portrait of the Clown (hexagonal SOM lattice)

Cannot be shown for data dim > 2

Can be shown for data dim > 2

Bonus: CONN shows topology violations
Deviations from the exact 4:2:1 proportions are due to the small size of the SOM, integer arithmetic, and the formation of inter-cluster gaps.

Discovery of small clusters with “SOM magnification”
Monitoring the learning of the 8-class synthetic data with TopoView (Merényi, Taşdemir, Zhang, Springer, LNAI 5400. 2009)

Top: All topology violating connections superimposed on mU-matrix
Bottom: Same with majority truth labels overlain.

Learning of topography not yet complete but SOM state is perfect for cluster capture.
SOM vs K-means clustering of multi-spectral image (8-d spectra as input data vectors)

18 clusters, found by ISODATA (K-means)  28 SOM clusters, extracted with CONN visualization

K-means does reasonably well on these relatively low-dimensional spectral data
SOM vs K-means clustering of hyperspectral image
(196-d spectra as input feature vectors)

Large building (SOM cluster D)

Coast guard bldg (SOM cluster a)

Buildings clearly outlined

Confusion of clusters

K-means does poorly – great confusion of clusters

Merényi et al., URBAN 2007

35 SOM clusters

21 K-means clusters
Spectral Statistics of Clusters, Ocean City
196-band Hyperspectral Image of Urban Area

K-means does poorly – great confusion of clusters

K-means
21 clusters
Interesting clusters not discovered; large variance of clusters

SOM
35 clusters
Many unique spectral types, “tight clusters”
CONN vs mU-matrix for identifying SOM clusters - effect on real data of moderate complexity

Clusters obtained from an 40 x 40 SOM

Data: Ocean City, Maryland
Daedalus AADS 1260 scanner
bands 3 – 10 used
(Csathó, Krabill, Lucas and Schenk, 1998)

Clear separation of rare clusters from other clusters

Taşdemir & Merényi, IEEE TNN 2009

Rare clusters, hard to separate from this representation
ALMA hyperspectral image of HD 142527

Data credit: JVO, project 2011.0.00318.5

ALMA: Atacama Large Millimeter Array

66 dishes at ~ 5,500 m

Sample image planes from ALMA Band 7, HD 142527

170 channels: C^{18}O, ^{13}CO, CS lines stacked
Spectral resolution: 0.122 MHz

ALMA spectra from combined C^{18}O, ^{13}CO, CS lines, showing differences in composition, Doppler shift, temperature (Data credit: JVO, project 2011.0.00318.5)

Artist’s concept of planet formation in HD 142527
Data and connectivity statistics

Passport, ALMA data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># vectors</td>
<td>5,625</td>
</tr>
<tr>
<td># dim</td>
<td>170</td>
</tr>
<tr>
<td># clusters</td>
<td>? (many)</td>
</tr>
<tr>
<td>Noise</td>
<td>Moderate</td>
</tr>
<tr>
<td>Similarity</td>
<td>Variable</td>
</tr>
<tr>
<td># V neighbors</td>
<td>17</td>
</tr>
<tr>
<td># R-local</td>
<td>2</td>
</tr>
</tbody>
</table>

Passport, 6d 8-class data

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td># vectors</td>
<td>16,386</td>
</tr>
<tr>
<td># dim</td>
<td>6</td>
</tr>
<tr>
<td># clusters</td>
<td>8</td>
</tr>
<tr>
<td>Noise</td>
<td>Moderate</td>
</tr>
<tr>
<td>Similarity</td>
<td>High</td>
</tr>
<tr>
<td># V neighbors</td>
<td>20</td>
</tr>
<tr>
<td># R-local</td>
<td>2</td>
</tr>
</tbody>
</table>
Clusters from 20 x 20 SOM of ALMA image

Prototypes in the lower left corner

The extracted clusters in the SOM

CONNvis

The clusters shown in the disk

Center detail
SOM clusters of HD 142527
First-cut hyperspectral analysis of ALMA data compared to Casassus et al. (2013)

Simultaneous CS, $^{13}$CO, & C$^{18}$O

SOM clusters
- capture general Doppler structure found in single-species lines.
- incorporate line intensities, widths, shapes, et cetera, as well as Doppler.
- contain more structure than single-line analysis, and more than can be shown here.

Coloring is arbitrary, not a heat map.

Thanks: Al Wootten
Data Credit: JVO Project Code 2011.0.00318.5

Single-line Doppler CO & HCO+

Mean cluster spectra
Layered Knowledge?

Superimposed cluster maps
Conclusions

- SOMs are powerful for structure discovery in complex data
  - The CONN(ectivity) similarity metric improves the segmentation of prototypes compared to distance-based metrics

- ALMA hyperspectral data cubes – a new type of complexity
- Showed intricate structure identified in ALMA data
- Emerging structure makes good sense, but it is also more complex than CONN seems to capture from the SOM
  - Motivates further development of metrics & visualization

- New types of astronomy data can present surprises we may not be ready for and will provide exciting opportunities for CI and ML research. 😊
Thank you

This one?  Or this?

Maybe both ...
References

http://www.ece.rice.edu/~erzsebet/publications-EMerenyi.pdf


References

http://www.ece.rice.edu/~erzsebet/publications-EMerenyi.pdf


