Efficient Visualization of Large-scale Data Tables through Reordering and Entropy Minimization

Nemanja Djuric, Slobodan Vucetic
Temple University, Philadelphia
December 10th, 2013, in Dallas, Texas
Data visualization

- Immediate feedback that can lead to faster knowledge discovery
  - Intuitive way of interacting with unknown data
  - Practical even for non-experts

- Visualizing large data matrices
  - Data given in a form of a large 2-D table
  - Long history, however novel methods required to tackle emerging Big Data problems
Visualizing data tables

- Existing approaches
Data reordering

- Idea: Reorder data matrix so that similar rows and columns are grouped together.

Data reordering: Related work

- Used in bioinformatics, anthropology, archeology, ...

- Low-dimensional projection approaches
  - PCA, LLE, Spectral Clustering (SC)

- Hierarchical clustering (HC) approaches
  - HC with optimal leaf ordering

- Traveling salesman solvers
  - Lin-Kernighan heuristic

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time</th>
<th>Space</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>$O(n \log(n))$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>LLE</td>
<td>$O(n^2)$</td>
<td>$O(n^2)$</td>
</tr>
<tr>
<td>SC</td>
<td>$O(n^2)$</td>
<td>$O(n^2)$</td>
</tr>
<tr>
<td>HC</td>
<td>$O(n^2)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>HC-olo</td>
<td>$O(n^3)$</td>
<td>$O(n^2)$</td>
</tr>
<tr>
<td>LK</td>
<td>$O(n^{2.2})$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>TSP-means</td>
<td>$O(n \log(n))$</td>
<td>$O(n)$</td>
</tr>
</tbody>
</table>
EM-ordering

- Reordering from the viewpoint of data compression
  - Assume data set $D = \{x_i, i = 1, \ldots, n\}$, where $x_i = [x_{i1}, x_{i2}, \ldots, x_{im}]$ is an $m$-dimensional example
  - Task: Reorder the data so that it is maximally compressible

- Differential Predictive Coding (DPC)
  - Use local context to code the value of $x_i$

\[
D = \{x_i, i = 1, \ldots, n\} \rightarrow D_{DPC} = \{x_1, \varepsilon_2, \ldots, \varepsilon_n\}
\]

where $\varepsilon_i = (x_i - x_{i-1}), \ i = 2, \ldots, n$
EM-ordering: Intuition

Before reordering:

```
3  3
5  1
2  4
4  2
1  5
```

After reordering:

```
1  5
2  4
3  3
4  2
5  1
```

DPC
EM-ordering

- Entropy of differences used to estimate data compressibility
  - Differences independent, sampled from $N(0, \sigma_j^2), j = 1, \ldots, m$

$$H(\varepsilon) = \frac{n}{2} (m \cdot \log(2\pi) + \sum_{j=1}^{m} \log(\sigma_j(\varepsilon))) +$$

$$0.5 \sum_{i=2}^{n} \sum_{j=1}^{m} \frac{(x_{\pi(i),j} - x_{\pi(i-1),j})^2}{\sigma_j^2}$$

- Solve the following optimization problem

$$(\pi^*, \{\sigma_1^*, \ldots, \sigma_m^*\}) = \arg \min_{\pi,\{\sigma_1, \ldots, \sigma_m\}} H(\varepsilon)$$
EM-ordering

- The optimization can be split into two parts
  1. Fix variance of differences → Minimize the overall distance between neighbors in the ordering (equivalent to TSP)
  2. Fix ordering → Find variance of the differences

- Or, more formally:

```
Algorithm 1 EM-ordering

Inputs: data set $D$; initial guess for $\{\sigma_j\}_{j=1,\ldots,m}$
Output: ordered set $D$; learned $\{\sigma_j\}_{j=1,\ldots,m}$

1. repeat until convergence
2. run TSP solver for current $\sigma_j$ to find $\pi$
3. calculate $\sigma_j$ for current ordering of $D$
```
The best TSP solvers have super-quadratic time complexity.

We propose an $O(n \log(n))$ method, called TSP-means

1. Create a $2^l$-ary tree through recursive runs of $k$-means ($k = 2$)
2. Traverse the tree breath-first, and solve TSP defined on children of the current node and their immediate neighbors.
Results: Synthetic data set

- Synthetic 2-D data set with data points located on two concentric circles of different radii

![Original](image1)
![PCA](image2)
![LLE](image3)
![SC](image4)
![HC](image5)
![HC-olo](image6)
![LK](image7)
![TSP-means](image8)
Results: Waveform data set

Figure of Merit scores are given in the parentheses:

\[ FOM(\pi) = \frac{1}{n-1} \sum_{i=1}^{n-1} I(y(\pi(i)) \neq y(\pi(i+1))) \]
Results: Real-world applications

- Minneapolis traffic data set

- Original data set
- Reordered data set
- Locations of the sensors
Results: Real-world applications

- Stocks data set

Original data set

Reordered data set
Conclusion

- Inadequacy of standard visualization tools in large-scale setting is apparent
  - Novel methods required to address Big Data problems

- EM-ordering and TSP-means
  - Fast, efficient knowledge discovery
  - Easily parallelizable
  - Interesting results on real-world data

- Future work
  - Binary, categorical data?
  - Development of an easy-to-use visualization software
Thank you!

- Questions?
LK vs. TSP-means

- Effect of user-set parameter $l$
  - Global vs. local solution

(a) LK solution
(b) TSP-means solution