Exploiting Sparsity in the Statistical Analysis of Gene Expression Data

Anirban Chatterjee
Padma Raghavan
Francesca Chiaromonte

The Pennsylvania State University

Supported by NSF through grants NSF-STHEC 0444345 and NSF CSR-SMA 0720749.
(Chatterjee, Bhowmick, Raghavan, Textmining at SDM 2008, longer version under review)
Outline

- Introduction
- Force-directed Graph Embedding
- Feature Subspace Transformation (FST)
- FST-K-Means Clustering
- Clustering Gene Expression Data
- Summary
Introduction

- Scientific dataset sizes growing rapidly with high throughput instruments, especially, in *Life Sciences*.
- FST-K-Means: Fast yet improved clustering by utilizing sparsity and structure. (Chatterjee, Bhowmick, Raghavan, Textmining at SDM 2008, longer version under review)
  - **SPARSITY**
    - An mxn matrix is **sparse** if the number of nonzeros is $O(m)$ or $O(n)$.
    - Many feature values are either zero or numerically insignificant.
    - Observed data are **sparse** and high-dimensional.
  - **STRUCTURE**
    - Similarity between observations indicates a relationship.
    - Relationship induces a **structure** in the data.
- **Feature subspace transformation**: Combines sparsity and structure through embedding in high-dimensional feature space.
Force-directed Graph Embedding

Fruchterman & Reingold Graph Drawing
(Prior work)

\[
\Delta d_{fr} = \sum \frac{u - v}{d_{uv}} fa_{uv} + \sum \frac{u - v}{d_{uv}} fr_{uv}
\]

where, \( d_{uv} = \text{Euclidean distance between document } u \text{ and } v \)

\[
fa_{uv} = \frac{d_{uv}^2}{k} \quad fr_{uv} = -\frac{k^2}{d_{uv}} \quad k = C \sqrt{\frac{\text{area}}{\text{number of vertices}}}
\]

Force-directed Graph Embedding

(Our modifications)

\[ \Delta d_{fr} = \sum \frac{u - v}{d_{uv}} fa_{uv} \times \frac{W_{uv}}{\text{iter}_i} \]

Calculated for vertices connected by an edge \( O(E) \)

where, \( d_{uv} = \) Euclidean distance between document \( u \) and \( v \)

\[ fa_{uv} = \frac{d_{uv}^2}{k} \quad k = C \sqrt{\frac{\text{area}}{\text{number of vertices}}} \]

During embedding only non-zero term entries are modified, i.e. entity vectors modified only in active dimensions or terms.
Computational Cost of Modified Embedding

- Gain in time complexity.
- Fruchterman-Reingold computational costs using attractive + repulsive forces is \( O(V^2 + E) \)
- Our approach
  - No repulsive force calculation
  - Reduced costs: \( O(E) \)
FST: Feature Subspace Transformation

- **Entity-Feature matrix**
  - An \( N \times R \) sparse matrix \( A \) of \( N \) entities and \( R \) features.
  - Each entry \( a_{ij} \) is the number of times feature \( t_j \) appears in entity \( d_i \).

- **Entity Graph**
  - An undirected graph \( G = (V,E) \) with \( N \) entities and \( E \) edges between these entities.
  - Edge weight \( e_{ij} \) is the number of common features between the two entities.
FST: Main Steps
Cluster Quality Metrics

Cluster A

Cluster B

Cluster Centroid “c”

Entity “u”

\( d_{uc} \)

\( d_{AB} \)
Quality Metrics: Internal

- External Quality Metric
  - Accuracy

\[ P = \frac{\mathcal{E}}{N} \]

- where,
  \( \mathcal{E} \rightarrow \) Number of correctly classified documents
  \( N \rightarrow \) Total number of documents
Quality Metrics: Internal

- **Internal Quality Metrics**
  - Measure of intra-cluster cohesiveness

\[
J(M_1, M_2 \cdots M_k) = \sum_{M_{ij} \in M_1} \left\| M_{ij} - w_1 \right\|^2 + \cdots + \sum_{M_{ij} \in M_k} \left\| M_{ij} - w_k \right\|^2
\]

- where,
  - \( M_c \rightarrow \text{Set of documents belonging to cluster } c \)
  - \( w_k \rightarrow \text{Centroid of cluster } M_c \)
  - \( M_{ij} \rightarrow \text{jth document of cluster } M_i \)
  - \( k \rightarrow \text{Total number of clusters} \)

Text and Gene Datasets

Text Datasets
- Features are words
- Multinomial distribution
- Feature extraction a difficult process
- Number of features less than number of observations
- Data acquisition is simple

Gene Expression Datasets
- Features are biological attributes (e.g., GC content)
- Difficult to decide distribution
- Feature extraction based on domain knowledge
- Number of features larger than observations
- Complicated data acquisition

Sparse Data
Datasets

<table>
<thead>
<tr>
<th>Name</th>
<th>Samples</th>
<th>Features</th>
<th>Source (Type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>adult_a2a</td>
<td>2,265</td>
<td>123</td>
<td>UCI (Census)</td>
</tr>
<tr>
<td>australian</td>
<td>690</td>
<td>14</td>
<td>UCI (Credit Card)</td>
</tr>
<tr>
<td>breast-cancer</td>
<td>683</td>
<td>10</td>
<td>UCI (Census)</td>
</tr>
<tr>
<td>dna</td>
<td>2,000</td>
<td>180</td>
<td>Starlog (Medical)</td>
</tr>
<tr>
<td>splice</td>
<td>1,000</td>
<td>60</td>
<td>Delve (Medical)</td>
</tr>
<tr>
<td>180txt</td>
<td>180</td>
<td>19,698</td>
<td>SMART (Text)</td>
</tr>
<tr>
<td>300txt</td>
<td>300</td>
<td>53,914</td>
<td>SMART (Text)</td>
</tr>
<tr>
<td>20news</td>
<td>1,061</td>
<td>16,127</td>
<td>Yahoo Newsgroup (Text)</td>
</tr>
</tbody>
</table>

Table 1: Test suite of datasets
FST-K-Means on DNA SPLICE

Figure 2: (a) Original data projected to first three principal components (b) Embedded data projected to first three principal components
FST-K-Means on DNA SPLICE

We observe an improvement of 25.27% in clustering accuracy relative to K-Means
## Classification Accuracy

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Classification Accuracy (P)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K-Means</td>
<td>MLKM</td>
</tr>
<tr>
<td>adult_a2a</td>
<td>70.60</td>
<td>52.49</td>
</tr>
<tr>
<td>australian</td>
<td><strong>85.51</strong></td>
<td>74.20</td>
</tr>
<tr>
<td>breast-cancer</td>
<td>93.70</td>
<td>69.69</td>
</tr>
<tr>
<td>dna</td>
<td>72.68</td>
<td>70.75</td>
</tr>
<tr>
<td>splice</td>
<td>55.80</td>
<td>53.20</td>
</tr>
<tr>
<td>180txt</td>
<td>73.33</td>
<td>91.67</td>
</tr>
<tr>
<td>300txt</td>
<td>78.67</td>
<td>64.33</td>
</tr>
<tr>
<td>20news</td>
<td>46.74</td>
<td>54.85</td>
</tr>
</tbody>
</table>

Table 2: Accuracy of classification of K-Means, MLKM and FST-K-Means
# Cluster Cohesiveness

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Cluster Cohesiveness (J)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K-Means</td>
<td>MLKM</td>
<td>FST-K-Means</td>
</tr>
<tr>
<td>Adult_a2a</td>
<td>24,013</td>
<td>16,665</td>
<td>16,721</td>
</tr>
<tr>
<td>australian</td>
<td>4,266</td>
<td>3,034</td>
<td>2,638</td>
</tr>
<tr>
<td>breast-cancer</td>
<td>2,475</td>
<td>2,203</td>
<td>1,366</td>
</tr>
<tr>
<td>dna</td>
<td>84,063</td>
<td>65,035</td>
<td>65,545</td>
</tr>
<tr>
<td>splice</td>
<td>31,883</td>
<td>31,618</td>
<td>31,205</td>
</tr>
<tr>
<td>180txt</td>
<td>25,681</td>
<td>23,776</td>
<td>24,131</td>
</tr>
<tr>
<td>300txt</td>
<td>47,235</td>
<td>44,667</td>
<td>45,052</td>
</tr>
<tr>
<td>20news</td>
<td>3,851,900</td>
<td>3,483,591</td>
<td>3,341,400</td>
</tr>
</tbody>
</table>

Table 3: Cluster cohesiveness of K-Means, MLKM and FST-K-Means
# X Chromosome Inactivation (XCI) Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Genes (E/I)</th>
<th>Features</th>
<th>Non-zeros</th>
</tr>
</thead>
<tbody>
<tr>
<td>50kb</td>
<td>399 (346/53)</td>
<td>248</td>
<td>28480</td>
</tr>
<tr>
<td>100kb</td>
<td>365 (318/47)</td>
<td>248</td>
<td>42655</td>
</tr>
<tr>
<td>250kb</td>
<td>315 (278/37)</td>
<td>248</td>
<td>57444</td>
</tr>
</tbody>
</table>
Results: XCI

Average “Improvements”:
- accuracy: 24.40%
- precision: 13.83%
- recall: 25.04%
- cohesiveness: 3.94%
Why does FST work?

Optimal clustering and bounds on cohesiveness:

\[ Y = \text{centered data matrix} \]
\[ \text{(Mean of data } A \text{ subtracted from each entity } a_i) \]
\[ N \bar{y}^2 = \text{trace of } Y^TY \]
\[ \lambda_i = i-th \text{ principal eigenvalue of } Y^TY \]

\[
N \bar{y}^2 - \sum_{i=1}^{k-1} \lambda_i \leq J \leq N \bar{y}^2
\]

Cohesiveness after FST-K-Means

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Lower Bound</th>
<th>Min</th>
<th>Max</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult_a2a</td>
<td>15,250</td>
<td>15,866</td>
<td>17,208</td>
<td>17,380</td>
</tr>
<tr>
<td>australian</td>
<td>2,442</td>
<td>2,638</td>
<td>3,008</td>
<td>3,493</td>
</tr>
<tr>
<td>breast-cancer</td>
<td>747</td>
<td>1,366</td>
<td>1,366</td>
<td>5,169</td>
</tr>
<tr>
<td>dna</td>
<td>63,890</td>
<td>65,525</td>
<td>65,865</td>
<td>67,190</td>
</tr>
<tr>
<td>splice</td>
<td>30,389</td>
<td>31,205</td>
<td>31,205</td>
<td>31,942</td>
</tr>
<tr>
<td>180txt</td>
<td>23,188</td>
<td>23,765</td>
<td>24,178</td>
<td>25,290</td>
</tr>
<tr>
<td>300txt</td>
<td>43,708</td>
<td>44,800</td>
<td>45,194</td>
<td>46,512</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Lower Bound</th>
<th>Min</th>
<th>Max</th>
<th>Upper Bound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adult_a2a</td>
<td>15,250</td>
<td>24,013</td>
<td>24,409</td>
<td>17,380</td>
</tr>
<tr>
<td>australian</td>
<td>2,442</td>
<td>4,266</td>
<td>4,458</td>
<td>3,493</td>
</tr>
<tr>
<td>breast-cancer</td>
<td>747</td>
<td>2,475</td>
<td>2,475</td>
<td>5,169</td>
</tr>
<tr>
<td>dna</td>
<td>63,890</td>
<td>84,062</td>
<td>84,123</td>
<td>67,190</td>
</tr>
<tr>
<td>splice</td>
<td>30,389</td>
<td>31,882</td>
<td>31,884</td>
<td>31,942</td>
</tr>
<tr>
<td>180txt</td>
<td>23,188</td>
<td>25,651</td>
<td>25,730</td>
<td>25,290</td>
</tr>
<tr>
<td>300txt</td>
<td>43,708</td>
<td>47,220</td>
<td>47,288</td>
<td>46,512</td>
</tr>
</tbody>
</table>

FST-K-Means satisfies the optimality bounds while K-Means fails to do so.
Questions?
Thank You
Email: achatte@cse.psu.edu