Comparative study of dimension reduction techniques for microarray gene expression data

Christoph Bartenhagen¹, Hans-Ulrich Klein¹, Christian Ruckert¹, Xiaoyi Jiang², Martin Dugas¹

¹Department of Medical Informatics and Biomathematics, University of Münster
²Department of Computer Science, University of Münster

GMDS-Jahrestagung 2010
Hochschule Mannheim

Motivation

- Quality assessment
- Data interpretation
- Hypothesis generation
- Reveal special patterns (e.g. cluster)
- Noise and outlier detection
Motivation

- Quality assessment
- Data interpretation
- Hypothesis generation
- Reveal special patterns (e.g., cluster)
- Noise and outlier detection
1. Dimension reduction:
   - Linear: Principal Component Analysis (PCA)
   - Nonlinear: Kernel PCA (KPCA), Isomap (IM), Maximum Variance Unfolding (MVU), Diffusion Maps (DM), Locally Linear Embedding (LLE), Laplacian Eigenmaps (LEM)
   - Focus on two and three dimensional visualization

2. Benchmark:
   - Ten microarray datasets + simulated data
   - Classification (SVM)
   - Cluster validation (Davis-Bouldin-Index)

3. Results:
   - Benchmark results for 1 of 10 microarray datasets
1. Dimension reduction:
   - Linear: Principal Component Analysis (PCA)
   - Nonlinear: Kernel PCA (KPCA), Isomap (IM), Maximum Variance Unfolding (MVU), Diffusion Maps (DM), Locally Linear Embedding (LLE), Laplacian Eigenmaps (LEM)
   - Focus on two and three dimensional visualization

2. Benchmark:
   - Ten microarray datasets + simulated data
   - Classification (SVM)
   - Cluster validation (Davis-Bouldin-Index)

3. Results:
   - Benchmark results for 1 of 10 microarray datasets

Linear vs. nonlinear dimension reduction techniques

Linear:
Introduction

Linear vs. nonlinear dimension reduction techniques

Linear:

Nonlinear:

Nonlinear dimension reduction techniques

Global approach:
- Maximum Variance Unfolding
- Diffusion Maps
- Isomap

Local approach:
- Laplacian Eigenmaps
- Locally Linear Embedding
Nonlinear dimension reduction techniques

Global approach:

- Maximum Variance Unfolding
- Diffusion Maps
- Isomap

Local approach:

- Laplacian Eigenmaps
- Locally Linear Embedding
Nonlinear dimension reduction techniques

Global approach:
- Maximum Variance Unfolding
- Diffusion Maps
- Isomap

Local approach:
- Laplacian Eigenmaps
- Locally Linear Embedding

Datasets

<table>
<thead>
<tr>
<th>Samples</th>
<th>Features</th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al. - Breast cancer</td>
<td>286</td>
<td>ER+</td>
<td>ER-</td>
</tr>
<tr>
<td>Verhaak et al. - Leukemia</td>
<td>863</td>
<td>NPMI pos</td>
<td>NPMI neg</td>
</tr>
<tr>
<td>Haberlach et al. - Leukemia</td>
<td>261</td>
<td>NPMI pos</td>
<td>NPMI neg</td>
</tr>
<tr>
<td>Haferlach et al. - Leukemia</td>
<td>77</td>
<td>AML with t(8;21)</td>
<td>AML with t(15;17)</td>
</tr>
<tr>
<td>Chiaretti et al. - Leukemia</td>
<td>72</td>
<td>TET2</td>
<td>AML</td>
</tr>
<tr>
<td>Alizadeh et al. - Lymphoma</td>
<td>36</td>
<td>Activated B-like DLBCL</td>
<td>GC B-like DLBCL</td>
</tr>
<tr>
<td>NuLi et al. - High-grade glioma</td>
<td>50</td>
<td>GBM</td>
<td>Anaplastic oligodendroglioma</td>
</tr>
<tr>
<td>Alon et al. - Colon cancer</td>
<td>82</td>
<td>Tumor</td>
<td>Normal</td>
</tr>
<tr>
<td>Singh et al. - Prostate cancer</td>
<td>90</td>
<td>Tumor</td>
<td>Normal</td>
</tr>
<tr>
<td>Simulated microarray data (100x)</td>
<td>50</td>
<td>10,000</td>
<td></td>
</tr>
</tbody>
</table>
Datasets

<table>
<thead>
<tr>
<th>Samples</th>
<th>Features</th>
<th>Class 1</th>
<th>Class 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al. - Breast cancer</td>
<td>286</td>
<td>ER+</td>
<td>ER-</td>
</tr>
<tr>
<td>Verhaak et al. - Leukemia</td>
<td>461</td>
<td>NPM1 pos.</td>
<td>NPM1 neg.</td>
</tr>
<tr>
<td>Haferlach et al. - Leukemia</td>
<td>251</td>
<td>NPM1 pos.</td>
<td>NPM1 neg.</td>
</tr>
<tr>
<td>Haferlach et al. - Leukemia</td>
<td>71</td>
<td>AML with t(8;21)</td>
<td>AML with t(15;17)</td>
</tr>
<tr>
<td>Haferlach et al. - Leukemia</td>
<td>72</td>
<td>ALL</td>
<td>AML</td>
</tr>
<tr>
<td>Lefrak et al. - Leukemia</td>
<td>22</td>
<td>t(8;21)</td>
<td>t(15;17)</td>
</tr>
<tr>
<td>Alizadeh et al. - Lymphoma</td>
<td>36</td>
<td>Activated B-like DLBCL</td>
<td>GCB-like DLBCL</td>
</tr>
<tr>
<td>Mott et al. - High-grade glioma</td>
<td>50</td>
<td>GBM</td>
<td>Anaplastic oligodendroglioma</td>
</tr>
<tr>
<td>Al-Shaikh et al. - Colon cancer</td>
<td>62</td>
<td>Tumor</td>
<td>Normal</td>
</tr>
<tr>
<td>Leng et al. - Prostate cancer</td>
<td>100</td>
<td>Tumor</td>
<td>Normal</td>
</tr>
<tr>
<td>Simulated microarray data (x100)</td>
<td>50</td>
<td>10.000</td>
<td>10.000</td>
</tr>
</tbody>
</table>

Benchmark

Gene expression data, Labels

- Estimate parameters
- Reduce dimension
- Classification
- Module validation
- Accuracy
- Cluster validation

Dimension reduction

- PCA, KPCA, LLE, M. Neigh., LEM, CM, MVU
- Gradient descent
- SVM & kernel parameter
- Train SVM
- Predict low dimensional test data
- Accuracies
- Cluster distances

Loo-cv with SVM: target dimension, neighbors, kernel parameter

- Separate training data
- Separate test data

- Grad. descent
- SVM & kernel parameter
- Reduce dimension
- Classification
- Dataset validation
Classification

### Estimated parameters

<table>
<thead>
<tr>
<th>Method</th>
<th>Dimension</th>
<th>Neighbours</th>
<th>loo-cv accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA</td>
<td>14</td>
<td></td>
<td>87.4</td>
</tr>
<tr>
<td>KPCA</td>
<td>15</td>
<td>5x5</td>
<td>87.1</td>
</tr>
<tr>
<td>LLE</td>
<td>12</td>
<td>14</td>
<td>88.5</td>
</tr>
<tr>
<td>IM</td>
<td>8</td>
<td>10</td>
<td>85</td>
</tr>
<tr>
<td>IM(mod)</td>
<td>15</td>
<td>4</td>
<td>87.4</td>
</tr>
<tr>
<td>LEM</td>
<td>5</td>
<td>4</td>
<td>85.3</td>
</tr>
<tr>
<td>DM</td>
<td>13</td>
<td>5x5</td>
<td>84.3</td>
</tr>
<tr>
<td>MVU</td>
<td>5</td>
<td>14</td>
<td>85</td>
</tr>
</tbody>
</table>

### Cluster validation

Cluster validation results for PCA, KPCA, LLE, IM, IM(mod), LEM, DM, and MVU methods. The DB-index values are shown for different target dimensions.
Noise evaluation

Simulated data

Introduction  Methods  Results  Conclusions
Most significant information can be captured well in very few dimensions; even for data with only a small number of (20 – 50) samples

Nonlinear dimension reduction methods yield better results in 2 & 3 dimensions than PCA

PCA and Diffusion Maps respond least sensitive to noise

LLE and Isomap performed best in classification and cluster validation on most datasets

LLE and Isomap need less differential features for good classification than PCA

LLE and Isomap in R:

http://bioconductor.org/packages/2.7/bioc/html/RDRToolbox.html
Thank you very much for your attention!