Accuracy of Distance Metric Learning Algorithms

Aurélien Sérandour

École Polytechnique
DMMT’09

June 28th, 2009
Motivation & problem

- recommendation problem in feature space
Motivation & problem

- recommendation problem in feature space
- use a simple Mahalanobis distance
Motivation & problem

- recommendation problem in feature space
- use a simple Mahalanobis distance
- real data are not perfect: try with imperfect ones
Many algorithms $\rightarrow$ only 4 chosen
Many algorithms → only 4 chosen

<table>
<thead>
<tr>
<th></th>
<th>Euclidean</th>
<th>Xing</th>
<th>Coding Similarity</th>
<th>ITML</th>
</tr>
</thead>
<tbody>
<tr>
<td>non iterative</td>
<td>iterative</td>
<td>non iterative</td>
<td>iterative</td>
<td>iterative</td>
</tr>
<tr>
<td>do nothing</td>
<td>minimize distance between similar pairs</td>
<td>shared information</td>
<td>respect thresholds</td>
<td></td>
</tr>
<tr>
<td>no data used</td>
<td>both</td>
<td>similar</td>
<td>both</td>
<td>infinite loops</td>
</tr>
<tr>
<td></td>
<td>infinite loops</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
- datasets have properties: size, dimensionality, # of classes.
Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
- datasets have properties: size, dimensionality, # of classes.
- create similarity sets
Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
- datasets have properties: size, dimensionality, # of classes.
- create similarity sets
- lots of dissimilar data
Experiments

- using 6 UCI datasets: Iris, Ionosphere, Wine, WDBC, Soybean-small, Balance-scale
- datasets have properties: size, dimensionality, # of classes.
- create similarity sets
- lots of dissimilar data
- inject some errors because no perfect-real datasets → flipping similarities (noise)
Experiments

Figure: Java application
### Results

<table>
<thead>
<tr>
<th>noise</th>
<th>Iris</th>
<th>Ionosphere</th>
<th>Wine</th>
<th>WDBC</th>
<th>Soybean</th>
<th>Balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>ITML</td>
<td>Xing</td>
<td>CS</td>
<td>ITML</td>
<td>CS</td>
<td>ITML</td>
</tr>
<tr>
<td>5</td>
<td>ITML</td>
<td>Xing</td>
<td>CS</td>
<td>Xing</td>
<td>ITML</td>
<td>CS</td>
</tr>
<tr>
<td>10</td>
<td>ITML</td>
<td>CS</td>
<td>CS</td>
<td>Euclid.</td>
<td>ITML</td>
<td>CS</td>
</tr>
<tr>
<td>20</td>
<td>Euclid.</td>
<td>CS</td>
<td>Xing</td>
<td>Euclid.</td>
<td>ITML</td>
<td>CS</td>
</tr>
<tr>
<td>30</td>
<td>Euclid.</td>
<td>CS</td>
<td>Xing</td>
<td>Euclid.</td>
<td>Euclid.</td>
<td>CS</td>
</tr>
</tbody>
</table>

**Table:** Best algorithms
Conclusion

- there is no unique solution
- difficult to predict the behavior
- data drives the result
- use other kind of distances?
Thank you!