Self-organising maps

Map high-dimensional data to a 2D grid of “units” according to similarity/distance (Kohonen, 1982).

“Spatially smooth version of k-means” (Ripley, PRNN, 1996).

Training SOMs

Data: 177 Italian wines

Object 1

Initial state
Training SOMs

Object 1

Winner 1

Algorithm:
- Pick random object
- Determine winner in map
- Update winner and environment
- Periodically, decrease environment and learning rate

R code:

```r
> library(kohonen)
> data(wines)
> somnet <- som(scale(wines), gr = somgrid(5, 5),rlen=100)
> plot(somnet, "codes")
```
Supervised SOMs

- use of all information
- better reproducibility
- better interpretability
- better predictions


---

Supervised SOMs

- use of all information
- better reproducibility
- treat Y as a special (set of) variables
- separate range scaling of distances in X and Y
- better interpretability
- explicit weighting of distances in X and Y
- better predictions
- for regression as well as classification


---

Supervised SOMs

- use of all information
- better reproducibility
- treat Y as a special (set of) variables
- separate range scaling of distances in X and Y
- better interpretability
- explicit weighting of distances in X and Y
- better predictions
- for regression as well as classification


---

X-ray powder patterns

Descriptor of crystal structure: similar patterns should correspond to similar structures

Self-organising maps for powder patterns

- Supervised and unsupervised mapping
- Special similarity function (WCC) with one parameter: triangle width

Mapping using cell volume

```r
xyfnet <- xyf(X[training,,] Y[training,]
+ gr = somgrid(20, 20, "hexagonal"),
+ rlen = 250, xweight = .5)
plot(xyfnet, "predict")
```

Training time: 1 h 20' (P 3.2GHz)

Data set: steroids

<table>
<thead>
<tr>
<th>Space group</th>
<th># compounds</th>
<th>label</th>
</tr>
</thead>
<tbody>
<tr>
<td>P212121</td>
<td>978</td>
<td>19</td>
</tr>
<tr>
<td>P21</td>
<td>843</td>
<td>4</td>
</tr>
<tr>
<td>P1</td>
<td>93</td>
<td>5</td>
</tr>
<tr>
<td>C2</td>
<td>99</td>
<td>1</td>
</tr>
</tbody>
</table>

Total: 2013

Training set (1342 compounds) and a test set (671 compounds).

Mapping using space group

```
sompredictions <-
+ predict(somnet, trainY = classvec2classmat(Yc[training]))
plot(somnet, "property",
+ property = sompredictions$unit.predictions)
plot(xyfnet, "predict")
```
Prediction results (test set)

Volume prediction (correlation coefficients)

<table>
<thead>
<tr>
<th>Seed 7</th>
<th>Seed 13</th>
<th>Seed 31</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM</td>
<td>.01</td>
<td>-.04</td>
</tr>
<tr>
<td>XYF (class only)</td>
<td>.36</td>
<td>.41</td>
</tr>
<tr>
<td>XYF (class and volume)</td>
<td>.72</td>
<td>.28</td>
</tr>
</tbody>
</table>

Space group prediction (percentage correct)

<table>
<thead>
<tr>
<th>Seed 7</th>
<th>Seed 13</th>
<th>Seed 31</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM</td>
<td>43%</td>
<td>43%</td>
</tr>
<tr>
<td>XYF (class only)</td>
<td>87%</td>
<td>86%</td>
</tr>
<tr>
<td>XYF (class and volume)</td>
<td>79%</td>
<td>46%</td>
</tr>
</tbody>
</table>

Conclusions

- SOMs (supervised and unsupervised) are ideally suited for analysing databases of chemical structures
- Special distance measures can/must be used
- Supervised SOMs have many advantages: better predictions, easier to interpret, and better stability
- Training can take a long time but mapping is relatively fast
- Including space group information is important in predicting properties of crystals

Acknowledgements

- Library 'class' by B.D. Ripley
- Edwards & Oman, RNews 3(3), 2003
- René de Gelder
- Willem Melssen
- Egon Willighagen