Supervised Self-Organising Maps

Ron Wehrens

Institute of Molecules and Materials, IMM

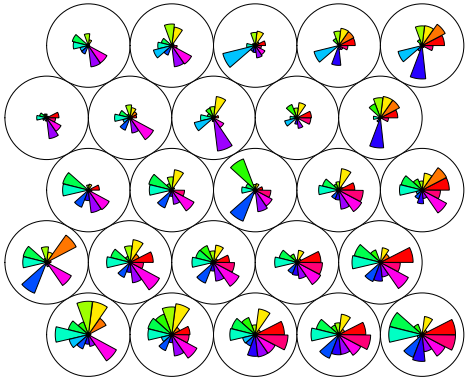
Radboud University

Nijmegen, The Netherlands

Self-organising maps

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

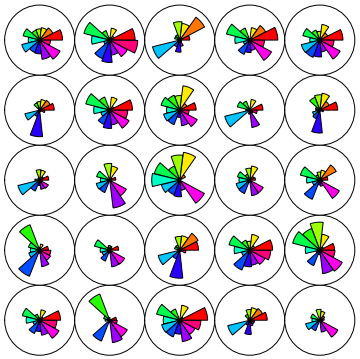
“Spatially smooth version of k-means” (Ripley, PRNN, 1996).



Training SOMs

Data: 177 Italian wines

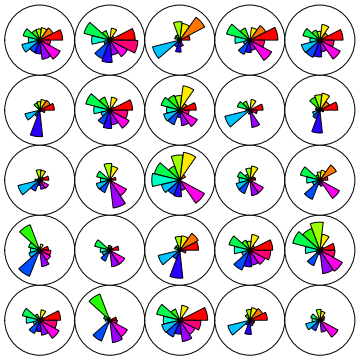
Initial state



Training SOMs

Object 1

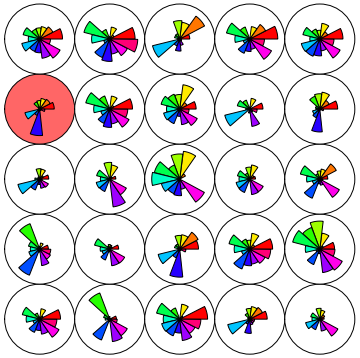
Initial state



Training SOMs

Winner 1

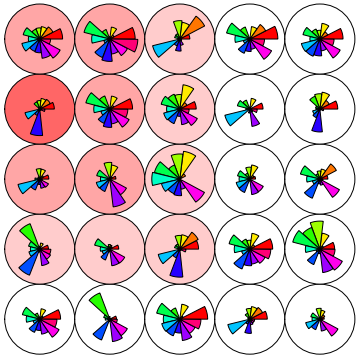
Object 1



Training SOMs

Update 1

Object 1



Training SOMs

Algorithm:

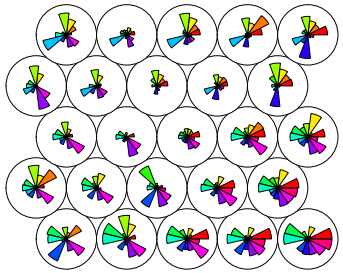
- Pick random object
- Determine winner in map
- Update winner and environment
- Periodically, decrease environment and learning rate

R code:

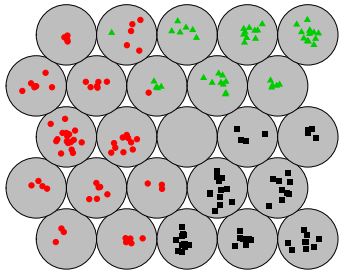
```
> library(kohonen)
> data(wines)
> somnet <- som(scale(wines), gr = somgrid(5, 5), rlen=100)
> plot(somnet, "codes")
```

Mapping

Wines: codebook vectors



mapping



Supervised SOMs

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

W.J. Melssen, R. Wehrens and L.M.C Buydens, *Chemom. Intell. Lab. Syst.* (2006), *in press*.

Supervised SOMs

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

- treat Y as a special (set of) variables
- separate range scaling of distances in X and Y
- explicit weighting of distances in X and Y
- for regression as well as classification

W.J. Melssen, R. Wehrens and L.M.C Buydens, *Chemom. Intell. Lab. Syst.* (2006), *in press*.

Supervised SOMs

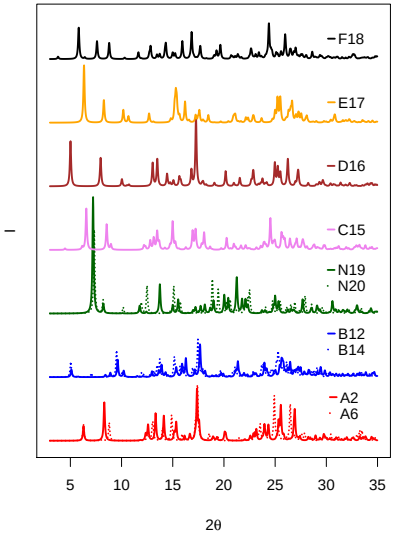
- use of all information
- better reproducibility
- better interpretability
- better predictions

- treat Y as a special (set of) variables
- separate range scaling of distances in X and Y
- explicit weighting of distances in X and Y
- for regression as well as classification

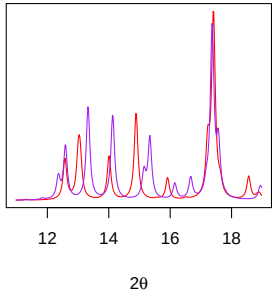
```
> library(kohonen)
> data(wines)
> xyfnet <- xyf(scale(wines), classvec2classmat(wine.classes),
                gr = somgrid(5, 5), rlen=100, xweight = .5)
```

W.J. Melssen, R. Wehrens and L.M.C Buydens, *Chemom. Intell. Lab. Syst.* (2006), *in press*.

X-ray powder patterns

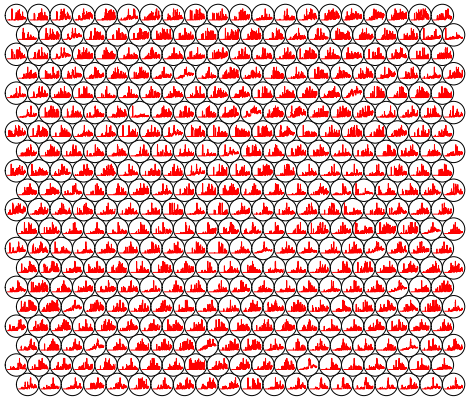


Descriptor of crystal structure:
similar patterns should
correspond to similar structures

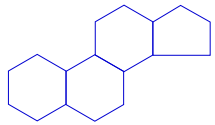


Package wccsom

- Self-organising maps for powder patterns
- Supervised and unsupervised mapping
- Special similarity function (WCC) with one parameter: triangle width



Data set: steroids



Space group	# compounds	label
P212121	978	19
P21	843	4
P1	93	5
C2	99	1
Total	2013	

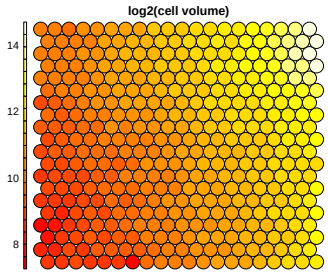
Training set (1342 compounds) and a test set (671 compounds).

Mapping using cell volume

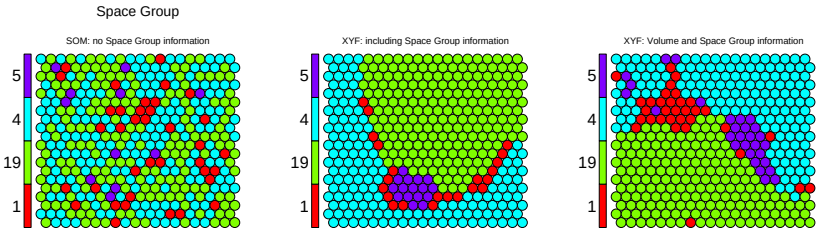
```
> xyfnet <- xyf(X[training,], Y[training],
+               gr = somgrid(20, 20, "hexagonal"),
+               rlen = 250, xweight = .5)
> plot(xyfnet, "predict")
```

Training time:

1 h 20' (P 3.2GHz)



Mapping using space group




```
> sompredictions <-
+    predict(somnet, trainY = classvec2classmat(Ycl[training]))
> plot(somnet, "property",
+      property = sompredictions$unit.predictions)
> plot(xyfnet, "predict")
```

Prediction results (test set)

Volume prediction (correlation coefficients)

	Seed 7	Seed 13	Seed 31
SOM	.01	-.04	.01
XYF (class only)	.36	.41	.41
XYF (class and volume)	.72	.28	.68

Space group prediction (percentage correct)

	Seed 7	Seed 13	Seed 31
SOM	43%	43%	24%
XYF (class only)	87%	86%	85%
XYF (class and volume)	79%	46%	66%

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