Supplement on data analysis and statistical methods

Self-organizing maps (SOMs)

SOMs were introduced by Teuvo Kohonen in 1982 (1) and constitute self-organized formation of topologically correct feature maps (2) as a simple model for biological neural networks. They are a form of unsupervised artificial neural networks and learn statistical patterns in a recursive manner. SOMs consist of a (usually two-dimensional) grid of learning units (which are also called nodes or neurons). Each node i has an associated reference vector, w_i, which stands for the pattern represented in the specific node.

Throughout the SOM learning algorithm, data records are presented to the SOM. The best-matching node for a record x is determined by evaluating the Euclidean distance $|w_i - x|$ to all the reference vectors. Now, the best-matching node j and it's neighboring nodes (on the two dimensional grid) learn the information of the new data record by adapting their weight vectors towards the direction of x, namely

$$w_i(t+1) = w_i(t) + \alpha(t) \cdot h_{ji}(t) \cdot (x - w_i(t))$$

where h_{ji} is the neighboring function of node i to the best-matching node j of the data vector x and α is the learning rate of the SOM. In our case we have

$$h_{ji}(t) = exp\left(-\frac{|z_j - z_i|}{2\sigma(t)}\right)$$

where z_i and z_j are the coordinates of the corresponding nodes on the two-dimensional grid. The radius σ is called tension.

By presenting all the available data vectors to the SOM, the winning nodes and their neighbours will be attracted to the corresponding parts of the multi-dimensional data distribution. By doing this repeatedly, the nodes represent the data distribution increasingly well, and neighboring nodes on the grid will represent similar parts of the data. We obtain a topology-preserving, two-dimensional representation of the multi-dimensional data distribution.

The nodes of the trained SOM can be interpreted as micro-clusters, which group very similar data records (here: patients) into the representing nodes. The SOM then interpolates through the data distribution like a non-linear regression surface. Figuratively speaking, a SOM can be seen as an elastic membrane having an inner tension, defined by the neighboring function, that moves freely within the data space, and becomes "attracted" to individual data vectors, respectively dense regions in the data space, until it reaches an equilibrium in which the inner tension and the forces pulling the nodes to the individual data vectors balance out.

Since adjacent nodes of the trained SOM correspond to neighboring regions in the data space, properties of the data distribution can be viewed by visualizing single variables over the SOM and comparing them with each other. This way, correlations and dependencies can easily be observed. In addition, the representation of the SOM allows clustering algorithms to be applied to the data. By clustering the frequency weighted SOM nodes. This has the advantage that clustering techniques can be used, which were not feasible using the original data due to runtime issues and susceptibility to statistical noise.

Learning parameters of the SOM model

The most precise training schedule of Viscovery SOMine 7.1, “Accurate” was used to calculate the SOM model on a hexagonal grid of 21 rows and 23 columns (where every second row has one node less, such that the SOM consists of 472 nodes).

Both, the learning rate α and the time dependence of the tension σ, are managed internally by Viscovery SOMine to guarantee optimal convergence properties. Only the final tension can be set by the user. In our model, a tension of 0.5 was used, which corresponds to a weak remaining interaction between directly adjacent nodes and a small smoothing effect by the SOM.

The following list of 58 biomarkers was chosen to constitute the data space for the training of the SOM: A2macro, AAT, Adiponectin, ANG2, B2M, BDNF, Bilirubin, C3, Eotaxin2, EPO, FactorVII, Ferritin, FetuinA, Fibrinogen, Haptoglobin, HDL, hsCRP, ICAM1, IFN, IgA, IGFBP1, IGFBP2, IgM, IL12p40, IL1beta, IL1ra, IL23, IL2Ralpha, IL6, IL8, IP10, ITAC, LAPTGFB1, LDL, Leptin, Leucocytes, MCP1, MCP4, MIF, MMP3, MMP9, Myoglobin, OCL1, ONN1, OPG, OPN1, PAI1, RANTES, SAA, SCF, SICAM1, SVCAM1, TARC, TIMP1, TNF-R1, TNF-R2, VDBP, and VEGF.

Hence, only these variables have been given a weight > 0 to influence the order of the resulting SOM and the cluster model. All other variables mentioned are mapped onto the nodes after the ordering process has been finished. All weighted variables have been scaled to equal variance in order to make the scales comparable and assigned the same weight, namely 1, to contribute equally to the order of the SOM. Eventual correlations between weighted variables have been compensated internally by Viscovery SOMine, adapting the weights accordingly to get a more balanced map.

Clustering method

Based on the created SOM model, clusters were generated using the SOM-Ward Cluster algorithm of Viscovery, a hybrid algorithm that combines the local ordering information of the map, i.e. the SOM topology, with Ward's classical hierarchical cluster algorithm producing very cohesive clusters.

The method begins by defining each individual node as a separate cluster. In each step of the algorithm, two clusters with minimal distance according to the SOM-Ward distance measure are merged. This measure heeds the Ward distances as well as the positioning of two clusters in the map picture by defining that the distance of non-adjacent clusters is always infinite, limiting merging to topologically neighboring clusters. (3)

References

(1) Kohonen, T. (2001). Self-Organizing Maps. Springer Series in Information Sciences, vol 30. Springer, Berlin, Heidelberg.

(2) Kohonen, T. (1982), Self-organized formation of topologically correct feature maps. *Biol. Cybern.* **43**, 59–69 (1982)

(3) Viscovery Software GmbH, Manual of Viscovery SOMine 7.1