4 The Median SOM

Teuvo Kohonen

The model vectors $\mathbf{m}_i$ in the basic SOM were determined as conditional averages over selected subsets of samples $\mathbf{x}(t)$. Thereby, however, sharp structures in the patterns formed by the components of $\mathbf{x}(t)$ will be smoothed out. Also, if the SOM is used to represent sets of statistical descriptors of a discrete set of items (cf. Sec. 11), the models will no longer be exact replica of any descriptor sets.
An alternative way for the construction of the SOM is to use the Batch Map principle, but instead of updating the old models $\mathbf{m}_i(t)$ by the respective means over the unions of the Voronoi sets, one can take the so-called *set medians* over the unions for the updated values of the $\mathbf{m}_i(t)$.
The set median M over the set $S = \{X(t)\}$ is defined to be that member of S, the sum of distances of which from all the other elements of S is minimum:

$$\sum_t d[X(t), M] = \min ! \tag{53}$$

The reason for calling M the "median" is that if the $X(t)$ are scalar numbers, and if $d[X(t), M] = |X(t) - M|$, then M is easily seen to be the arithmetic median of S. However, in the most general case, the $X(t)$ need not even be vector-valued.

The Batch Map algorithm is thus modified in the following:

1. Initialize the models M_i in some proper way.
2. Input all the available samples $X(t)$, and list each of them under the respective winner unit c (for which $d[X(t), M_i]$ is minimum). If there are several winners for $X(t)$, select one of them at random for listing.
3. Take for the updated value of M_i the median over the neighborhood N_c, i.e., over the union of the above lists associated with the winner unit c and its neighborhood N_c. If there are several medians, select one of them at random for the effective median.
4. Repeat from 2 a few times, until the values of the M_i can be regarded as steady.

In order to speed up the computations, if the usual SOM or Batch Map algorithm is applicable to the $X(t)$, it may be advisable to first construct the SOM in the traditional way and after that continue using the median algorithm.
As the set median is a replica of some of the members, the internal structures of this member will be preserved in the mapping, and thus every model resulting in the SOM will always represent some real input sample.