Adaptive Resonance Theory ART-2A algorithm

Automatic unsupervised classification for open-categorical problems

Literature

Primary	G.A. Carpenter, S. Grossberg and D.B. Rosen, Neural Networks **4** (1991) 493-504
Secondary	D. Wienke et al., Chemometrics and Intelligent Laboratory Systems **25** (1994) 367-387

Brief description

The algorithm clusters *n* data vectors $\vec{x}_1,...,\vec{x}_n$, each containing *m* components $x_{i1},...,x_{im}$, by grouping them into an a priori unspecified number of clusters, guided by the vigilance parameter ρ_{max}.

A priori component scaling

The values of the *m* components $x_{i1},...,x_{im}$ of the *n* data vectors $\vec{x}_1,...,\vec{x}_n$ should be mapped to interval [0,1], so that all components get the same significance. To do this, the respective minimal $x_{1,\min},...,x_{m,\min}$ and maximal components $x_{1,\max},...,x_{m,\max}$ can be determined with all components being scaled to interval [0,1] by using

$$x_{ij}^{(\text{scaled})} = \begin{cases} \dfrac{x_{ij} - x_{j,\min}}{x_{j,\max} - x_{j,\min}} & \text{for} \quad x_{j,\max} > x_{j,\min} \\ x_{ij} - x_{j,\min} & \text{for} \quad x_{j,\max} = x_{j,\min} \end{cases}$$

Starting point

Starting point are n data vectors $\vec{x}_1,...,\vec{x}_n$ with m components $x_{i1},...,x_{im}$ each ($x_{ij} \geq 0$: only positive, real values including zero are permitted). The data vectors can be combined to form a data matrix $\underline{\underline{X}}$ with n rows and m columns where each row corresponds to a data vector.

Initialization

- Set vigilance parameter ρ_{max}:

  $$0 < \rho_{max} < 1$$

  A ρ_{max} value close to zero leads to a coarse grouping (with few clusters), whereas a ρ_{max} value close to one leads to a fine grouping (with many clusters).

- Instantiate cluster matrix $\underline{\underline{W}}$ with $c_{cluster,max}$ rows and m columns. Parameter $c_{cluster,max}$ should be larger than the largest expected number of clusters to be formed. The individual rows of the cluster matrix $\underline{\underline{W}}$ will later be the individual cluster vectors $\vec{w}_k$, each with m components.

- Set parameters …

  … threshold for contrast enhancement (default value of offset a is 0.5)

  $$0 < \theta < \frac{1}{\sqrt{m}} \quad ; \quad e.g. \quad \theta = \frac{1}{\sqrt{m+a}} \quad ; \quad a > 0$$

  …learning rate ($\eta < 0.5$, default value is 0.01)

  $$0 < \eta << 1$$

  … scaling factor (default value of offset a is 0.5)

  $$\alpha < \frac{1}{\sqrt{m}} \quad ; \quad e.g. \quad \theta = \frac{1}{\sqrt{m+a}} \quad ; \quad a > 0$$

Training

- Randomly select a data vector $\vec{x}_i$ from data matrix $\underline{\underline{X}}$. Repeat this step until the randomly selected data vector is not a vector with length zero. Then normalize the data vector:

$$\vec{x}_i^0 = \frac{\vec{x}_i}{|\vec{x}_i|} \quad ; \quad |\vec{x}_i| = \sqrt{\sum_{j=1}^{m} x_{ij}^2}$$

- For contrast enhancement, all components of the normalized data vector $\vec{x}_i^0$ are transformed with a nonlinear threshold function. The transformed vector $\vec{y}_i$ is then normalized again:

$$y_{ij} = \begin{cases} x_{ij}^0 \ \text{für} \ x_{ij}^0 > \theta \\ 0 \ \text{für} \ x_{ij}^0 \leq \theta \end{cases}$$

$$\vec{y}_i^0 = \frac{\vec{y}_i}{|\vec{y}_i|}$$

Components of vector $\vec{x}_i^0$, that are small in magnitude, are suppressed (noise suppression). Therefore, components that are small in magnitude but relevant should be scaled in advance (see "A priori component scaling" above).

- If there are no clusters (first pass: $c_{cluster} = 0$), vector $\vec{y}_i^0$ is transferred to the cluster matrix $\underline{\underline{W}}$ and forms the first cluster: $\vec{w}_1 = \vec{y}_i^0$; $c_{cluster}^{new} = 1$. The new cluster vector $\vec{w}_1$ is the first row of cluster matrix $\underline{\underline{W}}$.

- If clusters already exist ($c_{cluster} \geq 1$), a maximum ρ_{winner} is determined:

$$\rho_{winner} = \max(\rho_i)$$

$$\rho_i = \alpha \sum_{j=1}^{m} y_{ij}^0 \quad \text{and} \quad \rho_i = \vec{y}_i^0 \cdot \vec{w}_k \quad \text{with} \quad k = 1,\ldots,c_{cluster}$$

Note: $\rho_i = \vec{y}_i^0 \cdot \vec{w}_k = |\vec{y}_i^0| \cdot |\vec{w}_k| \cos(\phi) = \cos(\phi)$ *(* $\vec{y}_i^0$ *and* $\vec{w}_k$ *are unit vectors,* ϕ *is the angle between both vectors). Since all vectors obey* $x_{ij} \geq 0$*:*
$0 \leq \phi \leq 90^\circ \ \Rightarrow \ 0 \leq \cos(\phi) \leq 1$.

If $\rho_{winner} = \alpha \sum_{j=1}^{m} y_{ij}^0$, the number of clusters is increased by one: $c_{cluster}^{new} = c_{cluster}^{old} + 1$. Vector $\vec{y}_i^0$ is transferred to the new cluster vector $\vec{w}_{c_{cluster}^{new}}$: $\vec{w}_{c_{cluster}^{new}} = \vec{y}_i^0$.

No assignment to one of the existing cluster vectors was convincing.

For $\rho_{winner} = \vec{y}_i^0 \cdot \vec{w}_{k_{winner}}$, ρ_{winner} is compared to ρ_{max}: For $\rho_{winner} < \rho_{max}$ the number of clusters is increased by one: $c_{cluster}^{new} = c_{cluster}^{old} + 1$. Vector $\vec{y}_i^0$ is transferred to new cluster vector $\vec{w}_{c_{cluster}^{new}}$: $\vec{w}_{c_{cluster}^{new}} = \vec{y}_i^0$. For $\rho_{winner} \geq \rho_{max}$ the number of clusters remains unchanged, but the winning cluster vector $\vec{w}_{k_{winner}}$ is modified as follows:

$$\vec{w}_{k_{winner}}^{new} = \vec{s}$$

$$\vec{s} = \frac{\vec{t}}{\left|\vec{t}\right|}$$

$$\vec{t} = \vec{u} + (1-\eta)\vec{w}_{k_{winner}}^{old}$$

$$\vec{u} = \eta \frac{\vec{v}}{\left|\vec{v}\right|}$$

$$v_j = \begin{cases} y_{ij}^0 & \text{für } w_{k_{winner}j}^{old} > \theta \\ 0 & \text{für } w_{k_{winner}j}^{old} \leq \theta \end{cases}$$

The learning rate η determines the incremental learning. For $\eta = 0$, all $\vec{w}_k$ remain constant forever, so that there is no incremental learning. For $\eta = 1$, all $\vec{w}_k$ are directly forgotten and only the new vector $\vec{y}_i^0$ is learned. The learning rate η mediates between these two extremes. The threshold vector $\vec{v}$ ensures that a feature, that has once fallen below the threshold θ, can never be learned again (stabilization).

- All training steps are repeated until the cluster matrix $\underline{\underline{W}}$ shows no significant changes after one epoch (i.e., after all *n* data vectors $\vec{x}_i$ have each run through the training phase once in random order), i.e., the individual cluster vectors $\vec{w}_k$ remain practically constant. This can be checked by the scalar product of old and new cluster vectors being above a convergence threshold ε (default value is 0.99)

$$\vec{w}_i^{old} \cdot \vec{w}_i > \varepsilon \quad \text{with} \quad i = 1,...,c_{cluster} \quad \text{and} \quad 0 < \varepsilon < 1$$

Alternatively, the training steps can be repeated until the assignment of the individual data vectors $\vec{x}_i$ to their respective clusters remains unchanged after one epoch.

Clustering (last pass)

- Choose a data vector $\vec{x}_i$ from the data matrix $\underline{\underline{X}}$. If it is a vector with length zero, assign it to the zero cluster and choose a new data vector, otherwise normalize it: $\vec{x}_i^0 = \frac{\vec{x}_i}{|\vec{x}_i|}$.

- Transform $\vec{x}_i^0$ with the nonlinear threshold function and normalize the transformed vector $\vec{y}_i$ again:

$$y_{ij} = \begin{cases} x_{ij} & \text{für } x_{ij} > \theta \\ 0 & \text{für } x_{ij} \leq \theta \end{cases}$$

$$\vec{y}_i^0 = \frac{\vec{y}_i}{|\vec{y}_i|}$$

- Determine the maximum ρ_{winner}:

$$\rho_{winner} = \max(\rho_i)$$

$$\rho_i = \vec{y}_i^0 \cdot \vec{w}_k \text{ mit } k = 1,...,c_{cluster}$$

- Data vector $\vec{x}_i$ belongs to cluster k_{winner} which is determined by $\rho_{winner} = \vec{y}_i^0 \cdot \vec{w}_{k_{winner}}$.