SELF - ORGANIZING MAP and DATA MINING (KDD)
in BIOMEDICAL SCIENCE

A. Mossolov

Innovazione Tecnologica e Ricerca Applicata
ITERA S.r.l., Castelletto Stura, ITALIA
Primary Goal of Data Mining

**Prediction** - using some variables to predict unknown values of other variables

**Description** - finding human-interpretable patterns describing the data
Data Mining Tasks

- Classification
- Clustering
- Summarization
- Regression
- Dependency Modeling
- Deviation Detection
Popular Data Mining Methods

- Decision Trees and Rules
- Nonlinear Regression
- Pattern based Methods
- Probabilistic Graphical models
- Relational Learning Models
- . . .
... is one of the simplest neural network KDD models which implements of unsupervised competitive learning.

Self-Organizing map (SOM) introduced by prof. T. Kohonen (Finland) in 1981.
SELF-ORGANIZING MAP - an idea

“Experimental” data
N-dimensional space

“Human” 2D plane
SOM algorithm:

- Geometry definition
- Initialization (Pre-Processing)
- Map Training
- Testing (Post Processing)
Structure of a SOM
SOM algorithm:

1. Geometry definition
SOM algorithm:

2. PreProcessing

- Component Scaling
  \[ x_i \rightarrow \bar{x}_i = \frac{x_i - m_i}{\sqrt{s_i}} \]

- Map Initialization
3. Training

1) $t := 0$: initialize $w_{ij}$ randomly for $i = 1, \ldots, m$, $j = 1, \ldots, n$

2) Choose input vector $x \in X$ randomly in the training set

3) Determine the neuron $i^*$ such that its weight vector $w_{i^*}$ is the closest to the input vector

$$\|w_{i^*}(t) - x\| = \min \{ \|w_i(t) - x\| \} \quad \text{for all } i$$

4) Update the weight vectors $w_i$, $i = 1, \ldots, m$:

$$w_i(t+1) := w_i(t) + \eta(t) A(i, i^*, t) (x - w_i(t))$$

5) Increment the time $t := t + 1$

5) if for several time intervals $\sqrt{\sum_{i=1}^{m} \left| \Delta w_{ij} \right|^2} > \varepsilon$ then

   go to 2 else STOP

$H$-function
SOM algorithm:

3. Training

Formally the BMU is defined as the neuron for which

$$\| \mathbf{x} - \mathbf{m}_w \| = \min_z \| \mathbf{x} - \mathbf{m}_z \|,$$

where $\| \cdot \|$ is the distance measure.

The SOM update rule for the weight vector of the unit $i$ is:

$$\mathbf{m}_i(t+1) = \mathbf{m}_i(t) + h_{\psi_i}(t)[\mathbf{x}(t) - \mathbf{m}_i(t)],$$

where $t$ denotes time.
SOM algorithm: 3. Training
SOM algorithm: 4. Visualization - Testing

U - matrix
(distance matrix)
SOM algorithm: 4. Visualization - Testing

Vector component visualization
SOM algorithm: 4. Visualization - Testing

Cluster visualization
### SOM algorithm: Example

#### ANIMALS “DATABASE”

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>small</td>
<td>medium</td>
<td>big</td>
<td>2legs</td>
<td>4legs</td>
<td>hair</td>
<td>hooves</td>
<td>mane</td>
<td>feathers</td>
<td>hunt</td>
<td>run</td>
<td>fly</td>
<td>swim</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>dove</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>hen</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>duck</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>goose</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>owl</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>hawk</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>eagle</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>fox</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>dog</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>wolf</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>cat</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>tiger</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>lion</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>horse</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>zebra</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>cow</td>
</tr>
</tbody>
</table>
SOM algorithm: Example - Initialization

SOM 12 x 8
SOM algorithm: Example - Training
SOM algorithm: Example - Testing
SOM algorithm: Example - Testing
SOM Example - "Biomed" dataset
(American Statistical Association, 1982)

SOM 30 x 25

209 observation:
134 “normals”
75 “carriers”

Initial

Trained
SOM Example - “Biomed” dataset
(American Statistical Association, 1982)
SOM Example - “Biomed” dataset
Headache and/or facial pain (“HFpain”) data collection includes a standardized questionnaire.

- The questionnaire is divided in the following sections:
- family history,
- physiological history (and, for women, the gynecological history), remote and recent medical history.
- General and neurological examination.
- Two psychometric tests: MMPI and STAI.

Totally the “HFpain” database contains 258+928 clinical headache and/or facial pain cases plus 400 cases of control groups. Each case in database is represented as a complex record with 50 fields. First 7 fields contain general information (ID, name, age, sex, diagnosis, etc.), other 26 fields contain in binary form answers on the standardized questionnaire questions (yes=1, no=0), the last fields contains information relay to MMPI and STAI psychometric tests. So, in general, the “HFpain” database has dimension 1586x50.
SOM Example

HF pain data base

Trained SOM Map
SOM Example

HFpain data base

Trained SOM Map
SOM Example

HF pain data base

Trained SOM Map
SOM Example

HFpain data base

Trained SOM Map
SOM Example

HFpain data base

Trained SOM Map - second group
**SOM Example**

"Protein" data base

<table>
<thead>
<tr>
<th>Country</th>
<th>RedMeat</th>
<th>WhiteMeat</th>
<th>Eggs</th>
<th>Milk</th>
<th>Fish</th>
<th>Cereals</th>
<th>Starch</th>
<th>Nuts</th>
<th>Fr&amp;Veg</th>
</tr>
</thead>
<tbody>
<tr>
<td>Albania</td>
<td>10.1</td>
<td>1.4</td>
<td>0.5</td>
<td>8.9</td>
<td>0.2</td>
<td>42.3</td>
<td>0.6</td>
<td>5.5</td>
<td>1.7</td>
</tr>
<tr>
<td>Austria</td>
<td>8.9</td>
<td>14.0</td>
<td>4.3</td>
<td>19.9</td>
<td>2.1</td>
<td>28.0</td>
<td>3.6</td>
<td>1.3</td>
<td>4.3</td>
</tr>
<tr>
<td>Belgium</td>
<td>13.5</td>
<td>9.3</td>
<td>4.1</td>
<td>17.5</td>
<td>4.5</td>
<td>26.6</td>
<td>5.7</td>
<td>2.1</td>
<td>4.0</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>7.8</td>
<td>6.0</td>
<td>1.6</td>
<td>8.3</td>
<td>1.2</td>
<td>56.7</td>
<td>1.1</td>
<td>3.7</td>
<td>4.2</td>
</tr>
<tr>
<td>Czechoslovakia</td>
<td>9.7</td>
<td>11.4</td>
<td>2.8</td>
<td>12.5</td>
<td>2.0</td>
<td>34.3</td>
<td>5.0</td>
<td>1.1</td>
<td>4.0</td>
</tr>
<tr>
<td>Denmark</td>
<td>10.6</td>
<td>10.8</td>
<td>3.7</td>
<td>25.0</td>
<td>9.9</td>
<td>21.9</td>
<td>4.8</td>
<td>0.7</td>
<td>2.4</td>
</tr>
<tr>
<td>E Germany</td>
<td>8.4</td>
<td>11.6</td>
<td>3.7</td>
<td>11.1</td>
<td>5.4</td>
<td>24.6</td>
<td>6.5</td>
<td>0.8</td>
<td>3.6</td>
</tr>
<tr>
<td>Finland</td>
<td>9.5</td>
<td>4.9</td>
<td>2.7</td>
<td>33.7</td>
<td>5.8</td>
<td>26.3</td>
<td>5.1</td>
<td>1.0</td>
<td>1.4</td>
</tr>
<tr>
<td>France</td>
<td>18.0</td>
<td>9.9</td>
<td>3.3</td>
<td>19.5</td>
<td>5.7</td>
<td>28.1</td>
<td>4.8</td>
<td>2.4</td>
<td>6.5</td>
</tr>
<tr>
<td>Greece</td>
<td>10.2</td>
<td>3.0</td>
<td>2.8</td>
<td>17.6</td>
<td>5.9</td>
<td>41.7</td>
<td>2.2</td>
<td>7.8</td>
<td>6.5</td>
</tr>
<tr>
<td>Hungary</td>
<td>5.3</td>
<td>12.4</td>
<td>2.9</td>
<td>9.7</td>
<td>0.3</td>
<td>40.1</td>
<td>4.0</td>
<td>5.4</td>
<td>4.2</td>
</tr>
<tr>
<td>Ireland</td>
<td>13.9</td>
<td>10.0</td>
<td>4.7</td>
<td>25.8</td>
<td>2.2</td>
<td>24.0</td>
<td>6.2</td>
<td>1.6</td>
<td>2.9</td>
</tr>
<tr>
<td>Italy</td>
<td>9.0</td>
<td>5.1</td>
<td>2.9</td>
<td>13.7</td>
<td>3.4</td>
<td>36.8</td>
<td>2.1</td>
<td>4.3</td>
<td>6.7</td>
</tr>
<tr>
<td>Netherlands</td>
<td>9.5</td>
<td>13.6</td>
<td>3.6</td>
<td>23.4</td>
<td>2.5</td>
<td>22.4</td>
<td>4.2</td>
<td>1.8</td>
<td>3.7</td>
</tr>
<tr>
<td>Norway</td>
<td>9.4</td>
<td>4.7</td>
<td>2.7</td>
<td>23.3</td>
<td>9.7</td>
<td>23.0</td>
<td>4.6</td>
<td>1.6</td>
<td>2.7</td>
</tr>
<tr>
<td>Poland</td>
<td>6.9</td>
<td>10.2</td>
<td>2.7</td>
<td>19.3</td>
<td>3.0</td>
<td>36.1</td>
<td>5.9</td>
<td>2.0</td>
<td>6.6</td>
</tr>
<tr>
<td>Portugal</td>
<td>6.2</td>
<td>3.7</td>
<td>1.1</td>
<td>4.9</td>
<td>14.2</td>
<td>27.0</td>
<td>5.9</td>
<td>4.7</td>
<td>7.9</td>
</tr>
<tr>
<td>Romania</td>
<td>6.2</td>
<td>6.3</td>
<td>1.5</td>
<td>11.1</td>
<td>1.0</td>
<td>49.6</td>
<td>3.1</td>
<td>5.3</td>
<td>2.8</td>
</tr>
<tr>
<td>Spain</td>
<td>7.1</td>
<td>3.4</td>
<td>3.1</td>
<td>8.6</td>
<td>7.0</td>
<td>29.2</td>
<td>5.7</td>
<td>5.9</td>
<td>7.2</td>
</tr>
<tr>
<td>Sweden</td>
<td>9.9</td>
<td>7.8</td>
<td>3.5</td>
<td>24.7</td>
<td>7.5</td>
<td>19.5</td>
<td>3.7</td>
<td>1.4</td>
<td>2.0</td>
</tr>
<tr>
<td>Switzerland</td>
<td>13.1</td>
<td>10.1</td>
<td>3.1</td>
<td>23.8</td>
<td>2.3</td>
<td>25.6</td>
<td>2.8</td>
<td>2.4</td>
<td>4.9</td>
</tr>
<tr>
<td>UK</td>
<td>17.4</td>
<td>5.7</td>
<td>4.7</td>
<td>20.6</td>
<td>4.3</td>
<td>24.3</td>
<td>4.7</td>
<td>3.4</td>
<td>3.3</td>
</tr>
<tr>
<td>USSR</td>
<td>9.3</td>
<td>4.6</td>
<td>2.1</td>
<td>16.6</td>
<td>3.0</td>
<td>43.6</td>
<td>6.4</td>
<td>3.4</td>
<td>2.9</td>
</tr>
<tr>
<td>W Germany</td>
<td>11.4</td>
<td>12.5</td>
<td>4.1</td>
<td>18.8</td>
<td>3.4</td>
<td>18.6</td>
<td>5.2</td>
<td>1.5</td>
<td>3.8</td>
</tr>
<tr>
<td>Yugoslavia</td>
<td>4.4</td>
<td>5.0</td>
<td>1.2</td>
<td>9.5</td>
<td>0.6</td>
<td>55.9</td>
<td>3.0</td>
<td>5.7</td>
<td>3.2</td>
</tr>
</tbody>
</table>
SOM Example

“Protein” data base

Trained SOM Map
SOM Example

“Protein” data base

Trained SOM Map
SOM Example

“Protein” data base

1. RdMeat: Red meat
2. WhMeat: White meat
3. Eggs: Eggs
4. Milk: Milk
5. Fish: Fish
6. Cereal: Cereals
7. Starch: Starchy foods
8. Nuts: Pulses, nuts, and oil-seeds
9. Fr&Veg: Fruits and vegetables

Trained SOM Map    Plane 4
SOM Example

“Protein” data base

1. RdMeat: Red meat
2. WhMeat: White meat
3. Eggs: Eggs
4. Milk: Milk
5. Fish: Fish
6. Cereal: Cereals
7. Starch: Starchy foods
8. Nuts: Pulses, nuts, and oil-seeds
9. Fr&Veg: Fruits and vegetables

Trained SOM Map Plane 6
SOM Example

“Protein” data base

1. RdMeat: Red meat
2. WhMeat: White meat
3. Eggs: Eggs
4. Milk: Milk
5. Fish: Fish
6. Cereal: Cereals
7. Starch: Starchy foods
8. Nuts: Pulses, nuts, and oil-seeds
9. Fr&Veg: Fruits and vegetables

Trained SOM Map Plane 7 + 9
SOM Example

“Brainsize” data base