Full Bayesian Model Averaging of Naïve Bayes for Clustering

Guzmán Santafé
Jose Antonio Lozano
Pedro Larrañaga
Full Bayesian Model Averaging (FBMA):

\[ p(c, \mathbf{x} | D) = \sum_S p(S | D) \int p(c, \mathbf{x} | S, \theta) p(\theta | S, D) d\theta \]

Complete Data:
Dash and Cooper (2002) obtain a closed form to calculate a FBMA with supervised naïve Bayes classifiers

Incomplete Data:
No exact computation of FBMA is possible
We propose an approximation of FBMA with naïve Bayes for clustering
- **Naïve Bayes**: every predictive variable is conditional independent given $C$
- **Selective naïve Bayes**: every predictive variable can be independent or dependent on $C$

All selective naïve Bayes structures with two predictive variables
FBMA of Naïve Bayes for Clustering

- FBMA ≈ averaging over MAP parameter configurations for all selective naïve Bayes structures

- Approximation to FBMA computed in the same time complexity required to learn MAP parameters:
  
  Expectation and Model Averaging (EMA) algorithm allows to deal with missing data

  Dash and Cooper (2002) formula is extended to clustering

- A unique naïve Bayes for clustering is obtained
Assumptions

1. Multinomial variables: \( X_i = \{x_i^1, \ldots, x_i^r\} \) with \( i = 1, \ldots, n \) and \( C = \{c^1, \ldots, c^r\} \)

2. Complete dataset except for the latent cluster variable \((C)\)

3. Dirichlet priors over parameters
   \[ \theta_{ijk} \sim D(\alpha_{ijk}) \]

4. Parameter independence when the dataset is complete
   \[ p(\theta|S) = \theta_C \prod_{i=1}^{n} \prod_{j=1}^{q_i} p(\theta_{ij}|S) \]

5. Structure modularity
   \[ p(S) \propto p_s(C') \prod_{i=1}^{n} p_s(X_i, Pa_i) \]
EMA algorithm

- Adaptation of the EM algorithm (Dempster et al. 1979)
- Random initialization of the parameters
- Solution calculated iteratively in two steps:
  - Expectation
  - Model Averaging
- EMA is a greedy algorithm
EMA algorithm. E step

- The same E step as the one from EM algorithm
  - The dataset, $D$, is ‘completed’ obtaining $D_{CE}$

$$E(N_{ijk} | \theta_S, S) = \sum_{l=1}^{N} p(c^j | \mathbf{x}^{(l)}, x^k_i, \theta_S, S)$$

- $E(N_{ijk} | \theta_S, S)$ is considered as actual $N_{ijk}$
EMA algorithm. Example of E step

- Model

\[
\begin{align*}
\theta_{C-1} &= 0.4 \\
\theta_{100} &= 0.27 \\
\theta_{200} &= 0.70 \\
\theta_{300} &= 0.95 \\
\theta_{400} &= 0.30
\end{align*}
\]

- Data Set

<table>
<thead>
<tr>
<th></th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>?</td>
</tr>
</tbody>
</table>
EMA algorithm. Example of E step

- **Model**

  \[
  \begin{align*}
  \theta_{C-1} &= 0.4 \\
  \theta_{100} &= 0.27 & \theta_{110} &= 0.80 \\
  \theta_{200} &= 0.70 & \theta_{210} &= 0.10 \\
  \theta_{300} &= 0.95 & \theta_{310} &= 0.20 \\
  \theta_{400} &= 0.30 & \theta_{410} &= 0.70
  \end{align*}
  \]

- **Data Set**

<table>
<thead>
<tr>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
<th>(X_4)</th>
<th>(C)</th>
<th>(C_0)</th>
<th>(C_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.85</td>
<td>0.15</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.13</td>
<td>0.87</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.33</td>
<td>0.67</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.02</td>
<td>0.95</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.04</td>
<td>0.96</td>
</tr>
</tbody>
</table>

  \[
  \begin{align*}
  N_{C-0} &= 1.37 & N_{C-1} &= 3.63 \\
  N_{100} &= 0.17 & N_{101} &= 1.2 \\
  N_{110} &= 1.83 & N_{111} &= 1.8 \\
  N_{410} &= 3.48 & N_{411} &= 0.15
  \end{align*}
  \]
EMA algorithm. MA step

- MAP approximation for averaging over parameters (Heckerman, 1995)
  \[
  \int p(c^j, \mathbf{x}|S, \theta)p(\theta|S, D^{CE})d\theta \approx \tilde{\theta}^S_{Cj} \prod_{i=1}^n \tilde{\theta}^S_{ijk}
  \]

- FBMA approximation at each iteration:
  \[
  p(c^j, \mathbf{x}|D^{CE}) = \sum_S \int p(c^j, \mathbf{x}|S, \theta)p(\theta|S, D^{CE})d\theta p(S|D^{CE})
  \approx \kappa \sum_S \tilde{\theta}^S_{Cj} \prod_{i=1}^n \tilde{\theta}^S_{ijk} p(D^{CE}|S)p(S)
  \]

- Given assumptions (3) and (4) we can compute \( p(D^{CE}|S) \) (Cooper and Heskovits, 1992):
  \[
  p(D^{CE}|S) \approx \prod_{i=0}^n \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + E(N_{ijk}|\theta, S))} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + E(N_{ijk}|\theta, S))}{\Gamma(\alpha_{ijk})}
  \]

being, in this case, \( X_0 \) the cluster variable, \( C \)
FBMA of Naïve Bayes for Clustering

• Given assumption (5):

\[ p(c^j, x|D_{CE}) \approx \kappa \sum_S \rho_{C-j}^S \prod_{i=1}^n \rho_{ijk}^S \]

\[ \rho_{ijk}^S = \tilde{\theta}_{ijk}^S p_s(X_i, P_{ai}) \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + E(N_{ij}|\theta(t), S))} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + E(N_{ijk}|\theta(t), S))}{\Gamma(\alpha_{ijk})} \]

\[ \rho_{C-j}^S = \tilde{\theta}_{C-j}^S p_s(C) \frac{\Gamma(\alpha_C)}{\Gamma(\alpha_C + E(N_{C|\theta(t), S}))} \prod_{j=1}^{r_C} \frac{\Gamma(\alpha_{C-j} + E(N_{C-j}|\theta(t), S))}{\Gamma(\alpha_{C-j})} \]

• If we expand the summation of structures:

\[ p(c^j, x|D_{CE}) \approx \kappa \left( \rho_{C-j} \rho_{1-k} \rho_{2-k} \cdots \rho_{n-k} + \rho_{C-j} \rho_{1jk} \rho_{2-k} \cdots \rho_{n-k} + \cdots + \rho_{C-j} \rho_{1jk} \rho_{2jk} \cdots \rho_{njk} \right) \]

2^n terms
FBMA of Naïve Bayes for Clustering

- $\sum_{m}^{j}k$ denotes the sum of the product up to $m$-th variable:

$$
\sum_{m}^{j}k \equiv \rho_{c-j} \rho_{1-k} \rho_{2-k} \cdots \rho_{m-k} 
+ \rho_{c-j} \rho_{1^j k} \rho_{2-k} \cdots \rho_{m-k} 
+ \vdots 
+ \rho_{c-j} \rho_{1^j k} \rho_{2^j k} \cdots \rho_{m^j k}
$$

- Thus, $\sum_{i}^{j}k$ can be written as a recurrence relationship:

$$
\sum_{i}^{j}k = \sum_{i-1}^{j}k (\rho_{i-k} + \rho_{i^j k}), \quad \sum_{0}^{j}k = \rho_{c-j}
$$

- Then, $p(c^{j}, x|D^{CE})$ can be approximated by a closed form

$$
p(c^{j}, x|D^{CE}) \approx \kappa \rho_{c^{j}} \prod_{i=1}^{n} (\rho_{i-k} + \rho_{i^j k})
$$

- Parameters for the model in current iteration of the EMA:

$$
\theta^{*}_{i^j k} \propto (\rho_{i-k} + \rho_{i^j k})
$$
FBMA of Naïve Bayes for Clustering

Empirical Testing

- Empirical evidence of the approximation is close to FBMA
- Model learned with EMA algorithm, $P_{EMA}$
- Brute force model, $P_{BF}$:

$$p(c, x|D) = \sum_S \int p(c, x|S, \theta)p(\theta|S, D)d\theta \ p(D|S)P(S)$$

- Averaging over parameters $\approx$ MAP (EM)
- $P(D|S) \approx$ Candidate method (Chickering et al., 1997)
- Averaging over structures: Brute force

- Distance between models: $D_{KL}(P_{BF}, P_{EMA})$
- Distances between $P_{BF}$ and 10000 random naïve Bayes models
FBMA of Naïve Bayes for Clustering

Empirical Testing

Test for a model with 6 predictive variables
Empirical Testing

<table>
<thead>
<tr>
<th>Test</th>
<th>$D_{KL}(P_{BF}, P_{EMA})$</th>
<th>$D_{KL}(P_{BF}, P_{5%})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.15495</td>
<td>0.97264</td>
</tr>
<tr>
<td>2</td>
<td>0.55039</td>
<td>0.98286</td>
</tr>
<tr>
<td>3</td>
<td>0.41811</td>
<td>0.67852</td>
</tr>
<tr>
<td>4</td>
<td>0.56426</td>
<td>0.69513</td>
</tr>
<tr>
<td>5</td>
<td>0.33207</td>
<td>0.90145</td>
</tr>
<tr>
<td>6</td>
<td>0.42465</td>
<td>0.88862</td>
</tr>
<tr>
<td>7</td>
<td>0.40876</td>
<td>1.07019</td>
</tr>
<tr>
<td>8</td>
<td>0.45795</td>
<td>1.06288</td>
</tr>
<tr>
<td>9</td>
<td>0.24308</td>
<td>0.76765</td>
</tr>
<tr>
<td>10</td>
<td>0.13184</td>
<td>1.74327</td>
</tr>
</tbody>
</table>

10 independent tests for models with 6 predictive variables
Model Detection

- Test if the EMA model detects the real model structure
- Set a selective naïve Bayes model
- Sample a data set
- Learn an EMA model from the data set

\[
\begin{array}{cccc}
1 & 0 & 1 & 1 \\
2 & 0 & 1 & 0 \\
N & 1 & 0 & 0 & 1 \\
\end{array}
\]
• Measure of independence for $X_i$ in the EMA model

$$I_P(X_i) = \frac{\sum_{j=1}^{rC} D_{KL}(p(X_i), p(X_i|c^j))}{r_C}$$

• The bigger is $N \rightarrow \begin{cases} & \text{the better is the MAP approx.} \\ & \text{the better is the EMA approx.} \end{cases}$

• Example of model detection
Test for model detection with 6 predictive variables ($X_4$ and $X_6$ independent of $C$)
Conclusions

• Empirical test ⇒ $P_{EMA}$ good approximation to FBMA

• Approximate a FBMA with EMA is not much expensive than a classical MAP approach with EM

• EMA is able to detect independencies between variables:
  EMA can be used for FSS

• EMA can be extended in order to deal with incomplete data.

• EMA can be extended to more complicated model (TAN)