COSC 522 – Machine Learning

Lecture 15 – Fusion

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
http://www.eecs.utk.edu/faculty/qi
Email: hqi@utk.edu

Course Website: http://web.eecs.utk.edu/~hqi/cosc522/
Recap

- **Supervised learning**
  - Classification
    - Bayesian based - Maximum Posterior Probability (MPP): For a given $x$, if $P(w_1|x) > P(w_2|x)$, then $x$ belongs to class 1, otherwise 2.
    - Parametric Learning
      - Case 1: Minimum Euclidean Distance (Linear Machine), $\Sigma_i = \sigma^2 I$
      - Case 2: Minimum Mahalanobis Distance (Linear Machine), $\Sigma_i = \Sigma$
      - Case 3: Quadratic classifier, $\Sigma_i = \text{arbitrary}$
    - Nonparametric Learning
      - K-Nearest Neighbor
  - Regression
    - Linear regression (Least square-based vs. ML)
  - Neural network
    - Perceptron
    - BPNN
  - Kernel-based approaches
    - Support Vector Machine
  - Decision tree

- **Unsupervised learning** (Kmeans, Winner-takes-all)
- **Supporting preprocessing techniques** [Standardization, Dimensionality Reduction (FLD, PCA)]
- **Supporting postprocessing techniques** [Performance Evaluation (confusion matrices, ROC)]
  - Classifier Fusion

$$P(w_j | x) = \frac{P(x | w_j)P(w_j)}{p(x)}$$
Questions

• Rationale with fusion?
• Different flavors of fusion?
• The fusion hierarchy
• What is the cost function for Naïve Bayes?
• What is the procedure for Naïve Bayes?
• What is the limitation of Naïve Bayes?
• What is the procedure of Behavior-Knowledge-Space (BKS)?
• How does it resolve issues with NB?
• What is Boosting and what is its difference to committee-based fusion approaches?
• What is AdaBoost?
Motivation

• Combining classifiers to achieve higher accuracy
  – Combination of multiple classifiers
  – Classifier fusion
  – Mixture of experts
  – Committees of neural networks
  – Consensus aggregation
  – ...

• Reference:

Three heads are better than one.
Popular Approaches

- Data-based fusion (early fusion)
- Feature-based fusion (middle fusion)
- Decision-based fusion (late fusion)

Approaches

- Committee-based
  - Majority voting
  - Bootstrap aggregation (Bagging) [Breiman, 1996]
- Bayesian-based
  - Naïve Bayes combination (NB)
  - Behavior-knowledge space (BKS) [Huang and Suen, 1995]
- Boosting
  - Adaptive boosting (AdaBoost) [Freund and Schapire, 1996]
- Interval-based integration
Application Example – Civilian Target Recognition

Ford 250  Harley Motocycle  Ford 350  Suzuki Vitara
Consensus Patterns

- Unanimity (100%)
- Simple majority (50%+1)
- Plurality (most votes)
Example of Majority Voting - Temporal Fusion

- Fuse all the 1-sec sub-interval local processing results corresponding to the same event (usually lasts about 10-sec)
- Majority voting

\[
\rho_i^j = \arg \max \omega_c, \quad c \in [1, C]
\]

- number of local output \(c\) occurrence
- number of possible local processing results
Questions

- Rationale with fusion?
- Different flavors of fusion?
- The fusion hierarchy
- What is the cost function for Naïve Bayes?
- What is the procedure for Naïve Bayes?
- What is the limitation of Naïve Bayes?
- What is the procedure of Behavior-Knowledge-Space (BKS)?
- How does it resolve issues with NB?
- What is Boosting and what is its difference to committee-based fusion approaches?
- What is AdaBoost?
NB (the independence assumption)

Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>AAV</th>
<th>DW</th>
<th>HMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>C1</td>
<td>894</td>
<td>329</td>
<td>143</td>
</tr>
<tr>
<td>AAV</td>
<td>99</td>
<td>411</td>
<td>274</td>
</tr>
<tr>
<td>DW</td>
<td>98</td>
<td>42</td>
<td>713</td>
</tr>
<tr>
<td>HMV</td>
<td>98</td>
<td>42</td>
<td>713</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>AAV</th>
<th>DW</th>
<th>HMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>C2</td>
<td>1304</td>
<td>156</td>
<td>77</td>
</tr>
<tr>
<td>AAV</td>
<td>114</td>
<td>437</td>
<td>83</td>
</tr>
<tr>
<td>DW</td>
<td>13</td>
<td>107</td>
<td>450</td>
</tr>
<tr>
<td>HMV</td>
<td>13</td>
<td>107</td>
<td>450</td>
</tr>
</tbody>
</table>

The real class is DW, the classifier says it’s HMV

k

i = 1, 2 (classifiers)

Probability that the true class is k given that $C_i$ assigns it to $s$

Probability multiplication
NB – Derivation

- Assume the classifiers are **mutually independent**
- Bayes combination - Naïve Bayes, simple Bayes, idiot’s Bayes
- Assume
  - $L$ classifiers, $i=1,...,L$
  - $c$ classes, $k=1,...,c$
  - $s_i$: class label given by the $i$th classifier, $i=1,...,L$, $s=s_1,...,s_L$

\[
P(\omega_k|s) = \frac{p(s|\omega_k)P(\omega_k)}{p(s)} = \frac{P(\omega_k) \prod_{i=1}^{L} p(s_i|\omega_k)}{p(s)}
\]

\[
P(\omega_k) = \frac{N_k}{N}
\]

\[
p(s_i|\omega_k) = \frac{c m_{k,s_i}}{N_k}
\]

\[
P(\omega_k|s) \approx \frac{1}{N_k^{L-1}} \prod_{i=1}^{L} c m_{k,s_i}
\]
BKS

- Majority voting won’t work
- Behavior-Knowledge Space algorithm (Huang&Suen)

Assumption:
- 2 classifiers
- 3 classes
- 100 samples in the training set

Then:
- 9 possible classification combinations

<table>
<thead>
<tr>
<th>$c_1$, $c_2$</th>
<th>samples from each class</th>
<th>fused result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,1</td>
<td>10/3/3</td>
<td>1</td>
</tr>
<tr>
<td>1,2</td>
<td>3/0/6</td>
<td>3</td>
</tr>
<tr>
<td>1,3</td>
<td>5/4/5</td>
<td>1,3</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td></td>
</tr>
<tr>
<td>3,3</td>
<td>0/0/6</td>
<td>3</td>
</tr>
</tbody>
</table>
Questions

- Rationale with fusion?
- Different flavors of fusion?
- The fusion hierarchy
- What is the cost function for Naïve Bayes?
- What is the procedure for Naïve Bayes?
- What is the limitation of Naïve Bayes?
- What is the procedure of Behavior-Knowledge-Space (BKS)?
- How does it resolve issues with NB?
- What is Boosting and what is its difference to committee-based fusion approaches?
- What is AdaBoost?
Boosting

- Base classifiers are trained in sequence!
- Base classifiers as weak learners
- Weighted majority voting to combine classifiers

\[
Y_M(x) = \text{sign} \left( \sum_{m} \alpha_m y_m(x) \right)
\]

**AdaBoost**

1. Initialize the data weighting coefficients \( \{w_n^{(1)}\} \) by setting \( w_n^{(1)} = \frac{1}{N} \) for \( n = 1, \ldots, N \).
2. For \( m = 1, \ldots, M \):
   a. Fit a classifier \( y_m(x) \) to the training data by minimizing the weighted error function
      \[
      J_m = \frac{1}{N} \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n)
      \]
   b. Evaluate the quantities \( \epsilon_m = \frac{1}{N} \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n) \) and then use these to evaluate \( \alpha_m = \ln \left\{ \frac{1 - \epsilon_m}{\epsilon_m} \right\} \).
   c. Update the data weighting coefficients
      \[
      w_n^{(m+1)} = w_n^{(m)} \exp \left\{ \alpha_m I(y_m(x_n) \neq t_n) \right\}
      \]
AdaBoost

- Step 1: Initialize the data weighting coefficients \( \{w_n\} \) by setting \( w_n^{(1)} = 1/N \), where \( N \) is the \# of samples
- Step 2: for each classifier \( y_m(x) \)
  - (a) Fit a classifier \( y_m(x) \) to the training data by minimizing the weighted error function
    \[
    J_m = \sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n)
    \]
  - (b) Evaluate the quantities
    \[
    \epsilon_m = \frac{\sum_{n=1}^{N} w_n^{(m)} I(y_m(x_n) \neq t_n)}{N}
    \]
    \[
    \alpha_m = \ln \left( \frac{1 - \epsilon_m}{\epsilon_m} \right)
    \]
  - (c) Update the data weighting coefficients
    \[
    w_n^{(m+1)} = w_n^{(m)} \exp \{\alpha_m I(y_m(x_n) \neq t_n)\}
    \]
- Step 3: Make predictions using the final model
  \[
  Y_M(x) = \text{sign} \left( \sum_{m=1}^{M} \alpha_m y_m(x) \right)
  \]
Figure 14.2 Illustration of boosting in which the base learners consist of simple thresholds applied to one or other of the axes. Each figure shows the number $m$ of base learners trained so far, along with the decision boundary of the most recent base learner (dashed black line) and the combined decision boundary of the ensemble (solid green line). Each data point is depicted by a circle whose radius indicates the weight assigned to that data point when training the most recently added base learner. Thus, for instance, we see that points that are misclassified by the $m = 1$ base learner are given greater weight when training the $m = 2$ base learner. Instead of doing a global error function minimization, however, we shall suppose that the base classifiers $y_1(x),...,y_{m-1}(x)$ are fixed, as are their coefficients $\alpha_1,...,\alpha_{m-1}$, and so we are minimizing only with respect to $\alpha_m$ and $y_m(x)$. Separating off the contribution from base classifier $y_m(x)$, we can then write the error function in the form

$$E = \sum_{n=1}^{N} \exp\left\{ -t_n f_{m-1}(x_n) - \alpha_m y_m(x_n) \right\}$$

(14.22)
Value-based vs. Interval-based Fusion

- Interval-based fusion can provide fault tolerance
- Interval integration – overlap function
  - Assume each sensor in a cluster measures the same parameters, the integration algorithm is to construct a simple function (overlap function) from the outputs of the sensors in a cluster and can resolve it at different resolutions as required

Crest: the highest, widest peak of the overlap function
## A Variant of kNN

- Generation of local confidence ranges (For example, at each node $i$, use kNN for each $k \in \{5, \ldots, 15\}$)

<table>
<thead>
<tr>
<th>$k$</th>
<th>Class 1</th>
<th>Class 2</th>
<th>...</th>
<th>Class n</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>3/5</td>
<td>2/5</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>2/6</td>
<td>3/6</td>
<td>...</td>
<td>1/6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>15</td>
<td>10/15</td>
<td>4/15</td>
<td>...</td>
<td>1/15</td>
</tr>
</tbody>
</table>

- Apply the integration algorithm on the confidence ranges generated from each node to construct an overlapping function
Example of Interval-based Fusion

<table>
<thead>
<tr>
<th></th>
<th>stop 1</th>
<th>stop 2</th>
<th>stop 3</th>
<th>stop 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>c</td>
<td>acc</td>
<td>c</td>
<td>acc</td>
</tr>
<tr>
<td>class 1</td>
<td>1</td>
<td>0.2</td>
<td>0.5</td>
<td>0.125</td>
</tr>
<tr>
<td>class 2</td>
<td>2.3</td>
<td>0.575</td>
<td>4.55</td>
<td>0.35</td>
</tr>
<tr>
<td>class 3</td>
<td>0.7</td>
<td>0.175</td>
<td>0.5</td>
<td>0.25</td>
</tr>
</tbody>
</table>

The table above represents the confidence (c) and accuracy (acc) values for different stops and classes. Each row corresponds to a class, with columns indicating the confidence and accuracy values for stops 1, 2, 3, and 4.
Confusion Matrices of Classification on Military Targets

<table>
<thead>
<tr>
<th></th>
<th>AAV</th>
<th>DW</th>
<th>HMV</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAV</td>
<td>29</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>DW</td>
<td>0</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>HMV</td>
<td>0</td>
<td>2</td>
<td>23</td>
</tr>
</tbody>
</table>

Acoustic (75.47%, 81.78%)

Seismic (85.37%, 89.44%)

Multi-modality fusion (84.34%)

Multi-sensor fusion (96.44%)
Confusion Matrices

Acoustic

Seismic

Multi-modal
Reference

• For details regarding majority voting and Naïve Bayes, see