Multi-sensor classification with Consensus-based Multi-view Maximum Entropy Discrimination

Tianpei Xie*, Nasser M. Nasrabadi†, Alfred O. Hero*

*University of Michigan, Ann Arbor, †U.S. Army Research Lab
1. Problem Motivations

2. Consensus-constraint via information geometry

3. Consensus-based Multi-view Maximum Entropy Learning

4. Experiments

5. Conclusion
Outline

1. Problem Motivations
2. Consensus-constraint via information geometry
3. Consensus-based Multi-view Maximum Entropy Learning
4. Experiments
5. Conclusion
Motivations

• In many applications, samples can be represented in multiple ways (referred as *multi-view* samples).

• For instance,
  1. In web-page network, ...
  2. In muti-sensor network, ...
  3. In biometrics, ...
  4. etc.
Multi-view learning and Challenges

- We focus on *multi-view learning*: learning to predict or classify based on multi-view data.

- A few challenges arise in multi-view learning:
  1. Information Fusion? Robustness?
  2. Parsimony?
  3. Unlabeled samples?

- In this work, we consider the **semi-supervised multi-view learning** problem.
Previous works

1. Multi-view feature learning methods, e.g.
   - Canonical Correlation Analysis (CCA) [Rupnik and Shawe-Taylor, 2010], Bi-modal Deep Autoencoder (Bi-DAE) [Ngiam et al., 2011]; SVM-2K [Farquhar et al., 2005], etc.
   - Cons: sensitive to local outliers;

2. Decision-level fusion e.g.
   - Bayes-Fusion: e.g. MCMC, particle filter methods [Klein, 2004]
   - Model averaging: e.g. boosting methods [Collins and Singer, 1999], etc
   - Cons: between-view correlation not taken into account;

3. Consensus-based multi-view learning model, e.g.
   - Co-training [Blum and Mitchell, 1998], Bayesian Co-training (Bayes Co-trn) [Yu et al., 2007], Multi-View MED [Sun and Chao, 2013] etc.
Our contribution

We propose a **Consensus-based Multi-View Maximum Entropy Discrimination (CMV-MED)** framework:

- Features are view-specific posterior distributions;
- a *consensus-view* model $\iff$ proposed *dissimilarity measure* btw these posterior distributions.
- $\Rightarrow$ *centroid* in an intrinsic *non-Euclidean* space induced via K-L div.

![Diagram](attachment:image.png)
## Comparison of multi-view learning methods

<table>
<thead>
<tr>
<th>Method</th>
<th>fusion stage</th>
<th>parsimony</th>
<th>semi-sup.</th>
<th>noise tol</th>
<th>Bayes.</th>
<th>#. views</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCA</td>
<td>feature</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Bi-DAE</td>
<td>feature</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>SVM-2K</td>
<td>feature</td>
<td>x</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Bayes-Fusion</td>
<td>decision</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>≥ 2</td>
</tr>
<tr>
<td>Boosting</td>
<td>decision</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>x</td>
<td>≥ 2</td>
</tr>
<tr>
<td>Co-training</td>
<td>consens.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>2</td>
</tr>
<tr>
<td>Bayes Co-trn</td>
<td>consens.</td>
<td>x</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>≥ 2</td>
</tr>
<tr>
<td>MV-MED</td>
<td>consens.</td>
<td>✓</td>
<td>✓</td>
<td>x</td>
<td>✓</td>
<td>2</td>
</tr>
<tr>
<td>CMV-MED</td>
<td>consens.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>≥ 2</td>
</tr>
</tbody>
</table>
**Assumptions and Stochastic consensus**

- **Binary** classification task with $V$ views, with $\mathbf{x}^{[V]} \equiv (\mathbf{x}^1, \ldots, \mathbf{x}^V) \in \mathcal{X}^1 \times \ldots \mathcal{X}^V$ and $y = \{-1, +1\}$.

- Log-linear predictive model $\log p_i(y|\mathbf{x}^i, \mathbf{w}_i) \propto \frac{1}{2} y (\mathbf{w}_i^T \mathbf{x}^i)$ for view $i$.

- The proposed **stochastic consensus** measure is given as

  $$ R_\pi(\mathbf{w}_1, \mathbf{w}_2) $$

  $$ = \mathbb{E}_{(\mathbf{x}^1, \mathbf{x}^2)} \left[ D \left( p_1(y|\mathbf{x}^1, \mathbf{w}_1), p_2(y|\mathbf{x}^2, \mathbf{w}_2) \right) \right] \quad (V = 2) $$

  $$ = \mathbb{E}_{(\mathbf{x}^1, \mathbf{x}^2)} \left[ \min_{q(y|\mathbf{x}^2) \in \Delta(Y)} \sum_{i \in \{1,2\}} \pi_i \text{KL} \left( q(y|\mathbf{x}^2) \| p_i(y|\mathbf{x}^i, \mathbf{w}_i) \right) \right] $$

  where $\text{KL}(\cdot\|\cdot)$ denotes the K-L divergence, and the weight $\pi \in \Delta$. $R_{\pi > 0} = 0$ iff $p_1 = p_2$. The optimal sol. $q^*(y|\mathbf{x}_m^{[2]}) \Rightarrow$ consensus-view model.
Comparison with other consensus measure

(1) Stochastic consensus:

\[ D(p, q) = \exp(-\text{sign}(p) p - \text{sign}(q) p); \]

(2) Exp-consensus:

\[ D(p, q) = \|p - q\|^2 \]
Interpretation in information geometry

- \( q^*(y|\mathbf{x}^{[V]}) = \arg \min_{q(y) \in \Delta(Y)} \sum_{i=1}^{V} \pi_i \text{KL}(q(y|\mathbf{x}^{[V]}) || p_i(y|x^i, w_i)). \)

\( \Rightarrow \) The centroid of conv. \( \{p_i(y|x^i, w_i), i = 1, \ldots, V\} \) in log(\( \Delta(Y) \)).
Outline

1. Problem Motivations
2. Consensus-constraint via information geometry
3. Consensus-based Multi-view Maximum Entropy Learning
4. Experiments
5. Conclusion
Maximum Entropy Discrimination (MED)

MED framework introduced by Jaakkola et al. [1999].

Let $\Delta F(y_n, x_n; w) = \log \left( \frac{p(y_n|x_n, w)}{p(y \neq y_n|x_n, w)} \right)$ is the discriminative functions.

MED: learn a convex combination of discriminative functions via Maximum Entropy principle.

- Assume the prior on $w$ and $\gamma$ as $p_0(w)p_0(\gamma)$, the goal is to learn $q(w, \gamma|D)$ via solving the following

$$\min_{q(w, \gamma|D) \in \Delta} \text{KL} (q(w, \gamma|D)||p_0(w)p_0(\gamma))$$

s.t. $\mathbb{E}_{q(w, \gamma|D)} [\Delta F(y_n, x_n; w) - \gamma_n] \geq 0, \forall n$

- MED defines decision rule via Bayesian averaging

$$y^* = \arg \max_{w, \gamma} \int_p(y|x^*, w)q(w, \gamma|D)dw d\gamma$$

$\Rightarrow$ MED is robust compared to single classifier [Jaakkola et al., 1999].
Our solution for CMV-MED is based on variational EM [Sindhwani et al., 2006]

1. Given the $\hat{w}_i^{t-1} = E_{q_t^{t-1}(w_i)}[w_i], i = 1, \ldots, V$ from single-view MED, find the consensus view on unlabeled data $U$ via information projection, i.e.

$$\log q_t(y|x_n^{[V]}) = \frac{1}{V} \sum_{i=1}^{V} \log p_i(y|x_n, \hat{w}_i^{t-1}) - \log Z(x_n), \forall n \in U,$$

where $Z(x_n)$ is the normalization factor.

2. Given the consensus view $q_t(y|x_n), \forall n \in U$, solve for each view $i = 1, \ldots, V$ a MED problem independently to obtain the following optimal solution

$$q_t(w_i|\alpha^i, \beta^i) = \text{MED-Solver}((y_n, x_n^i)_{n \in L}, \{x_m^i\}_{m \in U}, \{\hat{y}_m \sim q_t(y|x_m)\}_{m \in U}),$$

where $(\alpha^i, \beta^i)$ are dual variables associated with the SVM-type solution.

3. Repeat 1 and 2 until converge.
We test on **ARL-Footstep** [Damarla et al., 2011] data.

- It is a multi-sensor data set that contains acoustic signals collected by four well-synchronized sensors (labeled as Sensor 1,2,3,4) in a natural environment.

- The task is to discriminate between human footsteps and human-leading animal footsteps.

- It involves 840 segments from human subjects and 660 segments from human-animal subjects. We choose 600 segments from each class as the training set with $|L| = 50$.

- In each view, the feature dimension $d = 200$.

- measure the classification accuracy vs. size of labeled samples.

- We compare the proposed CMV-MED model with the SVM-2K, MV-MED as well as the single-view MED for each view.
The **WebKB4** [Craven et al., 2000] data set is widely-used in multi-view learning literature.

- It consists of 1051 two-view web pages collected from computer science department web sites at four universities.
- The task is to discriminate between course page and non-course page.
- There are 230 course pages and 821 non-course pages. The two natural views are words in a web page and words appearing in the links pointing to that page.
- In each view, we compute the term frequency-inverse document frequency weights (TF-IDF) features from the document word matrix.
- measure the classification accuracy vs. size of labeled samples
Conclusion

- The proposed method maximizes the stochastic agreement btw different models on unlabeled samples.

- The learned consensus-view distribution is the centroid of all view-specific posterior distributions over the space of probability measures.

- The proposed multi-view learning algorithm has higher accuracy and lower variance compared to its single-view counterparts.
Acknowledgment

- This research was partially supported by US Army Research Office (ARO) grants W911NF-11-1-0391 and WA11NF-11-1-103A1.

- Thanks for Army Research Lab to provide data sets.


