# Probabilistic Combination of Classifier and Cluster Ensembles for Non-transductive Learning


**Ayan Acharya**, Eduardo R. Hruschka, Joydeep Ghosh, Badrul Sarwar, Jean-David Ruvini

Dept. of ECE, UT Austin

May 3, 2013

# Motivations and Applications

Motivations?

- Unsupervised models provide a variety of supplementary constraints for classifying new data.
- Similar new instances in the target set are more likely to share the same class label.

Applications?

- Improve performance given "weak" classifiers/few labeled data.
- Semi-supervised and transfer learning.

# Pedagogical Example

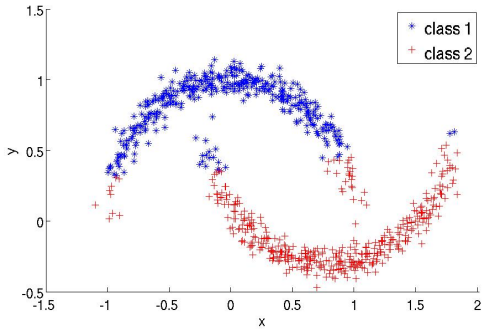



Figure: Class Labels from the Classifier Ensemble.

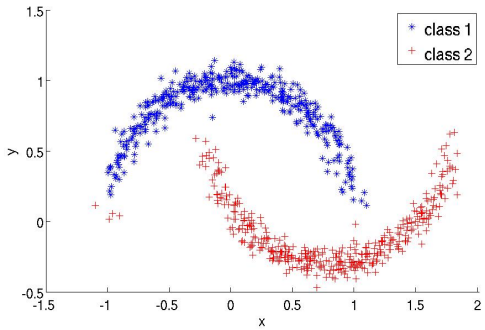
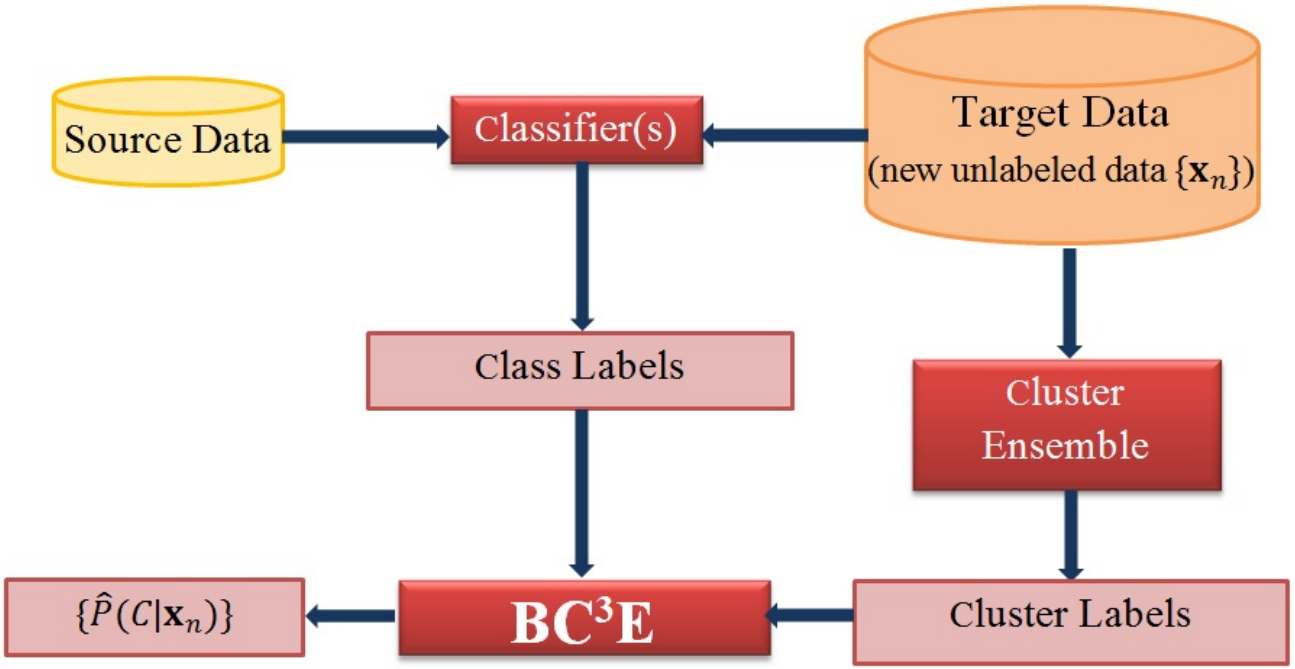



Figure: Class Labels from $\mathbf{BC^3E}$.

# Overview of BC$^3$E

**B**ayesian **C**ombination of **C**lassifier and **C**lustering **E**nsemble



# Overview of $\mathbf{BC^3E}$ – continued

| | $\mathbf{w}_1^{(1)}$ | $\mathbf{w}_2^{(1)}$ | $\cdots$ | $\mathbf{w}_{r_1}^{(1)}$ |
|---|---|---|---|---|
| $\mathbf{x}_1$ | 2 | 3 | $\cdots$ | 1 |
| $\mathbf{x}_2$ | 1 | 3 | $\cdots$ | 1 |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| $\mathbf{x}_N$ | 2 | 3 | $\cdots$ | 3 |

Table: From Classifiers

| | $\mathbf{w}_1^{(2)}$ | $\mathbf{w}_2^{(2)}$ | $\cdots$ | $\mathbf{w}_{r_2}^{(2)}$ |
|---|---|---|---|---|
| $\mathbf{x}_1$ | 4 | 5 | $\cdots$ | 4 |
| $\mathbf{x}_2$ | 2 | 4 | $\cdots$ | 4 |
| $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ | $\cdots$ |
| $\mathbf{x}_N$ | 2 | 4 | $\cdots$ | 2 |

Table: From Clustering Algorithms

# Limitations of Earlier Methods

- $C^3E$ (Acharya *et. al.*, 2012) has a tuning parameter $\alpha$ which controls how much weight the unsupervised information should be assigned – not always suitable for transfer learning applications.
- LWE (Gao *et. al.*, 2008) is relatively insensitive to such parameter tuning but performance is inferior.
- Neither of the above two methods offer privacy w.r.t. the class/cluster labels.

# Background – Bayesian Clustering Ensemble (**BCE**)

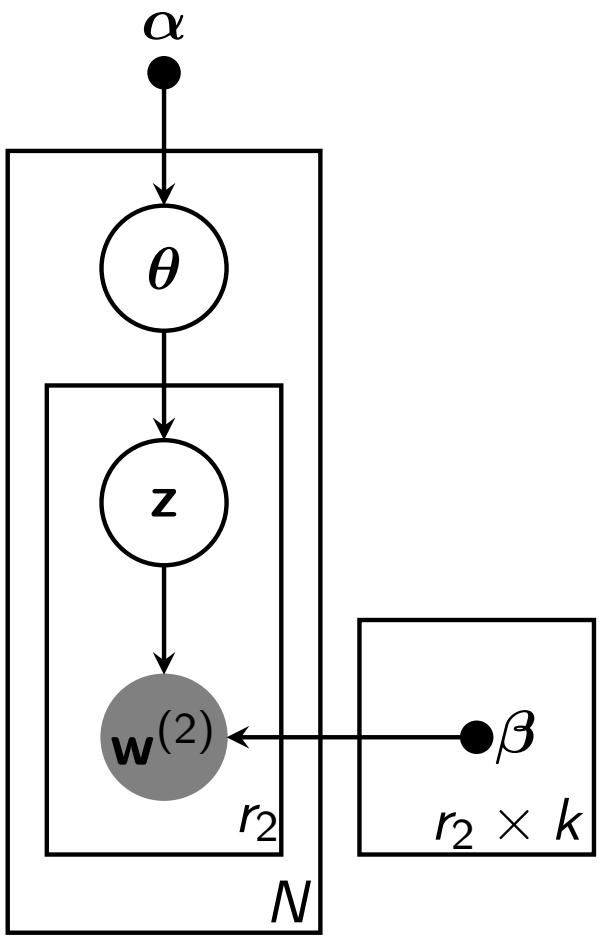
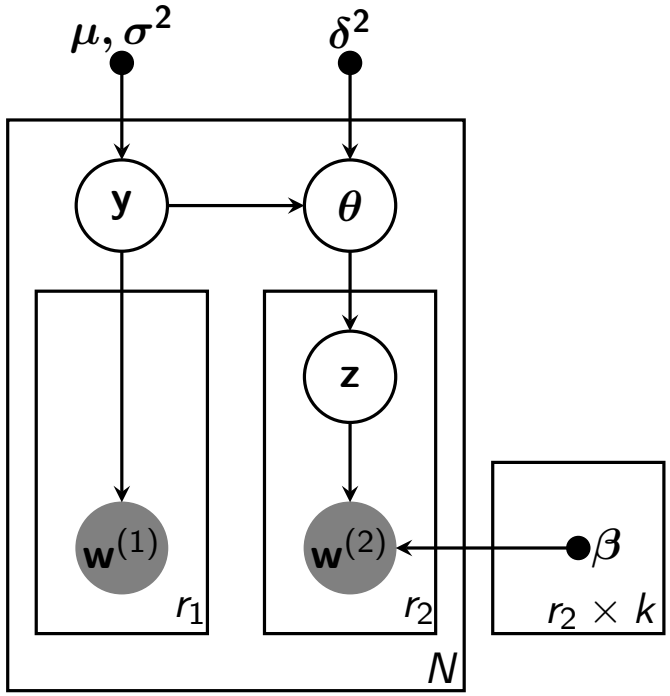



Figure: Graphical Model for **BCE** (Wang *et. al.*, 2011)

1. Choose $\theta_n \sim \text{Dir}(\alpha)$.
2. $\forall m \in \{1, 2, \cdots, r_2\}$.
   1. Choose $\mathbf{z}_{nm} \sim \text{multinomial}(\theta_n)$ where $\mathbf{z}_{nm}$ is a vector of dimension $k$ with only one component being unity and others being zero.
   2. Choose $w_{nm}^{(2)} \sim \text{multinomial}(\beta_{r\mathbf{z}_{nm}})$.

# Graphical Model of $\mathbf{BC^3E}$




$f(\mathbf{y}_n) = \left( \frac{\exp(y_{ni})}{\sum_i \exp(y_{ni})} \right)_{i=1}^{k}$ is the softmax function.

# Generative Process of $\mathbf{BC^3E}$

- For each $\mathbf{x}_n \in \mathcal{X}$, choose $\mathbf{y}_n \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^k$ is the mean and $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$ is the covariance.
- Choose $\boldsymbol{\theta}_n \sim \mathcal{N}(\mathbf{y}_n, \delta^2 I_k)$, where $\delta^2 \geq 0$ is the scaling factor of the covariance of the normal distribution centered at $\mathbf{y}_n$, and $I_k$ is the identity $k \times k$ matrix.
- $\forall l \in \{1, 2, \cdots, r_1\}$, choose $\mathbf{w}_{nl}^{(1)} \sim f(\mathbf{y}_n)$.
- $\forall m \in \{1, 2, \cdots, r_2\}$:
  1. Choose $\mathbf{z}_{nm} \sim f(\boldsymbol{\theta}_n)$, where $\mathbf{z}_{nm}$ is a $k$-dimensional vector with 1-of-$k$ representation.
  2. Choose $\mathbf{w}_{nm}^{(2)} \sim \text{multinomial}(\boldsymbol{\beta}_{r\mathbf{z}_{nm}})$.

# Distributions of $\mathbf{BC^3E}$

- Joint Distribution of $\mathbf{BC^3E}$:

$$p(\mathbf{X}, \mathbf{Z}|\zeta_0) = \prod_{n=1}^{N} p(\mathbf{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) p(\boldsymbol{\theta}_n|\mathbf{y}_n, \delta^2 I_k) \prod_{l=1}^{r_1} p(w_{nl}^{(1)}|f(\mathbf{y}_n)) \prod_{m=1}^{r_2} p(\mathbf{z}_{nm}|f(\boldsymbol{\theta}_n)) p(w_{nm}^{(2)}|\boldsymbol{\beta}, \mathbf{z}_{nm}). \quad (1)$$

- Variational Distribution of $\mathbf{BC^3E}$:

$$q(\mathbf{Z}|\{\zeta_n\}_{n=1}^{N}) = \prod_{n=1}^{N} q(\mathbf{y}_n|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) q(\boldsymbol{\theta}_n|\epsilon_n, \boldsymbol{\Delta}_n) \prod_{m=1}^{r_2} q(\mathbf{z}_{nm}|\phi_{nm}). \quad (2)$$

# Updates in BC$^3$E

$$\mathcal{L}_{[\mu_n]} = -\frac{1}{2}\sum_{i=1}^{k}\frac{(\mu_{ni}-\mu_i)^2}{\sigma_i^2} - \frac{1}{2\delta^2}\sum_{i=1}^{k}(\mu_{ni}-\epsilon_{ni})^2 \tag{3}$$
$$+ \sum_{l=1}^{r_1}\sum_{i=1}^{k} w_{nli}^{(1)}\mu_{ni} - \frac{r_1}{\xi_n}\sum_{i=1}^{k}\exp(\mu_{ni}+\sigma_{ni}^2/2).$$

$$\mathcal{L}_{[\epsilon_n]} = \sum_{m=1}^{r_2}\sum_{i=1}^{k}\phi_{nmi}\epsilon_{ni} - \frac{1}{\xi_n}\sum_{i=1}^{k}\exp(\epsilon_{ni}+\delta_{ni}^2/2) \tag{4}$$
$$- \frac{1}{2}\sum_{i=1}^{k}\frac{(\epsilon_{ni}-\mu_{ni})^2}{\delta^2}.$$

$$\phi_{nmi}^{*} \propto \exp\left(\epsilon_{ni} + \sum_{j=1}^{k^{(m)}}\beta_{mij}w_{nmj}^{(2)}\right). \tag{5}$$

# Distributed Learning Scenarios

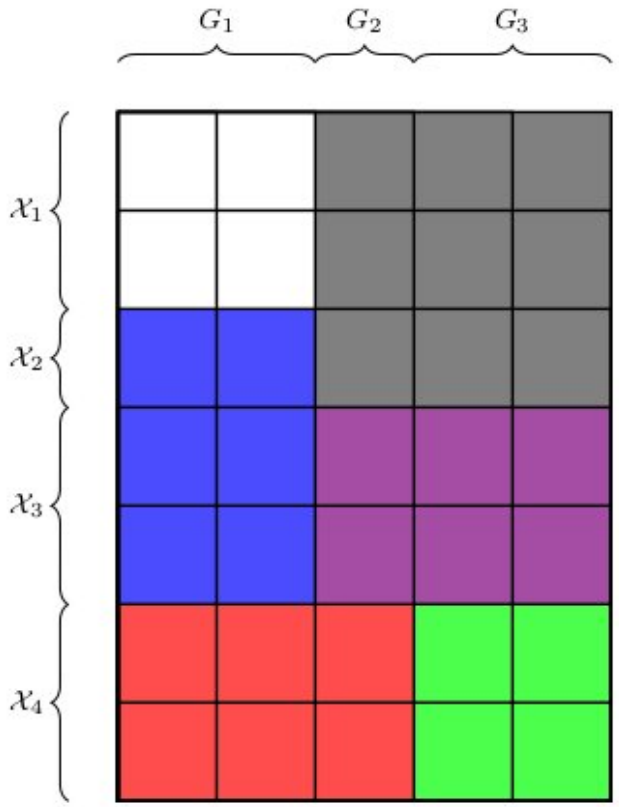
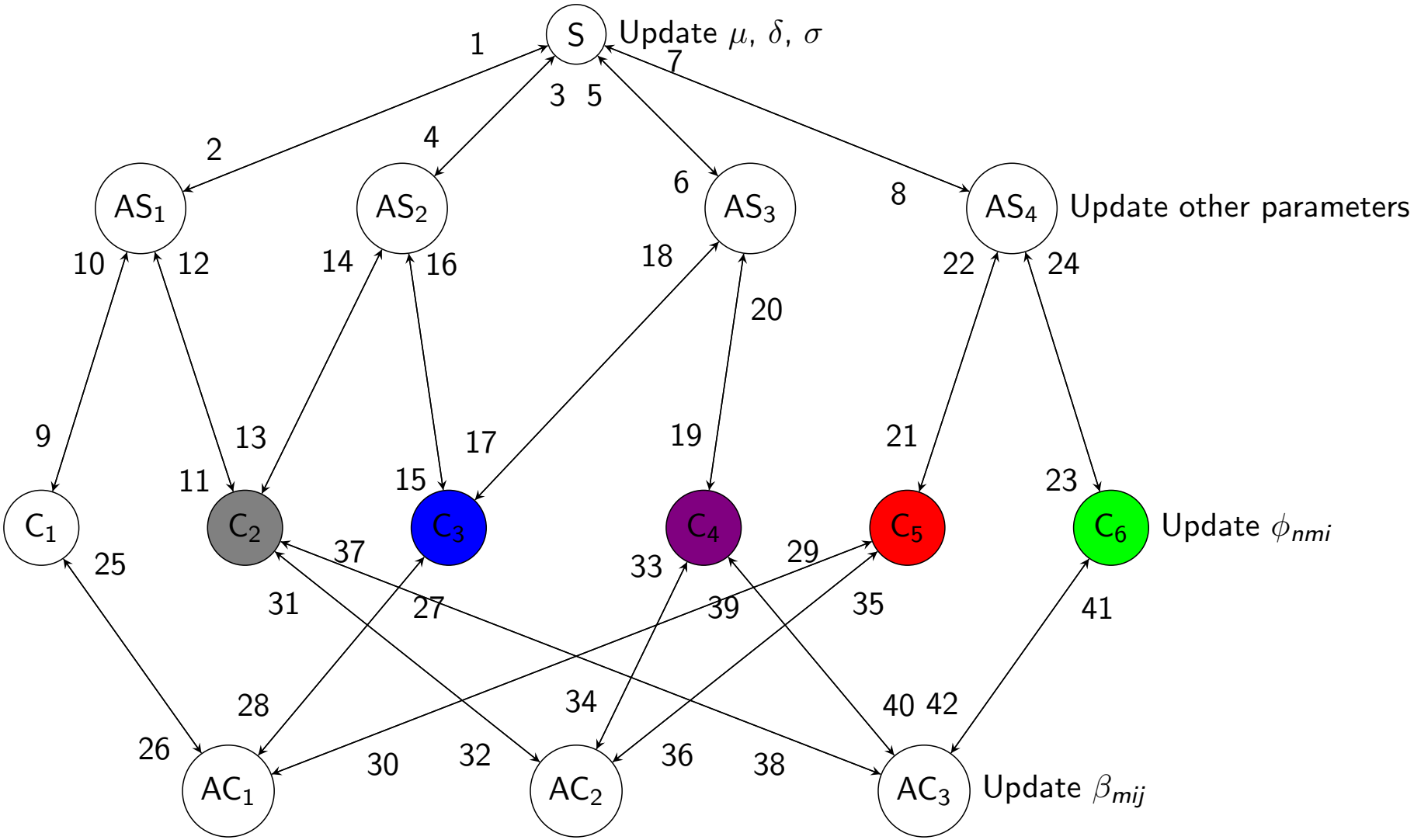
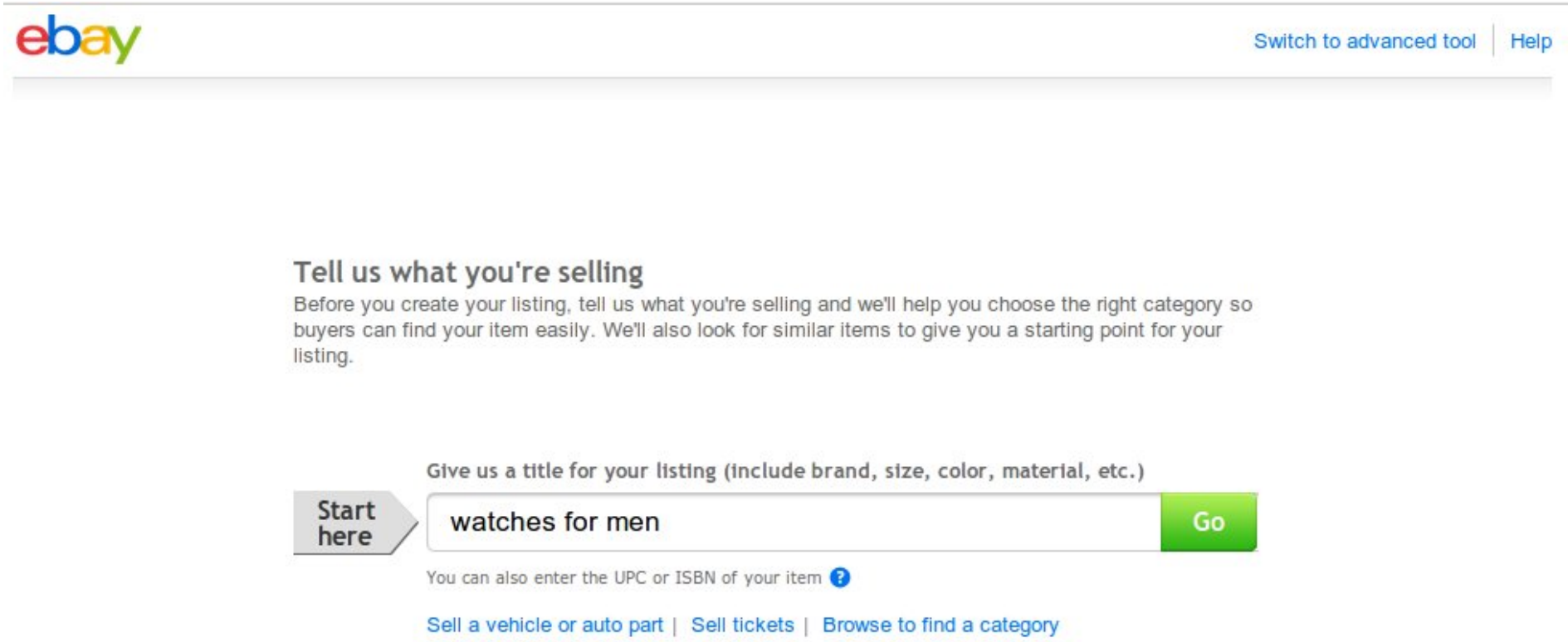



Figure: Arbitrarily Distributed Ensemble

# Parameter Updates in Distributed Learning

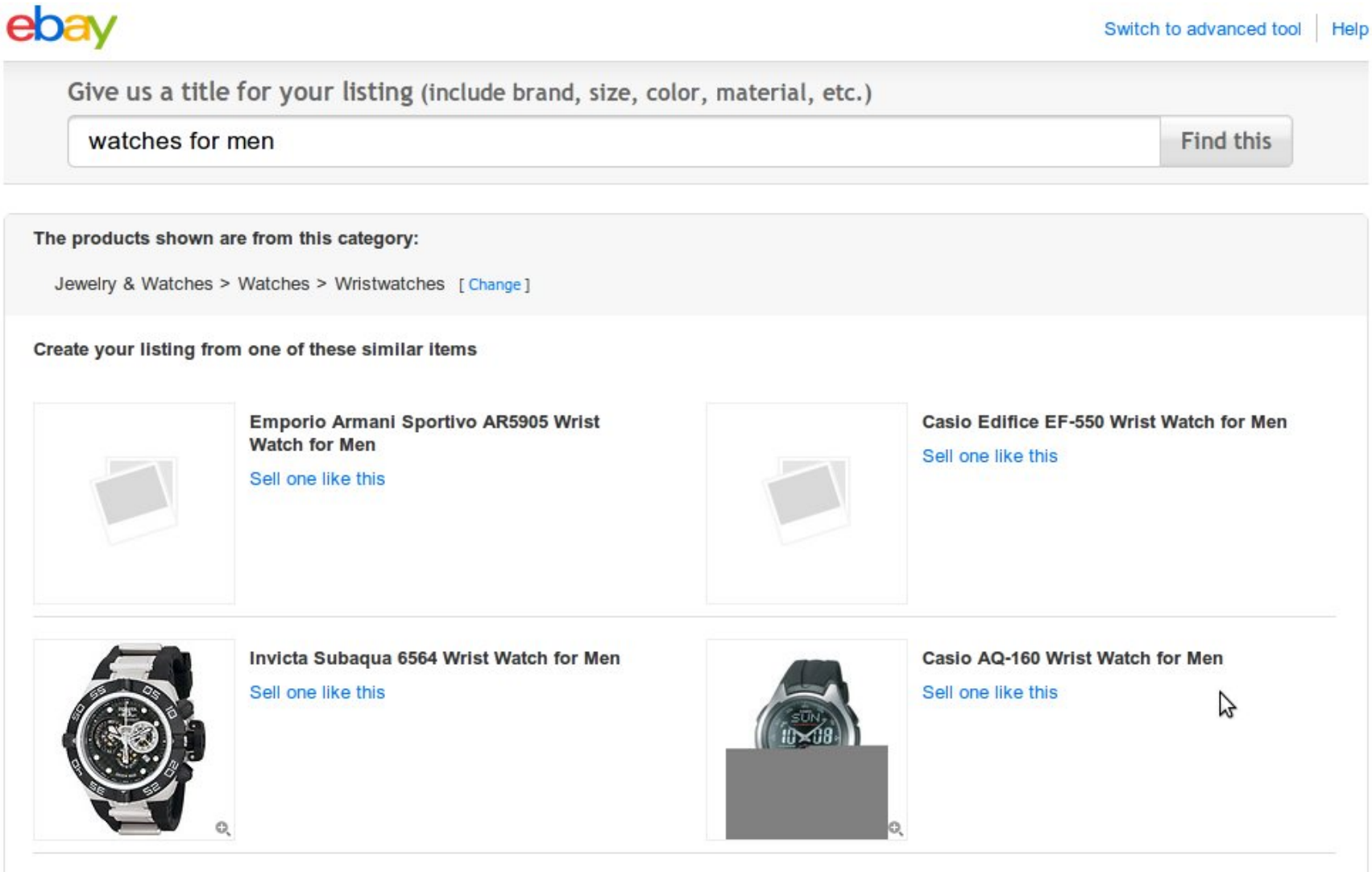


# Dataset from eBay Inc.



# Dataset from eBay Inc. – continued

39 top level nodes called *meta-categories* and 20K+ bottom level nodes called *leaf categories*.



# Transfer learning on text data from eBay Inc.

| Group ID | $\lvert\mathcal{X}\rvert$ | $k$-NN | BGCM | LWE | C$^3$E-Ideal | BC$^3$E |
|---|---|---|---|---|---|---|
| 42 | 1299 | 64.90 | 73.78 ($\pm$ 0.94) | 76.86 ($\pm$ 1.01) | 83.99 ($\pm$ 0.41) | 83.68 ($\pm$ 1.09) |
| 84 | 611 | 63.67 | 69.23 ($\pm$ 0.17) | 75.24 ($\pm$ 0.26) | 81.18 ($\pm$ 0.16) | 76.27 ($\pm$ 1.31) |
| 86 | 2381 | 77.66 | 84.33 ($\pm$ 2.74) | 83.29 ($\pm$ 1.02) | 92.78 ($\pm$ 0.35) | 87.20 ($\pm$ 0.91) |
| 67 | 789 | 72.75 | 72.75 ($\pm$ 0.07) | 78.03 ($\pm$ 0.72) | 82.64 ($\pm$ 0.82) | 81.75 ($\pm$ 1.37) |
| 52 | 1076 | 76.95 | 77.01 ($\pm$ 1.18) | 77.49 ($\pm$ 1.41) | 88.38 ($\pm$ 0.22) | 85.04 ($\pm$ 2.14) |
| 99 | 827 | 84.04 | 85.12 ($\pm$ 0.52) | 86.90 ($\pm$ 0.92) | 91.54 ($\pm$ 0.27) | 91.17 ($\pm$ 0.82) |
| 48 | 3445 | 86.33 | 86.19 ($\pm$ 0.25) | 90.38 ($\pm$ 1.03) | 92.71 ($\pm$ 0.31) | 92.71 ($\pm$ 1.16) |
| 94 | 440 | 79.32 | 81.08 ($\pm$ 0.73) | 82.52 ($\pm$ 0.83) | 85.45 ($\pm$ 0.09) | 85.45 ($\pm$ 0.79) |
| 35 | 4907 | 82.41 | 82.10 ($\pm$ 0.37) | 85.08 ($\pm$ 1.39) | 88.16 ($\pm$ 0.17) | 88.22 ($\pm$ 1.21) |
| 45 | 1952 | 74.80 | 73.12 ($\pm$ 0.81) | 73.64 ($\pm$ 1.68) | 84.32 ($\pm$ 0.23) | 77.97 ($\pm$ 0.47) |

Table: Performance of **BC$^3$E** on text classification data — Avg. Accuracies $\pm$(Standard Deviations).

# Semisupervised learning on UCI Data

| Dataset (% of tr. data) | $\lvert\mathcal{X}\rvert$ | Ensemble | Best | BGCM | $C^3E$ | $BC^3E$ |
|---|---|---|---|---|---|---|
| Half-moon(2%) | 784 | 92.53($\pm$1.83) | 93.02($\pm$0.82) | 92.16($\pm$1.47) | **99.64**($\pm$0.08) | 98.23($\pm$2.03) |
| Circles(2%) | 1568 | 60.03($\pm$8.44) | 95.74($\pm$5.15) | 78.67($\pm$0.54) | **99.61**($\pm$0.83) | 97.91($\pm$0.74) |
| Pima(2%) | 745 | 68.16($\pm$5.05) | 69.93($\pm$3.68) | 69.21($\pm$4.83) | 70.31($\pm$4.44) | **72.83**($\pm$0.49) |
| Heart(7%) | 251 | 77.77($\pm$2.55) | 79.22($\pm$2.20) | 82.78($\pm$4.82) | **82.85**($\pm$5.25) | 82.53($\pm$1.14) |
| G. Numer(10%) | 900 | 70.96($\pm$1.00) | 70.19($\pm$1.52) | 73.70($\pm$1.06) | 74.44($\pm$3.44) | **74.61**($\pm$1.62) |
| Wine(10%) | 900 | 79.87($\pm$5.68) | 80.37($\pm$5.47) | 75.37($\pm$13.66) | **83.62**($\pm$6.27) | 82.20($\pm$1.07) |

Table: Comparison of **$BC^3E$** with **$C^3E$** and **BGCM** — Avg. Accuracies $\pm$(Standard Deviations).

# Questions?

References:


- An Optimization Framework for Semi-Supervised and Transfer Learning using Multiple Classifiers and Clusterers, Acharya *et. al.* [Link].
- Knowledge Transfer via Multiple Model Local Structure Mapping, Gao *et. al.* [Link].
- Bayesian Cluster Ensembles, Wang *et. al.* [Link].
- Probabilistic Combination of Classifier and Cluster Ensembles for Non-transductive Learning [Link].