Domain Adaptation with Ensemble of Feature Groups

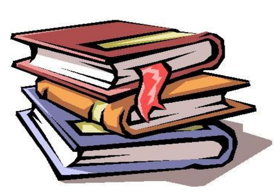
Rajhans Samdani[1] and Scott Wen-Tau Yih[2]

[1]University of Illinois at Urbana-Champaign [2]Microsoft Research

Simple and Effective Algorithm for Domain Adaptation

Domain Adaptation (DA)

- Supervision for one domain (Source $\mathcal{S}$): e.g. movie review sentiment detection
- Little/no supervision for different but related domain (Target $\mathcal{T}$): book reviews

Source  DA  Target

- Use data on S to improve accuracy on T
- Pervasive: labeled data is scarce!!!
- We show experiments on:
  - DA for Sentiment Analysis
  - Email spam detection: spammers adapt by hacking new computers and changing spam text – need for DA for detection

Our Approach: FEAD

- Feature Ensemble for Domain Adaptation
- Outputs weighted ensemble of classifiers based on different groups of features
  1. Classifiers are largely trained on source data (plus some labeled target data)
  2. Weights are tuned on small amount of labeled target data (not used in 1)
- Fast and very easy to implement
- Allows for incorporating knowledge about features via. feature groups
- Beats state-of-the-art algorithms experimentally

Feature Groups: Concept Drift

- Moving from source to target, features undergo a change in distribution
  - Features x_i change little – reliable
  - Features x_j change a lot – unreliable
- E.g. Email Spam Detection: time changes – domain changes
  - Email text features: easy to change; unreliable across time
  - Sender-ip features: hard to hack new computers; reliable
- Use "feature groups" as learned units with similar cross-domain behavior and learn ensemble

Our Algorithm: FEAD

Given: Data: L_S and L_T; feature groups: $X_0, \ldots, X_r$; convex loss function: Δ^c

// Local classifiers learned on source

for $i = 0$ to r **do**

  learn: $h_S^i \leftarrow \arg\min_h \Delta^c_{L_S}(h)$

end for

// Ensemble weights tuned on target

$\alpha_0, \ldots, \alpha_r \leftarrow \arg\min_{\alpha'_0, \ldots, \alpha'_r \geq 0} \sum_{(\mathbf{x}, y) \in L_T} \Delta^c(\sum_i \alpha'_i h_S^i(\mathbf{x}), y)$

// Local classifiers "re-learned" on source+target

for $i = 0$ to r **do**

  re-learn $h_S^i \leftarrow \arg\min_h \Delta^c_{L_S \cup L_T}(h)$

end for

return $\mathbf{w_t} = \sum_{i=0}^r \alpha_i h_S^i$ // final weighted ensemble

Discover Feature Groups by

- Domain knowledge: different feature generating functions
  - E.g. Email spam detection: email, sender-ip features, user-id features
- Simple ad-hoc measures: mutual information, frequency, etc.
  - E.g. Sentiment detection: features based on mutual information and frequency

Generalization Bound for Ensemble h

Expected error on target distribution

Ensemble coefficients for feature group i

Empirical source error for feature group i

d_H-distance between source and the source distribution of feature group i

$$\boxed{\Delta_{D_T}(h)} \leq (1-\beta)\Big(\epsilon^* + \sum_i \boxed{\alpha_i}\big(\boxed{\epsilon_i^S} + \boxed{d_H(D_S, D_S^i)}$$
$$+ \boxed{d_H(D_S^i, D_T)}\big)\Big) + \epsilon(p_H, \beta, \delta)$$
$$+ \beta \boxed{\Delta_{L_T}(h)}$$

d_H-distance between target and the source distribution of feature group i

Empirical target error

- d_h-distance: a measure of distance between distributions (Ben-David et al., 2006)
- Ensemble coefficients balance empirical target error, empirical source error(s), and distribution distances

Experiments: Sentiment Analysis

- 2000 labeled reviews for Movies, Books, Kitchen Appliances, DVDs and Electronics (Blitzer et al. 2006, Pang et al. 2002)
- Baselines: Logistic Regression on all data, EasyAdapt, and Multiview-Transfer
- Feature groups:
  - *Total*: All features
  - *Frequent*: features which have high M.I. with output labels in source *and* occur frequently in target
- FEAD stat. significantly better than
  - Multiview-Transfer and EA in 8 cases
  - Better than LR in 5 cases

Experiments: Sentiment Analysis

Setting	Algorithms			
Src-Tgt	LR	Multi-T	EA	FEAD
B-E	78.88	78.82	**79.38**	79.34
B-D	81.10	80.01	78.63	**81.80**
B-K	**82.29**	79.60	81.66	82.26
B-M	80.23	77.91	79.25	**80.70**
E-B	75.34	72.74	**75.85**	75.60
E-D	75.85	75.86	74.78	**76.76**
E-K	86.50	83.77	85.33	**87.59**
E-M	72.60	70.86	72.63	**73.54**
D-B	81.60	79.91	80.14	**82.46**
D-E	81.27	81.09	80.34	**81.54**
D-K	**82.95**	81.83	82.08	82.81
D-M	82.53	79.50	81.53	**82.53**
K-B	74.74	75.47	74.78	**75.75**
K-E	84.90	84.57	83.81	**85.24**
K-D	75.93	**76.97**	75.21	76.88
K-M	72.38	71.02	70.45	**72.62**
M-B	77.11	77.06	76.07	**78.88**
M-E	76.45	**80.18**	76.50	77.62
M-D	77.76	77.94	76.20	**79.52**
M-K	76.72	**79.41**	76.48	77.59
Avg.	78.86	78.23	78.06	**79.55**

Experiments: Spam Detection

- Hotmail Data: using historical data hence need DA as spam evolves
- Total 915,000 messages: first 765,000 as source data, next 30,000 as target tuning data, last 120,000 as target test data
- Feature Groups: Email-features, Sender-features, Use-features
- Evaluation: ROC curve at low FPR values