Multi-Task Learning for Boosting with Application to Web Search Ranking

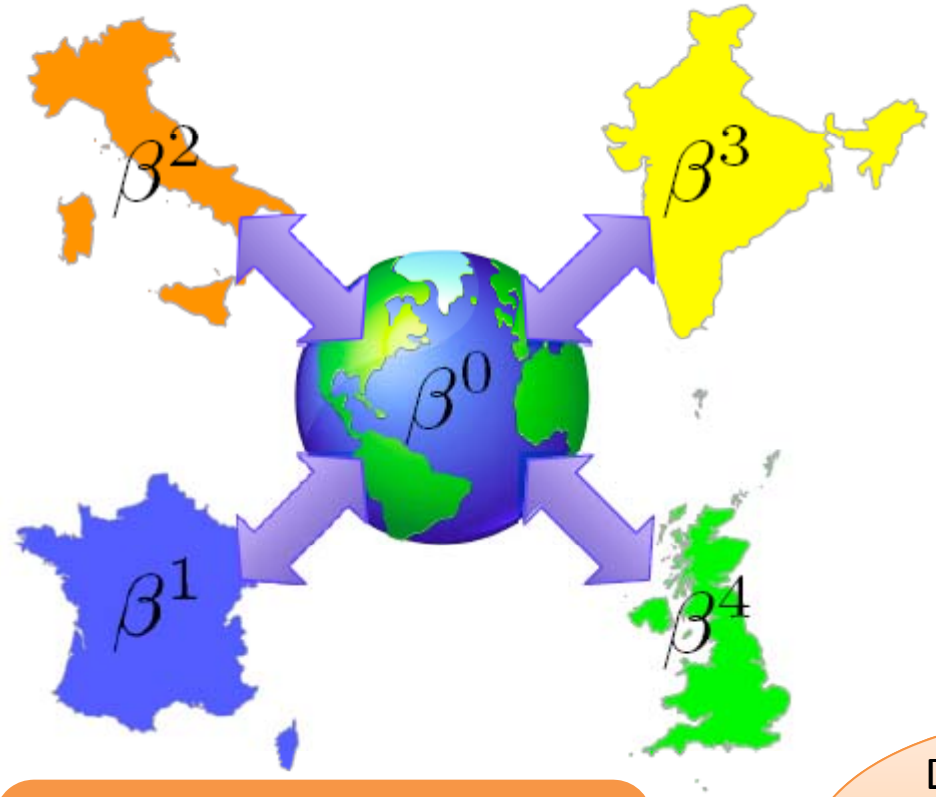
Olivier Chapelle et al.

Presenter: Wei Cheng

Outline

- Motivation
- Backgrounds
- Algorithm
  - From svm to boosting using L1 regularization
  - Є-boosting for optimization
  - Overall algorithm
- Evaluation
- Overall review and new research points discussion

Motivation



Different/same search engine(s) for different countries?

Domain specific engine is better!
e.g. ‘gelivable’
(very useful)

Motivation

- Should we train ranking model separately?
  - Corps in some domains might be too small to train a good model
  - Solution:

Multi-task learning

Backgrounds

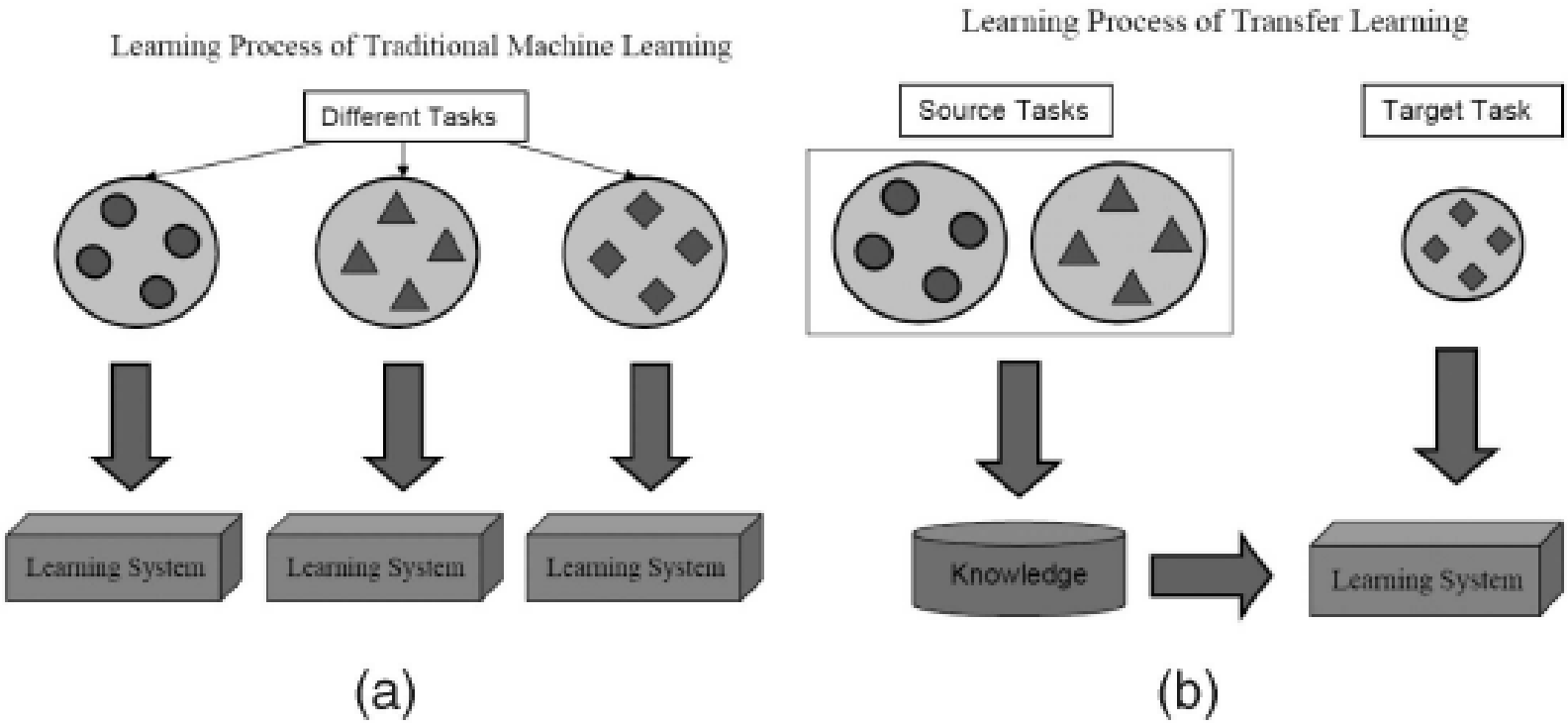


Fig. 1. Different learning processes between (a) traditional machine learning and (b) transfer learning.

Sinno Jialin Pan and Qiang Yang, TKDE'2010

Backgrounds

Different Settings of Transfer Learning

Transfer Learning Settings	Related Areas	Source Domain Labels	Target Domain Labels	Tasks
Inductive Transfer Learning	Multi-task Learning	Available	Available	Regression, Classification
	Self-taught Learning	Unavailable	Available	Regression, Classification
Transductive Transfer Learning	Domain Adaptation, Sample Selection Bias, Co-variate Shift	Available	Unavailable	Regression, Classification
Unsupervised Transfer Learning		Unavailable	Unavailable	Clustering, Dimensionality Reduction

Sinno Jialin Pan and Qiang Yang, TKDE'2010

Backgrounds

- Why transfer learning works?

Table 1: Cross-domain sentiment classification examples: reviews of *electronics* and *video games* products. Boldfaces are domain-specific words, which are much more frequent in one domain than in the other one. Italic words are some domain-independent words, which occur frequently in both domains. "+" denotes positive sentiment, and "-" denotes negative sentiment.

	electronics	*video games*
+	**Compact**; easy to operate; very *good* picture quality; looks **sharp**!	A very *good* game! It is action packed and full of *excitement*. I am very much **hooked** on this game.
+	I purchased this unit from Circuit City and I was very *excited* about the quality of the picture. It is really *nice* and **sharp**.	Very **realistic** shooting action and *good* plots. We played this and were **hooked**.
-	It is also quite **blurry** in very dark settings. I will *never buy* HP again.	The game is so **boring**. I am extremely unhappy and will probably *never buy* UbiSoft again.

Sinno Jialin Pan *et al. At WWW'10*

Backgrounds

- Why transfer learning works?(continue)

TRANSFER LEARNING DATA SETS (GENERATED FROM 20-NEWSGROUPS AND REUTERS-21578)

Data Set	$\mathcal{D}_s$	$\mathcal{D}_t$
comp vs rec	comp.graphics, comp.os.* rec.autos, rec.motorcycles	comp.sys.* rec.sport.*
comp vs sci	comp.graphics, comp.os.* sci.crypt, sci.med	comp.sys.* sci.electronics, sci.space
rec vs sci	rec.autos, rec.motorcycles sci.crypt, sci.med	rec.sport.* sci.electronics, sci.space
orgs vs people	orgs.{s}, people.{s}	orgs.{t}, people.{t}
orgs vs place	orgs.{s}, place.{s}	orgs.{t}, place.{t}
people vs place	people.{s}, place.{s}	people.{t}, place.{t}

Backgrounds

- Why transfer learning works?(continue)

Table: Average Learning Accuracy (%) based on 10 Repeated Runs

Accuracy	K-means	SVM	TSVM	CoCC	MTrick	**CCI**
comp vs rec	85.64	86.81	90.60	91.00	**97.18**	**97.42**
comp vs sci	68.92	62.02	67.63	80.80	82.54	**87.28**
rec vs sci	80.58	76.64	86.05	84.00	93.60	**94.88**
orgs vs people	69.74	69.34	73.80	76.40	74.69	**78.46**
orgs vs place	68.98	69.98	69.89	68.80	70.15	**72.15**
people vs place	60.37	56.94	58.43	66.94	63.31	**67.73**

Backgrounds

traditional learning

Input:

LearnerA

LearnerB

Target:

Dog/human

Girl/boy

Backgrounds

Multi-task learning

Input:

Joint Learning Task

Target:

Dog/human

Girl/boy

Algorithm

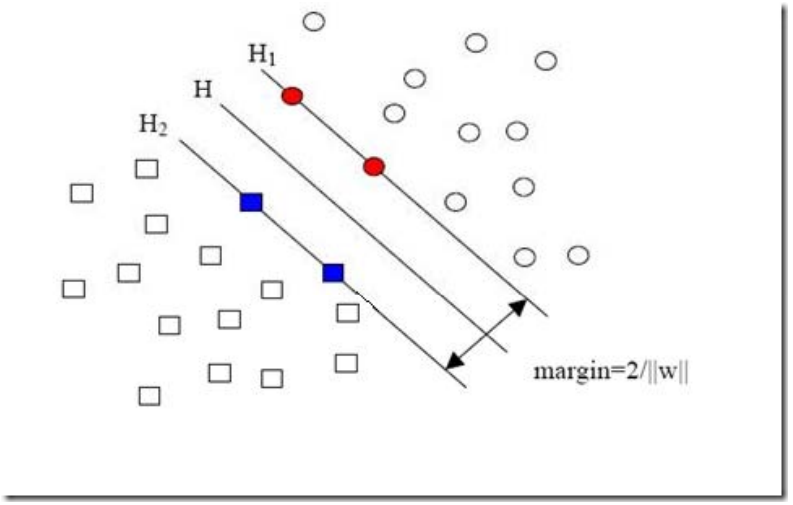
- The algorithm aims at designing an algorithm based on gradient boosted decision trees
- Inspired by svm based multi-task solution and boosting-trick.
- Using Є-boosting for optimization

Algorithm

- From svm to boosting using L1 regularization
  - Previous svm based multi-task learning:

$$\min_{w^0, w^1, \ldots, w^T} \sum_{t=0}^{T} \lambda_t \|w^t\|_2^2 + \sum_{t=1}^{T} C^t(w^0 + w^t)$$

$$C^t(w) = \sum_{i \in I^t} \max(0, 1 - y_i \langle w, x_i \rangle).$$



Algorithm

- Svm(kernel-trick)---→boosting (boosting trick)

Pick set of non-linear functions(e.g., decision trees, regression trees,….)

Apply every single function to each data point

$|H|=J$

X→ϕ(X)

$$\phi(x)=\begin{bmatrix} h_1(x) \\ . \\ . \\ h_J(x) \end{bmatrix}$$

Є-boosting for optimization

- Using L1 regulization

$$\min_{\beta^0,\beta^1,\ldots,\beta^T} \sum_{t=1}^{T} C^t(\beta^0+\beta^t) \quad \text{s.t.} \quad \sum_{t=0}^{T} \lambda_t \|\beta^t\|_1 \le \mu,$$

$$\min_{\|\beta\|_\lambda \le \mu} C(\beta), \qquad \|\beta\|_\lambda = \textstyle\sum_{t=0}^{T} \lambda_t \|\beta^t\|_1$$

Using Є-boosting

$$\min_{\Delta\beta} C(\beta+\Delta\beta) \quad \text{s.t.} \quad \|\Delta\beta\|_\lambda \le \epsilon$$

Algorithm

Algorithm 1 Multi-boost (S iterations)

$F^t = 0 \quad \forall 0 \leq t \leq T$

for $s \leftarrow 1$ to S **do**

$z_i = -\frac{\partial C(u)}{\partial u_i} \quad \forall 1 \leq i \leq n$

$\hat{h}^t \leftarrow \underset{h \in \mathcal{H}}{\operatorname{argmin}} \sum_{i \in I^t} (h(x_i) - z_i)^2, \quad 0 \leq t \leq T.$

$\hat{t} \leftarrow \underset{t}{\operatorname{argmax}} \frac{1}{\lambda_t} \sum_{i \in T^t} \hat{h}^t(x_i) z_i.$

$F^{\hat{t}} \leftarrow F^{\hat{t}} + \epsilon \hat{h}^{\hat{t}}$

$u_i \leftarrow u_i + \epsilon \hat{h}^{\hat{t}}(x_i) \quad \forall i \in I^t$

end for

Predict a new example x of task t as $F^0(x) + F^t(x)$

Evaluation

- Datasets

	Examples			Queries		
Country	Train	Valid	Test	Train	Valid	Test
A	72k	7k	11k	3486	477	600
B	64k	10k	–	4286	563	–
C	74k	4k	11k	5992	298	600
D	108k	12k	14k	7027	383	600
E	162k	14k	11k	7204	586	600
F	74k	11k	11k	7295	486	600
G	57k	5k	15k	7356	238	600
H	137k	11k	12k	7644	807	600
I	95k	12k	12k	8153	835	600
J	166k	12k	11k	11145	586	600
K	62k	10k	20k	11301	548	600
L	307k	–	–	12850	–	–
M	474k	–	–	15666	–	–
N	194k	16k	12k	18331	541	600
O	401k	–	–	33680	–	–

Evaluation

Country	*weighted*	*unweighted*	*pooling*	*cold-start*
A	0.561	1.444	-0.320	-0.282
C	1.135	1.295	0.972	1.252
D	-0.043	-0.233	-1.096	-2.378
E	0.222	0.342	-2.873	-3.624
M	-2.385	-0.029	-1.724	-6.376
N	-0.036	0.705	-1.160	-3.123

Table 2: Percentage change in DCG over *independent* ranking models for various baseline ranking models.

Evaluation

Country	% gain	Best Countries
A	+4.21	C D F H L M N
B	+2.06	N
C	+1.70	A M
D	+2.95	C H L N
E	+0.35	B C F L O
F	+1.43	A B E H L N
H	+1.11	A B D E F L
L	+0.57	A B C E F M
M	+0.45	A C N
N	+1.00	A F L
O	+0.61	A F

Table 3: Percentage improvement over *independent* for the best countries found on the validation set.

Evaluation

Country	Multi-GBDT		Multi-GBRank	
	Unweighted	Weighted	Unweighted	Weighted
A	+1.53	+0.72	+0.69	+0.75
B	+1.81	+1.58	+2.22	+1.64
C	+0.92	+0.52	+0.92	+0.01
D	+4.14	+3.62	+1.77	+1.84
E	-1.37	-1.45	-0.18	-0.91
F	+0.57	+1.80	+1.67	+2.22
G	+4.34	+4.68	+1.74	+0.75
H	+0.34	+0.96	+0.52	+0.85
I	-0.50	-0.80	-0.07	+0.32
J	+0.10	-0.69	+0.74	-0.64
K	+2.37	+2.38	+3.40	+2.01
N	+0.53	-1.23	+0.51	-0.92
Mean	+1.23	+1.01	+1.16	+0.66

Table 4: DCG-5 gains with *Multi-GBDT* and *Multi-GBRank* learning algorithms in two different weighting settings. The gains are over *indepenent-GBDT* and *independent-GBRank* respectively.

Evaluation

Country	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8
A	+0.33	+1.53				+0.46		
B	+1.87	+1.81		+1.60				
C	+1.81	+0.92				+0.64		
D	+2.90	+4.14		+2.29				
E	-1.02		-1.37					-1.24
F	+1.82	+0.57		-0.20				
G	+1.51		+4.34				+3.59	
H	+0.00	+0.34			+0.08			
I	-0.85	-0.50			-0.97			
J	+0.76		+0.10					+0.38
K			+2.37				+1.98	
L			†		†	†		
N			+0.53			-0.09		
O		†						†

Table 5: Improvement of multi-task models with various groupings of countries over *independent* ranking models. Each column in the table indicates a group that includes a subset of countries and each row in the table corresponds to a single country. The numbers in the cell are filled only when the corresponding country is part of the corresponding group. The symbol † indicates that this country was included for training but has not been tested.

Evaluation

Country	Improvements with F^0
A	+2.94
C	-0.20
D	-0.17
E	-0.33
F	+0.83
G	+0.49
H	+1.18
I	+0.73
J	+4.83
K	+0.85
N	-1.48
mean	+ 0.88

Table 7: DCG-5 gains of global models trained with multi-task approach compared with simple data combination from of all countries.

Overall review and new research point discussion

- Contributions:
  - Propose a novel multi-task learning method based on gradient boosted decision tree, which is useful for web-reranking applications. (e.g., personalized search).
  - Have a thorough evaluation on we-scale datasets.
- New research points:
  - Negative transfer: $P_s(\mathbf{x}) \approx P_t(\mathbf{x})$ $P_s(y|\mathbf{x}) \approx P_t(y|\mathbf{x})$
  - Effective grouping: flexible domain adaptation

Q&A

Thanks!