Efficient and Consistent Adversarial Bipartite Matching

Rizal Fathony*, Sima Behpour*, Xinhua Zhang, Brian D. Ziebart | {rfatho2,sbehpo2,zhangx,bziebart}@uic.edu

THE UNIVERSITY OF ILLINOIS AT CHICAGO

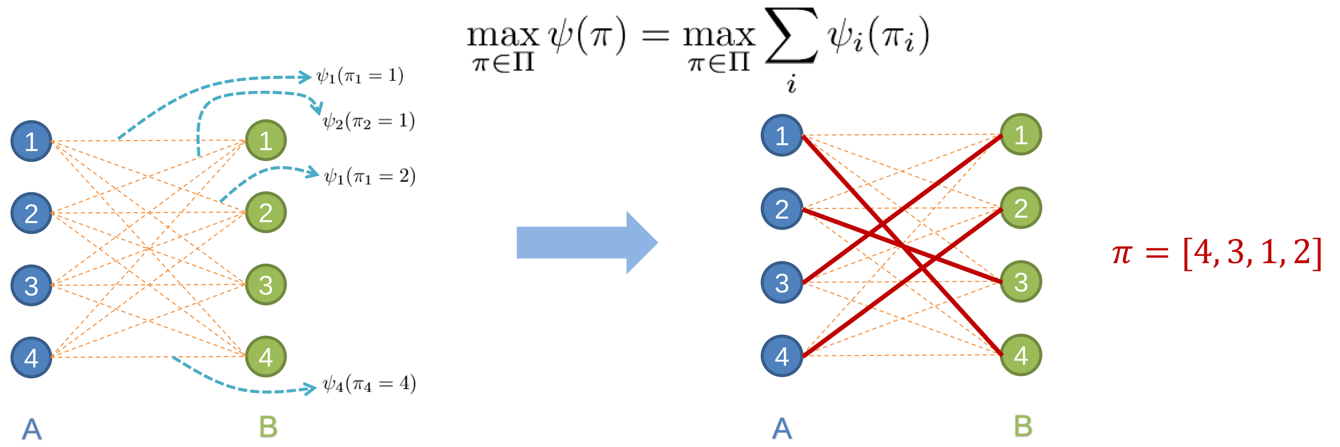
UIC COMPUTER SCIENCE

Bipartite Matching Task

Weighted Bipartite Matching

- Given: ① two sets of elements A and B with equal size,
  ② weights between the elements in A and the elements in B
- Task: find one-to-one mapping that maximize sum of potentials:

$$\max_{\pi\in\Pi} \psi(\pi) = \max_{\pi\in\Pi} \sum_i \psi_i(\pi_i)$$



Learning Bipartite Matching Task

- Given: training data → each sample: a bipartite graph (x) and a ground truth assignment (π)
- Task: learn weight function $\psi_i(\cdot)$ that minimizes miss-assignment metric (e.g. Hamming loss)

Applications:

① Word alignment (Taskar et. al., 2005; Pado & Lapta, 2006; Mac-Cartney et. al., 2008)
② Correspondence between images (Belongie et. al., 2002; Dellaert et. al., 2003)
③ Learning to rank documents (Dwork et. al., 2001; Le & Smola, 2007)

Previous Methods and Shortcomings

Desiderata for a Predictor:

① Efficiency: learning & prediction runtime is in a (low degree) polynomial time
② Consistency: must also minimizes Hamming loss under ideal condition
(given the true distribution and fully expressive model parameters)

① Exponential Family Random Field Approach (Petterson et. al., 2009; Volkovs & Zemel, 2012)

Probabilistic model:

$$P_\psi(\pi) = \frac{1}{Z_\psi} \exp\left(\sum_{i=1}^{n} \psi_i(\pi_i)\right)$$

$$Z_\psi = \sum_\pi \prod_{i=1}^{n} \exp\left(\psi_i(\pi_i)\right) = \mathrm{perm}(\mathbf{M}) \quad \text{where} M_{i,j} = \exp\left(\psi_i(j)\right)$$

Consistent? Yes!
produce Bayes optimal prediction over the Hamming loss in an ideal condition

Efficient? No!
normalization term Z_ψ involves matrix permanent computation (a #P-hard problem)

② Maximum Margin Approach (Tsochantaridis et. al., 2005)

Max-margin model:

$$\min_\psi \mathbb{E}_{\pi\sim\tilde{P}} \left[\max_{\pi'} \{\mathrm{loss}(\pi, \pi') + \psi(\pi')\} - \psi(\pi)\right]$$

$\tilde{P}$ is the empirical distribution

Efficient? Yes!
polynomial algorithm for computing the maximum violated constraint (Hungarian algorithm)

Consistent? No!
based on the CS multiclass SVM: not consistent for distributions with no majority label

Adversarial Bipartite Matching

Formulation

- Our method seeks a predictor that robustly minimizes Hamming loss, against the worst-case permutation mixture probability that is consistent with the statistics of the training data

$$\min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}} [\mathrm{loss}(\hat{\pi}, \check{\pi})]$$

$$\text{s.t. } \mathbb{E}_{x\sim\tilde{P};\check{\pi}|x\sim\check{P}} \left[\sum_{i=1}^{n} \phi_i(x, \check{\pi}_i)\right] = \mathbb{E}_{(x,\pi)\sim\tilde{P}} \left[\sum_{i=1}^{n} \phi_i(x, \pi_i)\right]$$

- Predictor: - makes a probabilistic prediction $\hat{P}(\hat{\pi}|x)$ and aims to minimize the loss
  - is pitted with an adversary instead of the empirical distribution
- Adversary: - makes a probabilistic prediction $\check{P}(\check{\pi}|x)$ and aims to maximize the loss
  - constrained to select probability that match the statistics of empirical distribution ($\tilde{P}$) via moment matching on the features $\phi(x, \pi) = \sum_{i=1}^{n} \phi_i(x, \pi_i)$

Dual Formulation

$$\min_\theta \mathbb{E}_{x,\pi\sim\tilde{P}} \min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{\substack{\hat{\pi}|x\sim\hat{P}\\ \check{\pi}|x\sim\check{P}}} \left[\mathrm{loss}(\hat{\pi}, \check{\pi}) + \theta \cdot \sum_{i=1}^{n} \left(\phi_i(x, \check{\pi}_i) - \phi_i(x, \pi_i)\right)\right]$$

where θ is the dual variable for moment matching constraints

Augmented Hamming loss matrix for $n = 3$ permutations:

	$\check{\pi}=123$	$\check{\pi}=132$	$\check{\pi}=213$	$\check{\pi}=231$	$\check{\pi}=312$	$\check{\pi}=321$
$\hat{\pi}=123$	$0+\delta_{123}$	$2+\delta_{132}$	$2+\delta_{213}$	$3+\delta_{231}$	$3+\delta_{312}$	$2+\delta_{321}$
$\hat{\pi}=132$	$2+\delta_{123}$	$0+\delta_{132}$	$3+\delta_{213}$	$2+\delta_{231}$	$2+\delta_{312}$	$3+\delta_{321}$
$\hat{\pi}=213$	$2+\delta_{123}$	$3+\delta_{132}$	$0+\delta_{213}$	$2+\delta_{231}$	$2+\delta_{312}$	$3+\delta_{321}$
$\hat{\pi}=231$	$3+\delta_{123}$	$2+\delta_{132}$	$2+\delta_{213}$	$0+\delta_{231}$	$3+\delta_{312}$	$2+\delta_{321}$
$\hat{\pi}=312$	$3+\delta_{123}$	$2+\delta_{132}$	$2+\delta_{213}$	$3+\delta_{231}$	$0+\delta_{312}$	$2+\delta_{321}$
$\hat{\pi}=321$	$2+\delta_{123}$	$3+\delta_{132}$	$3+\delta_{213}$	$2+\delta_{231}$	$2+\delta_{312}$	$0+\delta_{321}$

size:
$n! \times n!$

Intractable
for modestly-sized n

Algorithms

① Double Oracle Method

- Based on the observation: equilibrium is usually supported by small number of permutations
- Iterative method: - start from a single permutation for each player
  - alternately: * compute predictor's (/adversary's) strategy in the current game
    * compute adversary's (/predictor's) best response, add to the game
  - until no improvement in the game value
- Use the game solution to compute the gradient and perform gradient update
- No formal polynomial bound is known → the whole runtime cannot be characterized as polynomial

② Marginal Distribution Formulation

- Reformulation: from permutation mixture distributions to marginal distributions:

Predictor

$$\mathbf{P} = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline \hat{\pi}_1 & p_{1,1} & p_{1,2} & p_{1,3} \\ \hat{\pi}_2 & p_{2,1} & p_{2,2} & p_{2,3} \\ \hat{\pi}_3 & p_{3,1} & p_{3,2} & p_{3,3} \end{array}$$

$p_{i,j} = \hat{P}(\hat{\pi}_i = j)$

Adversary

$$\mathbf{Q} = \begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline \check{\pi}_1 & q_{1,1} & q_{1,2} & q_{1,3} \\ \check{\pi}_2 & q_{2,1} & q_{2,2} & q_{2,3} \\ \check{\pi}_3 & q_{3,1} & q_{3,2} & q_{3,3} \end{array}$$

$q_{i,j} = \check{P}(\check{\pi}_i = j)$

- Birkhoff – Von Neumann theorem:

The convex hull of the set of permutations forms a convex polytope whose points are doubly stochastic matrices: $\mathbf{P1} = \mathbf{P}^\top\mathbf{1} = \mathbf{Q1} = \mathbf{Q}^\top\mathbf{1} = \mathbf{1}$

- Reduce the space of optimization from $O(n!)$ to $O(n^2)$
- Marginal optimization (after adding regularization and smoothing penalty):

$$\max_{\mathbf{Q}\geq 0} \min_\theta \frac{1}{m} \sum_{i=1}^{m} \min_{\mathbf{P}_i \geq 0} \left[\langle \mathbf{Q}_i - \mathbf{Y}_i, \textstyle\sum_k \theta_k \mathbf{X}_{i,k}\rangle - \langle \mathbf{P}_i, \mathbf{Q}_i\rangle + \frac{\mu}{2}\|\mathbf{P}_i\|_F^2 - \frac{\mu}{2}\|\mathbf{Q}_i\|_F^2\right] + \frac{\lambda}{2}\|\theta\|_2^2$$

$$\text{s.t. : } \mathbf{P}_i\mathbf{1} = \mathbf{P}_i^\top\mathbf{1} = \mathbf{Q}_i\mathbf{1} = \mathbf{Q}_i^\top\mathbf{1} = \mathbf{1}, \quad \forall i$$

- Techniques: - Outer (Q): projected Quasi-Newton with a projection to doubly-stochastic matrix
  - Inner (θ): closed-form solution
  - Inner (P): projection to doubly-stochastic matrix
  - Projection to doubly-stochastic matrix : ADMM

Consistency

Empirical Risk Perspective of Adversarial Bipartite Matching

- Adversarial Bipartite Matching can be viewed as an ERM method with surrogate loss AL_f^{perm}

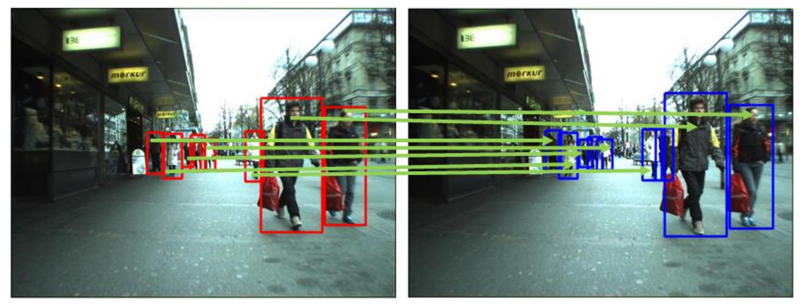
$$\min_\theta \mathbb{E}_{\substack{x\sim P\\ \pi|x\sim\tilde{P}}} \left[AL_{f_\theta}^{\mathrm{perm}}(x, \pi)\right]$$

$$\text{where: } AL_{f_\theta}^{\mathrm{perm}}(x, \pi) \triangleq \min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{\substack{\hat{\pi}|x\sim\hat{P}\\ \check{\pi}|x\sim\check{P}}} \left[\mathrm{loss}(\hat{\pi}, \check{\pi}) + f_\theta(x, \check{\pi}) - f_\theta(x, \pi)\right]$$

- We show that minimizing AL_f^{perm} also minimizes the Hamming loss given true distribution, and f is optimized over the set of all measurable functions on the input space (x, π)
- The consistency result also holds when f is an additive function over individual assignment π_i

Experiments

Experiments | Video Tracking Tasks



- Predict object correspondence in two different frames
- Public benchmark datasets
- 5 datasets in 2 groups (TUD and ETH)
- 48 different features for each pair of objects
- Train on one dataset, test on another dataset from the same group

Experiment results

Table 4. The mean and standard deviation (in parenthesis) of the average accuracy (1 - the average Hamming loss) for the adversarial bipartite matching model compared with Structured-SVM.

Training/ Testing	Adv DO	Adv Marg.	SSVM	Adv DO #Perm.
Campus/ Stadtmitte	0.662 (0.09)	0.662 (0.08)	0.662 (0.08)	11.4
Stadtmitte/ Campus	0.672 (0.12)	0.667 (0.11)	0.660 (0.12)	7.4
Bahnhof/ Sunnyday	**0.758** (0.12)	**0.754** (0.10)	0.729 (0.15)	5.8
Pedcross2/ Sunnyday	**0.760** (0.08)	**0.750** (0.10)	0.736 (0.13)	8.2
Sunnyday/ Bahnhof	**0.755** (0.20)	**0.751** (0.18)	0.739 (0.20)	9.8
Pedcross2/ Bahnhof	**0.760** (0.12)	**0.763** (0.16)	0.731 (0.21)	10.8
Bahnhof/ Pedcross2	**0.718** (0.16)	**0.714** (0.16)	0.701 (0.18)	8.5
Sunnyday/ Pedcross2	**0.719** (0.18)	**0.712** (0.17)	0.700 (0.18)	14.4

Empirical runtime (until convergence)

Table 5. Running time (in seconds) of the model for various number of elements n with fixed number of samples ($m = 50$)

Dataset	# Elements	Adv Marg.	SSVM
Campus	12	1.96	0.22
Stadtmitte	16	2.46	0.25
Sunnyday	18	2.75	0.15
Pedcross2	30	8.18	0.26
Bahnhof	34	9.79	0.31

Adv. Marginal Formulation:
grows (roughly) quadratically in n

CRF: impractical even for $n = 20$
(Petterson et. al., 2009)

Adv. Double Oracle:
small number of permutations in the equilibrium

Adversarial Bipartite Matching:

6 pairs of datasets: significantly outperforms SSVM

2 pairs of datasets: competitive with SSVM

Conclusions

	Efficient?	Consistent?	Perform well?
Exponential Family Random Field (Petterson et. al., 2009; Volkovs & Zemel, 2012)	✗	✓	?
Maximum Margin (Tsochantaridis et. al., 2005)	✓	✗	??
Adversarial Bipartite Matching (our approach)	✓	✓	✓

Acknowledgement

This research was supported in part by NSF Grants RI-#1526379 and CAREER-#1652530.

*equal contribution