Extreme Multi-Label Classification for Ad Targeting using Factorization Machines

Martin Pavlovski[1], Srinath Ravindran[1], Djordje Gligorijevic[2,*], Shubham Agrawal[1], Ivan Stojkovic[1], Nelson Segura-Nunez[1], Jelena Gligorijevic[1]

[1] Yahoo Research, Yahoo Inc. [2] eBay

* *This work was done when the author was at Yahoo.*

Challenges

- Handling Large number of labels (segments)
- Prediction latency / SLA requirements
- (Near) Real-Time prediction
- Handling relationship among features/labels

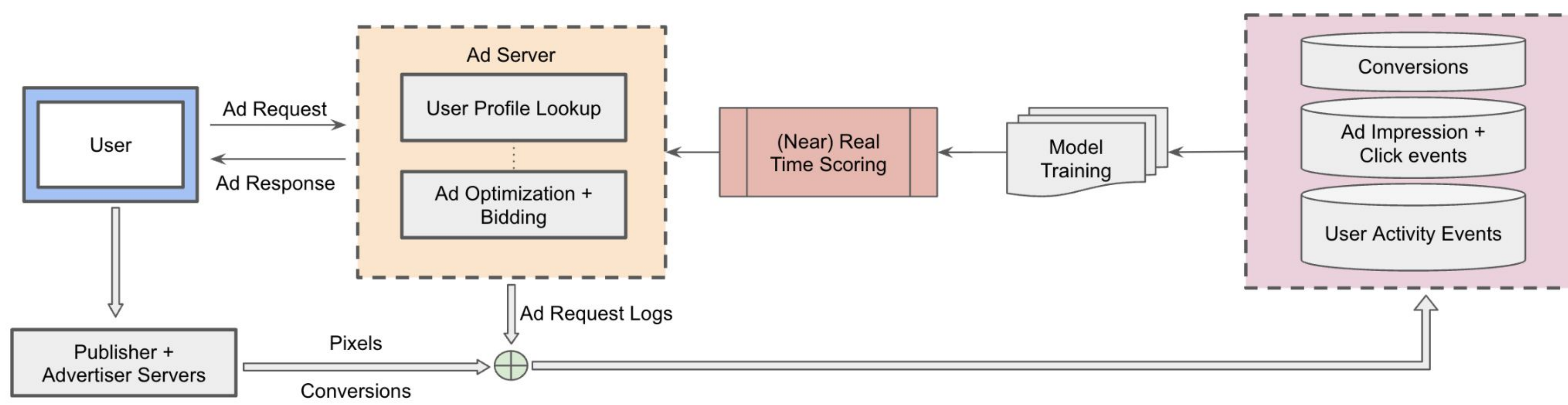
Assigning users to a large number of targeting segments

↓

An extreme multi-label classification (XMLC) problem

Contributions

- Generalization of FM models to XMLC problems (MLFM)
- A lightweight formulation of MLFM (*asymptotically* linear time)
- Extensive experiments
  - Comparison to several OVA and XMLC baselines
  - Evaluation on both benchmark and proprietary datasets
  - Demonstration of computational efficiency
- Application to ad targeting involving a large number of segments



Problem Formulation

Given: A dataset containing $(\mathbf{x},\mathbf{y})$ pairs, where $\mathbf{x}$ has with D features and $\hat{D}$ fields (depending on the application) such that only one feature value x_i can be active per field $F(i)$ and a feature f_i belongs to one and only one field $F(i)$. $\mathbf{y} \in \{0,1\}^L$ denotes the labels; multiple labels can be active for $\mathbf{x}$.

Objective

Learn a function $\phi : \mathrm{R}^D \rightarrow \mathrm{R}^L$ that maps an example $\mathbf{x}$ to a probability vector $[P(y_1|\mathbf{x}), \ldots, P(y_L|\mathbf{x})]$.

Related Approaches

$$\phi_{LR}(\mathbf{x}) = w_0 + \sum_{i=1}^{D} w_i x_i$$

$$\phi_{Poly2}(\mathbf{x}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=i+1}^{D} w_{ij} x_i x_j$$

$$\phi_{FM}(\mathbf{x}) = w_0 + \sum_{i=1}^{D} w_i x_i + \sum_{i=1}^{D}\sum_{j=i+1}^{D} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j$$

Multi-Label Factorization Machine (MLFM)

MLFM – *Original formulation*

$$\phi_{MLFM}(\mathbf{x}) = \mathbf{w}_0 + \left[\sum_{i=1}^{D} w_i^l x_i\right]_{l=1}^{L} + \left[\sum_{i=1}^{D}\sum_{j=i+1}^{D} r^l_{F(i)F(j)} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i x_j\right]_{l=1}^{L}$$

MLFM – *Alternative (lightweight) formulation*

$$\phi'_{MLFM}(\mathbf{x}) = \left[\sum_{i=0}^{D} w_i^l x_i\right]_{l=1}^{L} + \left[\frac{1}{2}\sum_{k=1}^{H*M}\left(\left(\sum_{i=1}^{D} q^l_{ik} x_i\right)^2 - \sum_{i=1}^{D}\left(q^l_{ik} x_i\right)^2\right)\right]_{l=1}^{L}$$

M: feature embedding dimension and H: field embedding dimension

Parameter Learning

CCE loss $\ell(\mathbf{x}_n, \mathbf{y}_n)$ for the n-th data point:

$$-\frac{1}{L}\left(\sum_{l=1}^{L} y_{nl}(\log(p_{nl})) + (1-y_{nl})(1-\log(p_{nl}))\right)$$

where $p = P(y|\mathbf{x}) = 1/(1+e^{-\phi(\mathbf{x})})$.

The optimal parameters $\mathbf{w}_0$, w^l_i, $r^l_{F(i)F(j)}$, $\mathbf{v}_i$ are obtained by minimizing $\Sigma\, \ell(\mathbf{x}_n, \mathbf{y}_n)$.

Inference Time Complexity

$O(DL + LDMH)$

*For sparse multi-field categorical data (where $\hat{D} \ll D$)

$O(\hat{D}L + L\hat{D}MH) \approx O(L\hat{D}MH)$

Experiments

Model	MediaMill				RCV1				EURLex			
	AUC		Inference time		AUC		Inference time		AUC		Inference time	
	Macro	Stratified	CPU	GPU	Macro	Stratified	CPU	GPU	Macro	Stratified	CPU	GPU
OVA-LR	0.6582	0.6689	0.0110	0.0053	0.6197	0.9466	0.2612	0.0104	0.7900	0.9168	6.6999	0.0133
OVA-SVM	0.5090	0.4944	0.0834	–	0.7697	0.7490	5.3943	–	0.7723	0.7474	67.2408	–
OVA-MLP	0.6141	0.7130	0.0443	0.0085	0.9020	0.9618	0.3896	0.0176	0.8967	0.9703	6.7850	0.0237
FastXML	0.6789	0.7946	0.1071	–	0.6906	0.9485	0.2992	–	0.8359	0.9448	0.9486	–
PfastreXML	*0.8354*	*0.8081*	0.4330	–	**0.9354**	**0.9873**	1.8260	–	*0.9282*	*0.9734*	2.8259	–
Parabel	0.7963	0.8024	0.0536	–	0.8439	0.9623	0.7467	–	0.8687	0.9558	2.3686	–
MLFM	**0.8456**	**0.8248**	0.0200	0.0113	*0.9179*	***0.9808***	0.3955	0.0192	**0.9613**	**0.9847**	7.9049	0.0298

Model	Macro Test AUC	Strat. Test AUC
OVA-LR	0.7928	0.7477
OVA-SVM	0.7707	0.7028
FastXML	0.8011	0.7528
PfastreXML	0.8065	0.7804
Parabel	0.8239	0.7892
MLFM	**0.8421**	**0.8006**

Model	Parameters	Dimensions	Example	Size
OVA-LR	Feature weights & Bias terms	$(D+1) \times L$	$200\text{K} \times 1098$	~**3.9 GB**
MLFM	Feature weights	$D \times L$	$200\text{K} \times 1098$	3.9 GB
	Bias terms	$1 \times L$	1×1098	20.4 KB
	Feature embeddings	$D \times M$	$200\text{K} \times 10$	48.7 MB
	Interaction weights	$\hat{D}^2 \times L$	324×1098	6.3 MB
	Grand total			~**4.00 GB**

Model	Inference time (ms)
OVA-LR	0.067447
MLFM (original)	0.159827
MLFM (lightweight)	0.101286

yahoo!

yahoo! research