A NON-IID FRAMEWORK FOR COLLABORATIVE FILTERING WITH RESTRICTED BOLTZMANN MACHINES

Kostadin Georgiev, VMware Bulgaria
Preslav Nakov, Qatar Computing Research Institute

ICML, June 17, 2013, Atlanta

Overview

1. Non-IID framework for collaborative filtering
   - based on Restricted Botzmann Machines (RBMs)
   - with user ratings modeled as real values (vs. multinomials)

2. A neighborhood method boosted by RBM

COLLABORATIVE FILTERING

Introduction

- Recommender systems
  - predict user preferences for new items
  - content-based vs. collaborative
- Collaborative filtering (CF)
  - predictions inferred from the preferences of other users
  - N x M user-item matrix of rating values
    - large and highly sparse (e.g., 95% of values are missing)

User-based vs. Item-based CF

- User-based
  - most of the early CF systems
- Item-based
  - *e.g., (Sarwar et al., 2001)*
- Joint user-item based
  - matrix factorization, joint latent factor space
    - (*Salakhutdinov & Mnih, 2008; Koren et al., 2009; Lawrence & Urtasun,2009*);
  - probabilistic latent model
    - (*Langseth & Nielsen, 2012*)

Boltzmann Machines for CF

- Restricted Botzmann Machines (*Salakhutdinov et al., 2007*)
  - user-based
  - ratings as multinomial variables
  - outperforms SVD

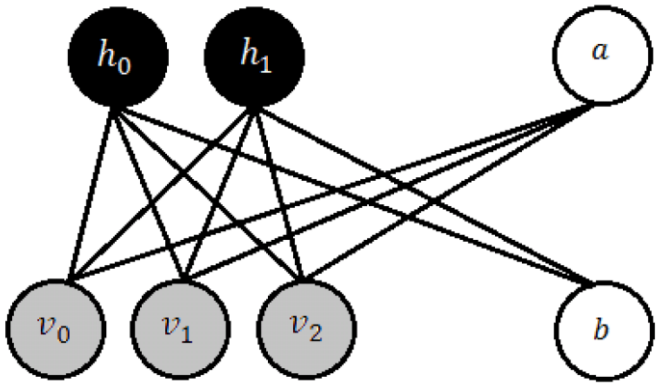
- Unrestricted Boltzmann Machines (*Truyen et al., 2009*)
  - joint user-item based
  - connections between the visible units
  - preprocessing, correlations computation, neighborhood formation
  - ordinal modeling of ratings better than categorical

THIS WORK

Outline

- User-based RBM (U-RBM)
- Item-based RBM (I-RBM)
- Hybrid non-IID RBM (UI-RBM)
- Neighborhood method boosted by I-RBM (I-RBM+INB)

User/Item-based RBM Model

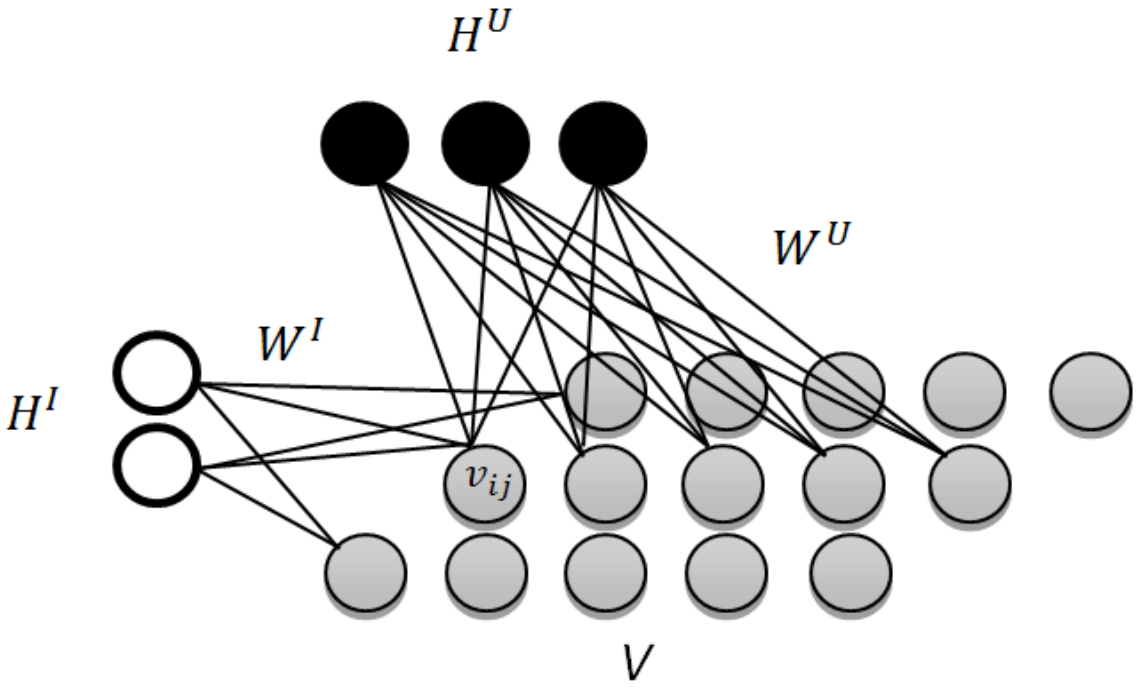


$$P(h_j = 1|v) = L(b_j + \sum_{i=1}^{M} w_{ij} v_i)$$

$$P(v_i|h) = \mathcal{N}(a_i + \sum_{j=1}^{F} w_{ij} h_j, \sigma_i^2)$$

- The visible layer
  - represents either all ratings by a user or all rating for an item
  - units model ratings as real values (vs. multinomial)
  - noise-free reconstruction is better

Non-IID Hybrid RBM Model (1)



- We remove the IID assumption for the training data
- Topology: Unit v_{ij} is connected to two independent hidden layers: one user-based and another item-based.

Non-IID Hybrid RBM Model (2)

- Missing values (ratings): the generated predictions are used during testing, but are ignored during training
- Training procedure: we average the predictions of the user-based and of the item-based RBM models

$$v_{ij} = \frac{1}{2}\left[a_i^U + \sum_{p=1}^{F^U} w_{ip}^U h_{ip}^U + a_j^I + \sum_{q=1}^{F^I} w_{jq}^I h_{jq}^I\right]$$

Neighborhood Boosted by I-RBM

- Use the I-RBM predictions from a neighborhood-based (NB) algorithm

$$w_{ij} = \frac{\sum_{u=1}^{N}(r'_{ui} - \overline{r}_i)(r'_{uj} - \overline{r}_j)}{\sqrt{\sum_{u=1}^{N}(r'_{ui} - \overline{r}_i)^2}\sqrt{\sum_{u=1}^{N}(r'_{uj} - \overline{r}_j)^2}}$$

- However, compute the averages from the original ratings

$$P_{ui} = \overline{r}_i + \frac{\sum_{j=1}^{M}(r_{uj} - \overline{r}_j)w_{ij}}{\sum_{j=1}^{M}|w_{ij}|}$$

EXPERIMENTS AND EVALUATION

Data

- Two *MovieLens* datasets:
  - **100k:**
    - 1,682 movies assigned
    - 943 users
    - 100,000 ratings
    - sparseness: 93.7%
  - **1M:**
    - 3,952 movies
    - 6,040 users
    - 1 million ratings
    - sparseness: 95.8%
- Each rating is an integer between 1 (worst) and 5 (best)

Evaluation

- Mean Absolute Error (MAE):

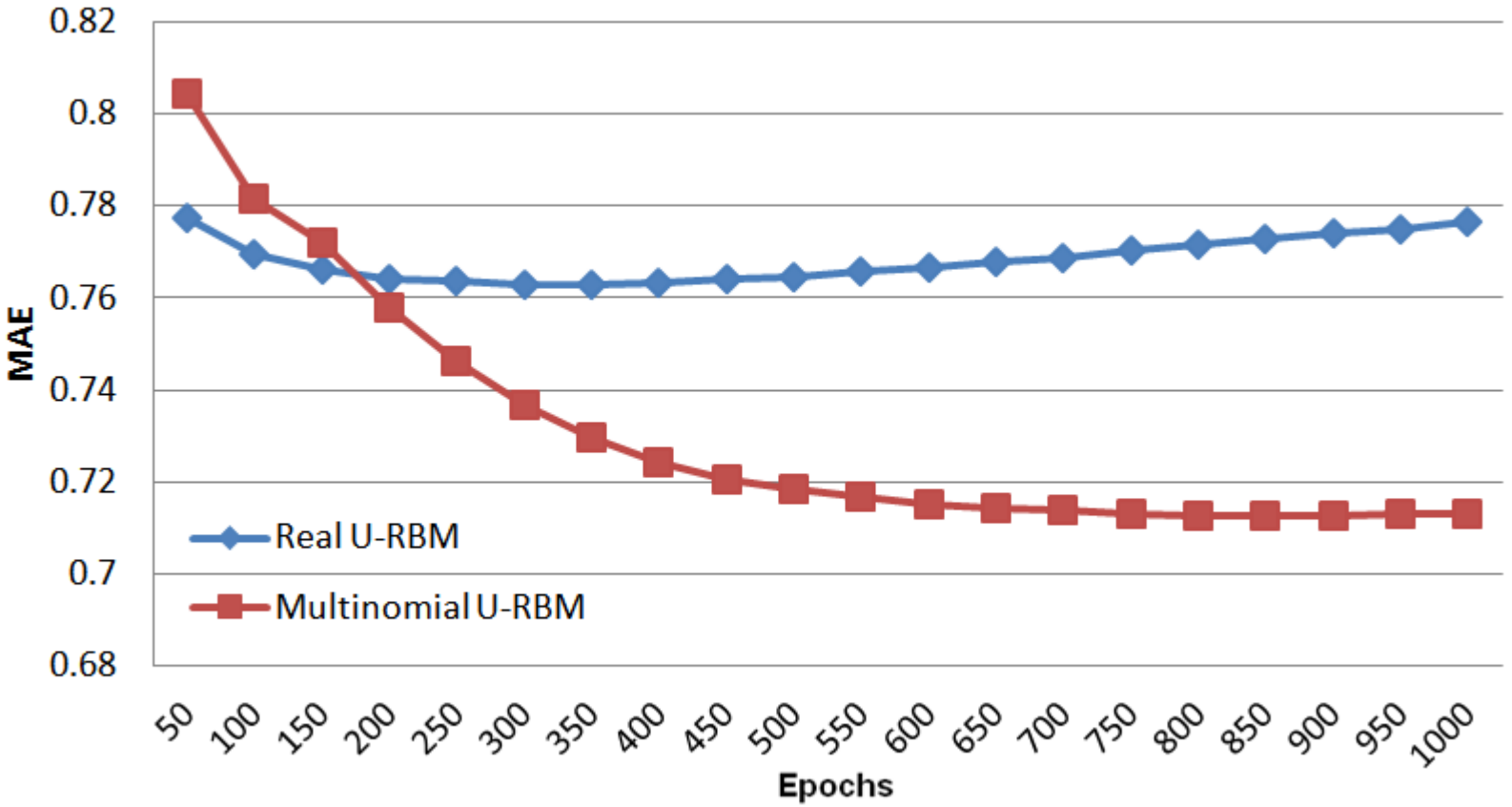
$$MAE = \frac{\sum_i^L |u_i - p_i|}{L}$$

- Cross-validation
  - 5-fold
  - 80%:20% training:testing data splits

Experiments

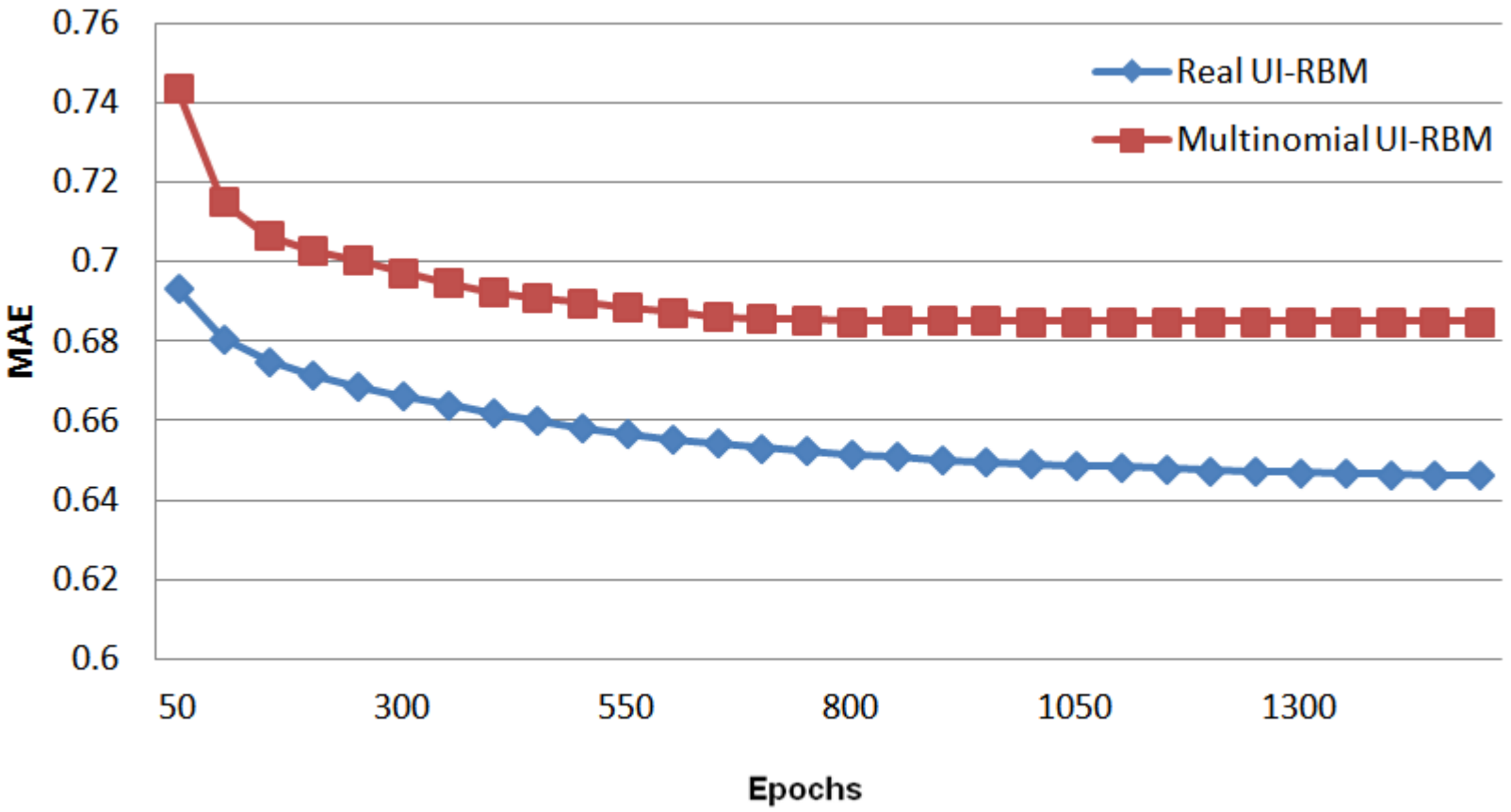
- Evaluated three RBM-based models:
  - User-based RBM (U-RBM)
  - Item-based RBM (I-RBM)
  - Hybrid non-IID RBM model (UI-RBM)
- Tested real-valued vs. multinomial visible units
  - for all above models
- Neighborhood model boosted by I-RBM (I-RBM+INB)

Types of Visible Units: the IID Case



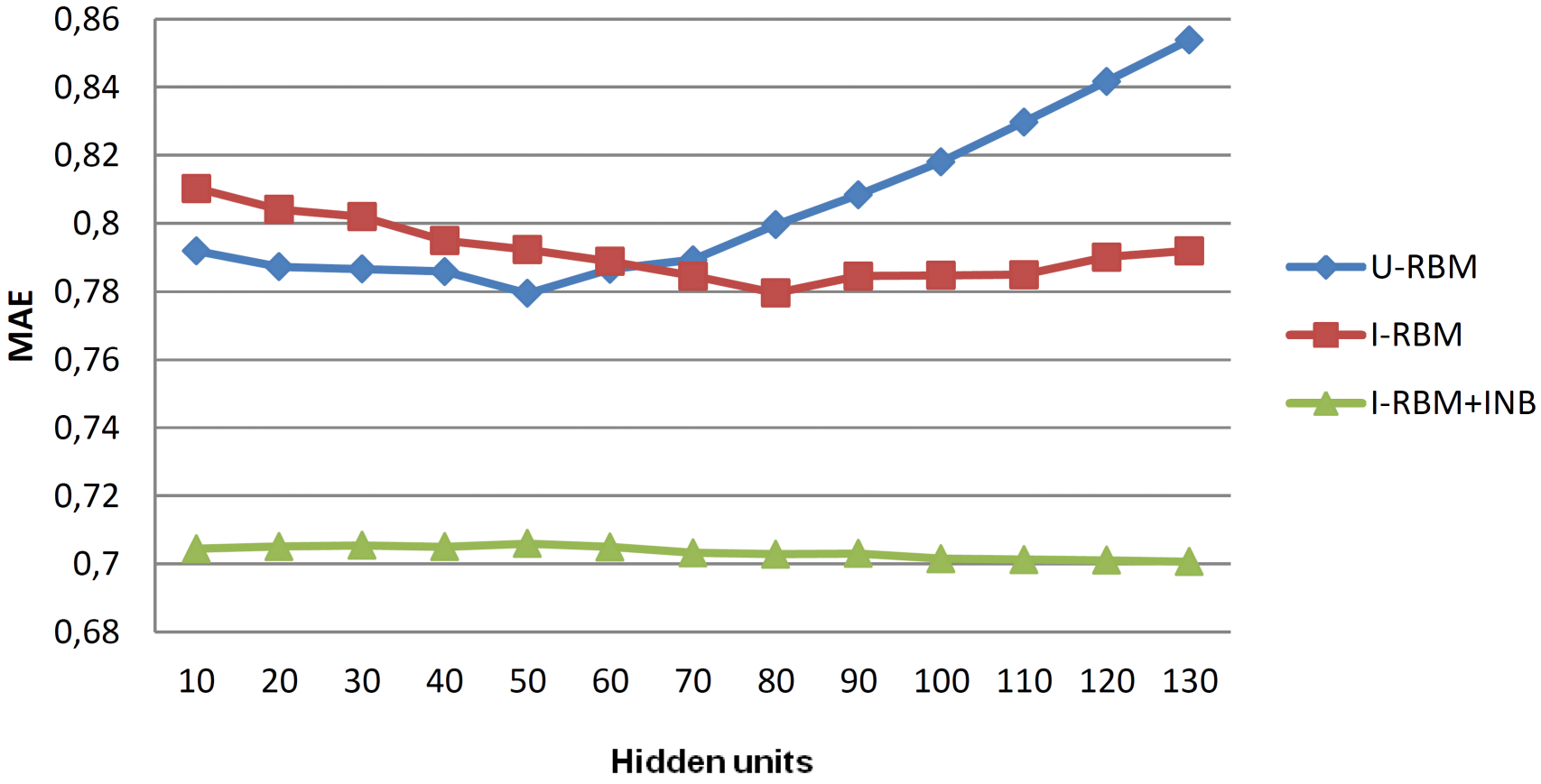
In the IID case, multinomial visible units are better than real-valued.

Types of Visible Units: the non-IID Case



In the non-IID case: real-valued visible units outperform multinomial.

Number of Units: the IID Case



- Item-based RBM model outperforms user-based, but not by much.
- Hybrid item-based RBM + NB model is relatively insensitive to number of units.

Results: MovieLens 100k

CF Model	MAE
SVD PCA (Vozalis et al., 2010)	0.793
H-NLPCA (Vozalis et al., 2010)	0.784
U-RBM	**0.779**
I-RBM	**0.775**
SVD (Sarwar et al., 2002)	0.733
Item-based CF (Sarwar et al., 2001)	0.726
Iter PCA + K-means (Kim & Yum, 2005)	0.712
Iter PCA + RRC (Kim & Yum, 2005)	0.700
I-RBM+INB	**0.699**
UI-RBM	**0.690**
Latent CF (Langseth & Nielsen, 2012)	0.685

Results: MovieLens 1M

CF Model	MAE
Real U-RBM	**0.762**
Real I-RBM	**0.761**
L8 (Taranto et al., 2012)	0.720
Multinomial U-RBM	0.711
Multinomial I-RBM	0.710
Multinomial UI-RBM	**0.685**
GAUSS-UI-BM (Truyen et al., 2009)	0.675
Real I-RBM+INB	**0.669**
ORD-UI-BM (Truyen et al., 2009)	0.657
Real UI-RBM	**0.645**
ORD-UI-BM-CORR (Truyen et al., 2009)	0.640

CONCLUSION AND FUTURE WORK

Conclusion and Future Work

- Conclusion
  - proposed a non-IID RBM framework for CF
  - results rival the best CF algorithms, which are more complex
- Future work
  - add an additional layer to model higher-order correlations
  - add content-based features, e.g., demographic

Thank you!