Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches

Maurizio Ferrari Dacrema
Paolo Cremonesi
Dietmar Jannach

Overview

Indications exist in other application areas of machine learning that the achieved progress of DL algorithms, measured in terms of accuracy improvements over existing models, is not always as strong as expected.

Goal: Evaluation and reproducibility study of recent top-n DL algorithms for RS

How easy is to reproduce the published results?

How competitive DL is against heuristic baselines?

Methodology

Collecting Relevant Papers:

- Conferences: RecSys, WWW, KDD, SIGIR
- Long paper, from 2015 to 2018
- DL applied to traditional Top-n
- Evaluation with accuracy metrics

Collecting Reproducible Papers:

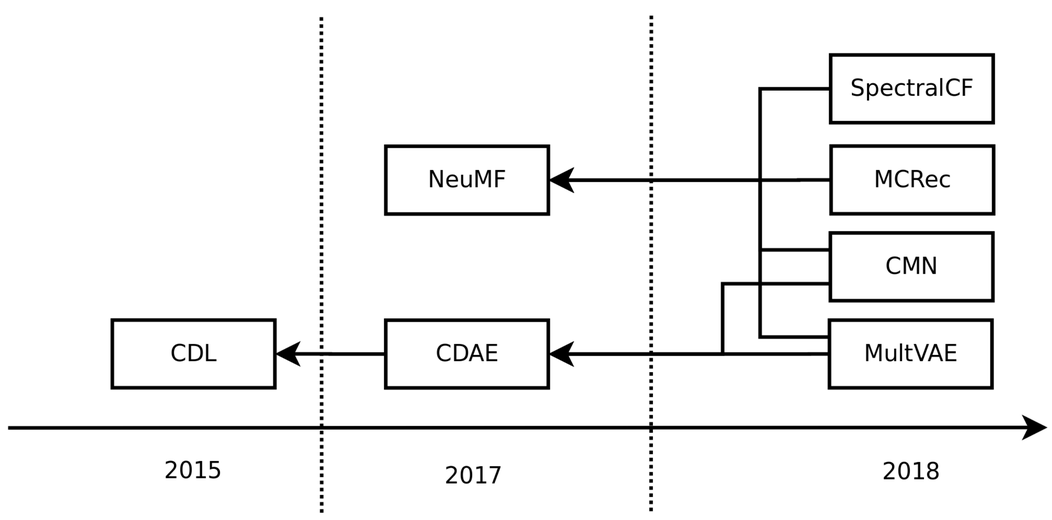
- Source code available and runnable
- Public dataset (original split preferably)
- If not available, ask the authors and wait a month

Reproducibility statistics

Conference	Repr. Ratio
KDD	3/4 (75 %)
WWW	2/4 (50 %)
SIGIR	1/3 (30 %)
RecSys	1/7 (14 %)
Total	7/18 (39 %)

What is the reason for such low reproducibilty?

(1) Lost, (2) NDA, (3) Not runnable, (8) No reply



Extended version!

Extended version on the way!

- 26 relevant papers 12 reproducible
- 253 days of Amazon AWS

Follow our lab on ResearchGate

Experimental evaluation

Same experimental procedure as the original paper: same data, train/test split, metrics, cutoffs

Hyperparameters:

DL	Use original hyperparameters
Baselines	Bayesian search, 40 cases

Baselines:

Top Popular, ItemKNN, UserKNN, P3alpha, RP3beta
ItemKNN CF + CBF
SLIM

Results and discussion

DL algorithms outperforming our baselines

Algorithm	CF + CBF	CF + CBF + NP	CF + CBF + NP + SLIM
MCRec	-	-	-
SpectralCF	-	-	-
CMN	4/12 - 30%	-	-
NeuMF	6/12 - 50%	6/12 - 50%	-
CDL	9/24 - 37%	9/24 - 37%	9/24 - 37%
CVAE	9/24 - 37%	9/24 - 37%	9/24 - 37%
Mult-VAE	12/12 - 100%	12/12 - 100%	10/12 - 83%

Why this discrepancy in our results vs the original ones?

- Weak baselines
- Poor tuning of baseline hyperparameters

Methodological issues:

- Number of epochs selected with test data
- All sorts of evaluation procedures with no motivation

How can we move forward?

Add simple baselines and properly tune them

Improve motivation of experimental design

Improve reproducibility

- Virtualization
- Preprocessing code