UNIVERSITÄT KLAGENFURT

Everyone's a Winner! On Hyperparameter Tuning of Recommendation Models

Faisal Shehzad and Dietmar Jannach

University of Klagenfurt, Austria

dietmar.jannach@aau.at

RecSys 2023, Singapore

Recommender Systems Publication Culture

- Content of most of published recommender systems research:
  1. A **novel** machine learning model to predict user actions or preferences
  2. Empirical **evidence** that it substantially advances the state-of-the-art

- Requirement for 1:
  - Technical novelty
    - preferably exhibiting a certain level of complexity/sophistication
- Requirement for 2:
  - Experiments showing that "Ours" is better than the state-of-the-art
    - preferably with respect to **all accuracy measures** and for **all datasets** that are examined

The problem with the state-of-the-art

- Providing evidence that "Ours" advances the state-of-the-art
  - **Problem:** A state-of-the-art in the sense of a *unique small set of models that is consistently better than other models* does not exist
    - Accuracy results and algorithm rankings depend on datasets, data pre-processing steps, choice of accuracy metrics, choice of baselines, and tuning efforts
  - **Solution:** Provide experimental results involving a selection of very recent and/or commonly used alternatives *to convince reviewers*
    - Researchers have some freedom, e.g., in terms of datasets, baselines, and metrics.

The problem with the state-of-the-art

- Providing evidence that "Ours" advances the state-of-the-art
  - **Problem:** A state-of-the-art in the sense of a *unique small set of models that is consistently better than other model*s does not exist
    - Accuracy results and algorithm rankings depend on datasets, data pre-processing steps, choice of accuracy metrics, choice of baselines, and tuning efforts
  - **Solution:** Provide experimental results involving a selection of very recent and/or commonly used alternatives *to convince reviewers*
    - Researchers have some freedom, e.g., in terms of datasets, baselines, metrics, etc.

Combining these aspects points to a potential problem already

Prerequisites for an insightful comparison

- Let us put the state-of-the-art problem aside
  - And accept that the experimental configuration (baselines, datasets, metrics, etc.) is done in good faith by the researcher
- Let us focus on what makes a *fair and, more importantly, insightful comparison*
- It is easy to agree that a comparison can only be insightful
  - if comparable effort is spent to optimize the performance of each model, including both "Ours" and the baseline models

Hyperparameter tuning

- The performance of machine learning model crucially depends on chosen hyperparameters
  - learning rate, embedding sizes, network structure, loss-related parameters, etc.
- Various techniques exist to find optimal/good hyperparameters for a given dataset

Hyperparameter tuning

- What is reported in many papers
  - Examined hyperparameter ranges for “Ours”
  - Tuning method for “Ours”, e.g., grid search
  - Best hyperparameters for “Ours”, sometimes per dataset
- And for the baselines?
  - Detailed information regarding hyperparameter ranges, process, and best values per dataset for the baselines **almost always missing**
  - Sometimes only vague and/or short statements are provided

Analyzing the current literature

- We scanned recent conference proceedings for papers reporting improved top-n recommendation results
  - KDD, RecSys, SIGIR, TheWebConf, WSDM
- Identified 21 relevant papers
- Analyzed what is documented regarding baseline tuning
  - We of course cannot know what has been actually done

- Disclaimer:
  - We observed similar documentation patterns regarding hyperparameter tuning process in our own previous work

Analyzing the current literature

- One somewhat positive example
  - Reports the searched ranges for “common” hyperparameters (e.g., learning rate, dropout ratio or the coefficient for L2 regularization)
  - Method-specific hyperparameters however taken from original papers
    - Where probably different datasets were used
  - Optimal values however not reported in the end
  - URL provided, but points to empty GitHub repository

Analyzing the current literature

- Another somewhat positive example
  - Reports some of the hyperparameter ranges and some chosen values also for the baselines
    - But does not report on how the parameters were found
  - However, only one set of hyperparameter values reported
    - Even though the evaluation was done on three datasets

- Overall
  - Only two of 21 papers report hyperparameter ranges and chosen values for "Ours" and baselines for each dataset
    - One of them points to an empty GitHub repository

Summing up

- We recall, for an insightful comparison: all hyperparameters of all models must be tuned for all datasets
  - This is usually a huge computational effort nowadays, often taking weeks
  - Still, many papers don't spend a single word on it

- Other common dark patterns
  - "We use the same embedding size for all models for fair comparison."
    - Clearly, the embedding size is a hyperparameter to tune for each model
  - About 50% of papers provide code
    - **None** of the papers provides the code for the baselines
  - Not providing code that produces the results in the paper (only training)

Consequences of improper tuning

- If we assume that some researchers actually did **not** tune the baselines, but only "Ours"
  - Beating the *chosen* baselines might become quite simple
  - The ranking of the baselines may more or less be random, because, e.g.,
    - Some parameters are taken from original papers (using different datasets and evaluation protocols and metrics)
    - Some parameters are just default parameters left in the code by the authors
    - Some parameters are chosen arbitrarily "for fair comparison"
  - In some cases, authors might not have even run the code of the baselines but copied the values from a previous paper

Consequences of improper tuning

- Every model can be declared "winner"
- Experiment (see paper):
  - We benchmarked eight recent neural models (the "state-of-the-art")
  - We tuned the hyperparameters for all of them for three datasets
  - We compared each tuned model against non-tuned ones
    - Non-tuned = using randomly chosen hyperparameters
  - Outcome:
    - Even the worst of the tuned models is better than all non-tuned models
    - *Thus:* Every model can be on the top of the ranking when compared to non-tuned alternatives/baselines

Every model could be “Ours”

Tuned models					
ML-1M		**AMZm**		**Epinions**	
Model	*nDCG@10*	*Model*	*nDCG@10*	*Model*	*nDCG@10*
Mult-DAE	0,300	NeuMF	0,056	Mult-VAE	0,149
Mult-VAE	0,294	Mult-VAE	0,054	Mult-DAE	0.146
GMF	0,280	GMF	0,051	GMF	0,128
NeuMF	0,277	Mult-DAE	0,048	NeuMF	0,118
ONCF	0,225	*MostPop*	*0,013*	ONCF	0,077
MostPop	*0,162*	ConvMF	0,011	*MostPop*	*0,045*
ConvMF	0,160	NGCF	0,008	ConvMF	0,043
NGCF	0,100	ONCF	0,009	NGCF	0,031
Non-tuned models	>		>		>
Mult-DAE	0,071	Mult-DAE	0,003	Mult-DAE	0,015
ONCF	0,037	Mult-VAE	0,002	ONCF	0,005
ConvMF	0,022	ConvMF	0,002	NGCF	0,003
NeuMF	0,021	GMF	0,0007	GMF	0,002
GMF	0,016	NGCF	0,0006	Mult-VAE	0,002
NGCF	0,013	ONCF	0,0004	NeuMF	0,0008
Mult-VAE	0,006	NeuMF	0,0004	ConvMF	0,0008

Table 1. Accuracy results (NDCG@10) for tuned and non-tuned models, sorted by NDCG in descending order.

A little secret ..

- We used the exact same experimental configuration from [1].
- But we omitted the "shallow" EASE^R model by Steck
  - Which would be the overall best performing one
  - It also only has one relevant hyperparameter

- Which brings us back to some previous observation, which may lead to problems
  - Researchers have some freedom, e.g., in terms of datasets, baselines, metrics, etc.

[1] Anelli, V. W., Bellogin, A., Noia, T. D., Jannach, D. and Pomo, C.: "Top-N Recommendation Algorithms: A Quest for the State-of-the-Art". In: 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP 2022). 2022

The general problem is actually well known

- … and limits the progress that we make
  - both in recommender systems research, as well as in information retrieval and other fields

- Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements that don't add up: ad-hoc retrieval results since 1998. In Proceedings of the 18th ACM conference on Information and knowledge management (CIKM '09).
- Makridakis S, Spiliotis E, Assimakopoulos V (2018) Statistical and Machine Learning forecasting methods: Concerns and ways forward. PLoS ONE 13(3): e0194889.
- Wei Yang, Kuang Lu, Peilin Yang, and Jimmy Lin. 2019. Critically Examining the "Neural Hype": Weak Baselines and the Additivity of Effectiveness Gains from Neural Ranking Models. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19).
- Steffen Rendle, Walid Krichene, Li Zhang, and John Anderson. 2020. Neural Collaborative Filtering vs. Matrix Factorization Revisited. In Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20).
- Ferrari Dacrema, M., Boglio, S., Cremonesi, P. and Jannach, D.: *"A Troubling Analysis of Reproducibility and Progress in Recommender Systems Research"*. ACM Transactions on Information Systems, Vol. 39(2). 2021
- …

Jacob Buckman

Please Commit More Blatant Academic Fraud

Posted on May 29, 2021

- A more drastic description of "day-to-day fraud"

"[…] Trying that shiny new algorithm out on a couple dozen seeds, and then only reporting the best few.

Running a big hyperparameter sweep on your proposed approach but using the defaults for the baseline.

Cherry-picking examples where your model looks good, or cherry-picking whole datasets to test on, where you've confirmed your model's advantage. "

D. Picard: **Torch.manual_seed(3407) is all you need**: On the influence of random seeds in deep learning architectures for computer vision, arxiv, 2021

What are the reasons?

- It is easy to agree that zero insights regarding relative algorithm performance can be obtained from such experiments

- Why is it *apparently* still common?
  - Implementing all baselines in a common framework is tedious
  - Tuning hyperparameters can be computationally complex
  - Some ideas probably just turn out not to work at all

But most importantly:
It is the accepted standard even for our top-level publication outlets!

What can we do – ways forward

- Change the publication culture and incentivization system
  - From leaderboard chasing to real-world problems
- Raise awareness, educate and train all involved stakeholders
  - Students, teachers, textbook authors, reviewers, grant evaluators, …
- Improve our methodological standards, various proposals exist
  - Do fair comparisons
  - Publish all code and data to exactly reproduce all reported results
  - Use validated evaluation frameworks

Bauer, C., Fröbe, M., Jannach, D., Kruschwitz, U., Rosso, P., Spina, D. and Tintarev, N.: "**Overcoming Methodological Challenges in Information Retrieval and Recommender Systems through Awareness and Education**". In: Dagstuhl Seminar 23031: Frontiers of Information Access Experimentation for Research and Education. 2023

Thank you!

- Congratulations to everyone who is a winner this year!

- Time for questions, contact: dietmar.jannach@aau.at