Integrating Confidence into Embedding-Based Models for Learn-to-Rank in Recommender Systems

Joel Machado Pires[1], Frederico Araújo Durão[2]

[1]Programa de Pós-Graduação em Ciência da Computação
Universidade Federal da Bahia (UFBA)
Avenida Milton Santos, Salvador, Bahia, 40170-110

[2]Institute of Computing (IC/UFBA)

joelpires@ufba.br, fdurao@ufba.br

Abstract. *Recommender systems produce uncertain predictions due to sparse and noisy user interactions, yet existing confidence estimation methods are mostly limited to matrix factorization models and remain underexplored for graph neural networks and learning-to-rank. This work presents a unified study of confidence estimation in recommender systems by benchmarking prior distribution-based approaches, analyzing their calibration and impact on predictive performance, and proposing probabilistic graph-based models and an uncertainty-aware pairwise ranking formulation based on the Gaussian cumulative distribution function. Experiments on multiple public datasets show that the proposed methods preserve or improve recommendation quality while producing confidence estimates that strongly correlate with prediction errors, providing reliable uncertainty quantification and a general framework for confidence-aware recommendation.*

1. Work Description

Recommender Systems (RS) are systems designed to help users discover relevant content by providing personalized recommendations. Particularly in the collaborative filtering (CF) approach, models are designed to capture latent user–item interaction patterns. Prominent methods in the literature include MF-based [Saia et al. 2016, D'Amico et al. 2022, Mesas and Bellogín 2020, Koren and Sill 2011, Knyazev and Oosterhuis 2023] and GNN-based approaches [Wang et al. 2018b, Wang et al. 2019, Tao et al. 2020, Feng et al. 2019]. However, the stochastic nature of user preferences and the sparsity of interaction data introduce uncertainty into these predictions [Papadopoulos et al. 2001, Koren and Sill 2011, Wang et al. 2018a]. This uncertainty may lead to unreliable recommendations, especially in high-risk domains or for decision-critical applications. The presence of unreliable recommendations can be exacerbated when the model fails to provide any indication of prediction reliability. To address this, confidence estimation has been proposed as a way to quantify the reliability of model predictions, providing an inverse measure of prediction uncertainty [Mesas and Bellogín 2020, Koren and Sill 2011, Knyazev and Oosterhuis 2023].

In statistics, confidence intervals are used to infer that a value will fall within a specified range with a given confidence level [Watt et al. 2007]. In RS, some works presented a new viewpoint: confidence as an estimation of the model's certainty level for each prediction [Koren and Sill 2011, Wang et al. 2018a, Knyazev and Oosterhuis 2023]. In particular, when a model predicts whether a user will like or dislike a specific item, we want to estimate the confidence in this prediction.

The literature often addresses confidence by imposing a distribution on the model's outcomes, such as a Gaussian, Gamma, or Beta distribution. Previous work has explored these confidence modeling strategies primarily in MF-based models, while GNN-based models remain underexplored in this context. Additionally, despite growing attention to confidence estimation in RS, existing methods have not been holistically benchmarked, limiting our understanding of their strengths, weaknesses, and applicability. In particular, the literature lacks a clear understanding of how previous confidence modeling techniques can be adapted to GNN-based architectures. For instance, OrdRec by Koren and Sill [Koren and Sill 2011], and CPMF and CBPMF by Wang et al. [Wang et al. 2018a], model uncertainty using predefined distributions such as Beta and Gaussian. LBD, proposed by Knyazev and Oosterhuis [Knyazev and Oosterhuis 2023], follows a similar approach. Other works—such as those by Mesas and Bellonín [Mesas and Bellogín 2020], Gohari et al. [Gohari et al. 2018], and Himabindu et al. [Himabindu et al. 2018]—employ heuristic methods to estimate confidence and use it for output calibration.

However, none of these approaches provides confidence estimation tailored for GNN-based models, which remain underrepresented in this area. Moreover, methods such as those proposed in [Koren and Sill 2011, Wang et al. 2018a] often degrade accuracy, while the approach in [Knyazev and Oosterhuis 2023] suffers from convergence issues. Finally, no prior work has systematically analyzed the quality of confidence estimates, for instance, by evaluating their relationship with prediction error across different model types. In particular, does not demonstrate whether these estimates are reliable.

This work investigates prior-distribution-based methods for confidence estimation to understand why confidence integration can degrade model performance and lead to poor calibration. Additionally, we propose a mitigation strategy for the identified issues and address the lack of confidence-aware graph neural network models by introducing probabilistic recommender GAT (PRGAT) and probabilistic recommender LightGCN (PRLIGHTGCN). We make the following key contributions:

- Benchmarking of four distribution-based confidence-aware models - OrdRec, CPMF, CBPMF, and LBD — on three public datasets (Amazon Movies and TVs, JesterJoke, and MovieLens) using standard metrics for rating prediction (RMSE) and ranking (NDCG, MAP).
- Proposing a model architecture with confidence integration.
- Analyzing the relationship between model confidence and prediction error across datasets
- Providing insights into confidence calibration and the limitations of existing confidence-aware approaches.

The comprehensive experimental study conducted in this work reveals that prior distribution-based confidence models exhibit pronounced sensitivity to dataset characteristics, leading to inconsistent performance in both prediction accuracy and ranking quality. Beyond performance reporting, our analysis identifies structural factors underlying these behaviors, including optimization instability and distributional constraints that hinder reliable calibration. In contrast, the proposed probabilistic graph-based framework demonstrates consistent robustness across all evaluated datasets, outperforming existing distribution-based approaches. Notably, the results also indicate that naively integrating distribution-based confidence formulations into embedding models can adversely affect predictive accuracy.

However, this aforementioned proposed solution and prior works focus on rating prediction. Attempts to extend confidence estimation to pair-wise or list-wise learn-to-rank are scarce. Post-processing solutions [Zhang et al. 2016] and heuristic-based approaches [Mesas and Bellogín 2020] introduce either high computational costs or model-specific assumptions. Wang et al. [Wang and Kadıoğlu 2023] advance the field with deep learning–based uncertainty modeling that distinguishes epistemic and aleatoric uncertainty. They also propose a solution based on distribution, but its inference overhead and restriction to neural architectures limit broader applicability.

To the best of our knowledge, learning confidence estimates for the relevance scores produced by pairwise or listwise learn-to-rank models are very scarce. To address this gap, we introduce uncertainty-aware learning-to-rank, a method that integrates confidence estimation directly into the pairwise ranking framework. Specifically, we reformulate the BPR loss by replacing its traditional logistic sigmoid with the cumulative CDF of the Normal distribution. This formulation allows the model to jointly learn the mean and variance of the predicted relevance scores, with the variance representing the model's uncertainty for each prediction. As a result, the method provides a calibrated quantification of how certain it is that an observed item should be ranked higher than a non-observed one. Importantly, this design maintains the computational efficiency, simplicity, and general applicability of embedding-based ranking models, making it compatible with various architectures such as MF and GAT.

This work performs an experimental evaluation with datasets from distinct contexts, such as Amazon Beauty, Jester Joke, Movie Lens 1M, and Rotten Tomatoes, and different metrics for assessing ranking quality, such as NDCG, MAP, Recall, and MRR. Additionally, we analyze the model's confidence output, including its distribution and its relationship with the model's error. Experimental results demonstrated that the proposed method is equivalent to, and in some cases, even outperforms the baselines. We used standard MF and a deep GAT model without confidence integration as baselines for assessing prediction performance. Therefore, this work also brings the following contributions:

- A unified uncertainty-aware listwise formulation using a Gaussian CDF.
- Confidence-aware variants that preserve or improve ranking performance across four datasets and two model families.
- Empirical evidence that uncertainty reflects predictive reliability with strong negative Pearson correlations between confidence and BPR error.
- Demonstrate uncertainty magnitudes and correlations reflect the model's reliability and dataset properties such as sparsity, interaction independence, and domain complexity.
- A cross-domain evaluation demonstrating general applicability for listwise learn-to-rank.

2. General Information

- Categoria: Mestrado
- Nome do(a) Autor(a): Joel Machado Pires
- Título da tese ou da dissertação: Integrating Confidence into Embedding-Based Models for Learn-to-Rank in Recommender Systems
- Programa de Pós-Graduação / Universidade: Programa de Pós-Graduação em Ciência da Computação / Universidade Federal da Bahia (UFBA)
- Nome do(a) orientador(a): Frederico Araújo Durão
- Data de defesa e aprovação: 20/01/2026
- Banca examinadora: André Levi Zanon (Insight Research Ireland Centre for Data Analytics), Yuri de Almeida Malheiros Barbosa (UFPB)

3. Highlights

- Confidence-aware GATs improve recommender system reliability
- Benchmark of distribution-based confidence integration methods
- Proposed models outperform prior confidence-aware baselines
- Cubic confidence-error relationship found in learn-to-rank models

References

D'Amico, E., Gabbolini, G., Bernardis, C., and Cremonesi, P. (2022). Analyzing and improving stability of matrix factorization for recommender systems. *Journal of Intelligent Information Systems*, 58(2):255–285.

Feng, C., Liu, Z., Lin, S., and Quek, T. Q. (2019). Attention-based graph convolutional network for recommendation system. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7560–7564.

Gohari, F. S., Aliee, F. S., and Haghighi, H. (2018). A new confidence-based recommendation approach: Combining trust and certainty. *Information Sciences*, 422:21–50.

Himabindu, T. V., Padmanabhan, V., and Pujari, A. K. (2018). Conformal matrix factorization based recommender system. *Information Sciences*, 467:685–707.

Knyazev, N. and Oosterhuis, H. (2023). A lightweight method for modeling confidence in recommendations with learned beta distributions. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 306–317.

Koren, Y. and Sill, J. (2011). Ordrec: an ordinal model for predicting personalized item rating distributions. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, page 117–124, New York, NY, USA. Association for Computing Machinery.

Mesas, R. M. and Bellogín, A. (2020). Exploiting recommendation confidence in decision-aware recommender systems. *Journal of Intelligent Information Systems*, 54(1):45–78.

Papadopoulos, G., Edwards, P., and Murray, A. (2001). Confidence estimation methods for neural networks: a practical comparison. *IEEE Transactions on Neural Networks*, 12(6):1278–1287.

Saia, R., Boratto, L., and Carta, S. (2016). A semantic approach to remove incoherent items from a user profile and improve the accuracy of a recommender system. *Journal of Intelligent Information Systems*, 47(1):111–134.

Tao, Z., Wei, Y., Wang, X., He, X., Huang, X., and Chua, T.-S. (2020). Mgat: Multimodal graph attention network for recommendation. *Information Processing Management*, 57(5):102277.

Wang, C., Liu, Q., Wu, R., Chen, E., Liu, C., Huang, X., and Huang, Z. (2018a). Confidence-aware matrix factorization for recommender systems. In *Proceedings of the AAAI Conference on artificial intelligence*, volume 32.

Wang, H., Zhang, F., Wang, J., Zhao, M., Li, W., Xie, X., and Guo, M. (2018b). Ripplenet: Propagating user preferences on the knowledge graph for recommender systems. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, page 417–426, New York, NY, USA. Association for Computing Machinery.

Wang, X., He, X., Cao, Y., Liu, M., and Chua, T.-S. (2019). Kgat: Knowledge graph attention network for recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 950–958, New York, NY, USA. Association for Computing Machinery.

Wang, X. and Kadıoğlu, S. (2023). Modeling uncertainty to improve personalized recommendations via bayesian deep learning. *International Journal of Data Science and Analytics*, 16(2):191–201.

Watt, T. A., McCleery, R. H., and Hart, T. (2007). *Introduction to statistics for biology*. CRC Press.

Zhang, M., Guo, X., and Chen, G. (2016). Prediction uncertainty in collaborative filtering: Enhancing personalized online product ranking. *Decision Support Systems*, 83:10–21.