Global-Prototype Disentanglement Network: Bridging Distributional Gaps in Multi-Scenario Recommendation

Abstract

The heterogeneity of user engagement across distinct interface modalities—ranging from main feeds to real-time streaming contexts—poses a significant challenge in recommendation systems. While these environments share underlying user interests, the statistical divergence in data distributions often leads to suboptimal performance in standard Multi-Scenario Recommendation (MSR) frameworks. Current architectures, which typically rely on rigid parameter sharing combined with branch networks, struggle to adequately decouple scenario-specific nuances from common patterns. This failure to account for distributional shifts frequently results in feature interference and negative transfer. To address these limitations, we introduce the Global-Prototype Disentanglement Network (GPD-Net). This framework establishes a novel paradigm based on distribution-adaptive latent spaces. Structurally, GPD-Net leverages a Mixture-of-Experts (MoE) backbone to extract generalized signals, complemented by standalone pathways for distinct scenario traits. A key innovation lies in projecting these signals into a prototype-guided manifold, yielding aligned representations that balance global coherence with local specificity. Furthermore, we incorporate an Unbalanced Optimal Transport (UOT) mechanism to dynamically align feature vectors with global anchors, treating the calculated transport plans as auxiliary supervision signals. To ensure robustness and prevent feature collapse, an orthogonality regularization is imposed on the prototype matrix. Validation via rigorous offline benchmarks on four datasets and live production A/B testing confirms that GPD-Net significantly outperforms state-of-the-art baselines by effectively mitigating distribution entanglement.

1. Introduction

Modern recommendation systems increasingly operate across multiple user interfaces such as short-video feeds, topic-specific channels, and live streaming environments. Each scenario demonstrates unique interaction dynamics and sampling biases, creating significant challenges for shared model architectures. The inherent distributional divergence among scenarios leads to entangled feature spaces, where scenario-specific knowledge interferes with generalizable patterns. Existing Multi-Scenario Recommendation (MSR) models often employ a combination of shared encoders and specialized subnetworks. However, rigid parameter sharing introduces strong coupling between unrelated contexts, while fully decoupled designs fail to exploit shared user intent. Therefore, a principled mechanism for disentangling shared and scenario-dependent semantics is essential.

2. Methodology

The proposed Global-Prototype Disentanglement Network (GPD-Net) introduces a three-stage modeling paradigm that jointly considers representation disentanglement and cross-scenario alignment. At its core, GPD-Net integrates a Mixture-of-Experts (MoE) backbone with prototype-guided latent alignment, optimized under the Unbalanced Optimal Transport (UOT) framework. This architecture enables flexible information flow between global and scenario-specific representations while preserving discriminative structure within each domain.

2.1 Mixture-of-Experts Backbone

The MoE module serves as a generalized encoder shared across scenarios. Given an input feature sequence X, a gating network assigns adaptive weights to K experts based on contextual embeddings, producing a global representation. This mechanism effectively captures cross-scenario commonality while enabling adaptive specialization. Scenario-specific branches complement the shared encoder to preserve unique contextual semantics, forming the dual-channel structure of GPD-Net.

2.2 Prototype-guided Manifold Projection

To achieve balanced representation learning, GPD-Net introduces a prototype-guided manifold where latent features are softly projected onto a global prototype matrix P = [p1, ..., pM]. Each prototype captures a semantic anchor corresponding to a shared behavioral pattern. The projection process enforces global coherence while maintaining local adaptability across scenarios. Orthogonality regularization on P prevents redundancy and ensures a well-conditioned prototype space, stabilizing training under highly heterogeneous distributions.

2.3 Unbalanced Optimal Transport Alignment

Unlike balanced transport that assumes equal mass between distributions, the Unbalanced Optimal Transport (UOT) module allows partial correspondence between feature and prototype distributions. Formally, given feature embeddings Z and prototypes P, UOT solves a relaxed assignment problem minimizing transport cost plus marginal regularization. The resulting transport matrix serves as auxiliary supervision, dynamically reweighting sample-to-prototype affinities. This process enables stable alignment under distributional imbalance, mitigating mode collapse and feature interference across scenarios.

3. Experiments

We evaluate GPD-Net across four benchmark datasets covering diverse user interaction modalities, including video feeds, live streaming, and e-commerce browsing. In offline experiments, GPD-Net consistently achieves superior AUC and NDCG compared to leading MSR baselines such as MMoE, PLE, and STAR. Furthermore, real-world A/B testing on a large-scale content platform demonstrates statistically significant gains in watch-time and click-through metrics. Ablation studies verify the importance of prototype orthogonality and UOT alignment in reducing feature entanglement.

4. Conclusion

In summary, the Global-Prototype Disentanglement Network (GPD-Net) offers a principled approach to addressing distributional heterogeneity in multi-scenario recommendation. By coupling mixture-of-experts generalization with prototype-guided manifold alignment and unbalanced optimal transport supervision, GPD-Net achieves robust, disentangled, and interpretable representations across scenarios. Its superior performance across both offline and online evaluations validates its effectiveness as a general framework for distribution-adaptive recommendation.