DeepCF: A Unified Framework of Representation Learning and Matching Function Learning in Recommender System

1.Main Contribution

The Authors designed a framework (DeepCF) which combines the representation learning-based CF methods and matching function learning-based CF methods to endow the model with a great flexibility of learning the matching function while maintaining the ability to learning low-rank relations.

2.Approach

2.1 Initial User-Item interaction matrix $\mathbf{Y}$

Suppose there are M users and N items, the shape of $\mathbf{Y}$ should be $\mathbb{R}^{M\times N}$. Since $\mathbf{Y}$ is made of implicit data, every element y_{ui} in $\mathbf{Y}$ might be inaccurate, especially when it is 0, which might mean an unobserved interaction. The authors try to uniformly sample negative instances from the unobserved interactions to tackle the problem.

y_{ui} could be ether 0 or 1, which is not enough to rank and recommend items. To tackle this, the authors assume y_{ui} obeys a Bernoulli distribution: $P(y_{ui}=k\mid p_{ui})=p_{ui}^{k}(1-p_{ui})^{1-k}$. By modeling p_{ui} instead of y_{ui} , the binary classification interaction prediction problem is converted to a rating prediction problem.

2.2 Representation Learning Part (CFNet-rl)

The initial representations of user u and item i are denoted as $\mathbf{v}_u^U$ and $\mathbf{v}_i^I$. Through MLP functions, we can get the latent representations $\mathbf{p}_u=f(\mathbf{v}_u^U)$ for user u and $\mathbf{q}_u=g(\mathbf{v}_i^I)$ for item i. Then a matching function is used to get the final $\hat{y}_{ui}$: $\hat{y}_{ui}=\sigma(\mathbf{W}_{out}^T(\mathbf{p}_u\odot\mathbf{q}_i))$.

This part still focuses on catching low-rank relations between users and items but is more expressive for that the matching function could be non-linear.

2.3 Matching Function Learning Part (CFNet-ml)

For $\mathbf{v}_u^U$ and $\mathbf{v}_i^I$, two linear embedding layers are applied to learn the latent representations for users and items: $\mathbf{p}_u=P^T\mathbf{y}_{u*},\mathbf{q}_u=Q^T\mathbf{y}_{*i}$. Then, MLP is used to learn the matching function and finally we can get the matching score $\hat{y}_{ui}$.

2.4 Fusion and Learning between the two parts

The purpose is to fuse CFNet-rl and CFNet-ml to get a better representation. The predictive vectors of the two layers are defined as the last layer of the matching function, specifically, it is

$\mathbf{p}_u \odot \mathbf{q}_i$ for CFNet-rl, which is denoted as a_Y^{rl} , and the last layer of MLP in the matching function for CFNet-ml, which is denoted as a_Y^{ml} . And a fully connected layer is applied here to get the final representation: $\hat{y}_{ui} = \sigma(W_{out}^T \begin{bmatrix} a_Y^{rl} \\ a_Y^{ml} \end{bmatrix})$.

2.5 Train

The binary cross-entropy loss function is applied here as the loss function:

$$l = -\sum_{(u,i)\in y^+ \cup y^-} y_{ui} \log \hat{y}_{ui} + (1 - y_{ui}) \log(1 - \hat{y}_{ui}) .$$

Where y^+ denotes all the observed interactions in $\mathbf{Y}$ and y^- denotes all the sampled unobserved interactions.

3.Experiments

The authors use HR (Hit Ratio) and NDCG (Normalized Discounted Cumulative Gain) as the measures. Experiments conducted on 4 public datasets show that DeepCF can achieve SOTA results compared to ItemPop, eALS, DMF, NeuMF.

For DeepCF itself, pre-training on CFNet-rl and CFNet-ml is necessary. Besides that, the optimal sampling ratio is around 3 to 7, and the more dimensions of a_Y^{rl} and a_Y^{ml} are, the better the performance of representation is.

4.Future Work

1) Combine auxiliary data to improve the initial representation performance.

2) Try pair-wise loss instead of the point-wise loss in the future.

3) The DeepCF framework can also be applied to some other data mining related areas.

HM-Modularity: A Harmonic Motif Modularity Approach for Multi-layer Network Community Detection

1.Main Contribution

The main contribution of the paper is that the authors propose a higher-order structural approach (harmonic motif) for multi-layer network community detection.

2.Challenges

The two challenges are 1) the most representative higher-order structure (motif) may vary from different layers; 2) the same node subset may exhibit different higher-order connectivity patterns in different layers.

3.Approach

Generally speaking, the authors use Harmonic Motif to tackle the first challenge and use a threshold to control the slightly different higher-order structural information between layers to tackle the second challenge.

3.1 Harmonic Motif

Firstly, Z-score is defined as $Z = \frac{N_{real} - mean(N_{rand})}{std(N_{rand})}$ to get the statistically significance of a motif in a single layer. So, the average Z-score over all the v layers is $\bar{Z} = \frac{1}{v}\sum_{p=1}^{v} Z^{p}$. Since we only consider 3-node and 4-node motif in this paper, we can pick up 2 motifs which have the largest $\bar{Z} - score$ for 3-node and 4-node motif. And they are denoted as $\bar{M}_3$ and $\bar{M}_4$. $\bar{M}_3$ and $\bar{M}_4$ can tackle the first challenge for that they include the structural information over all layers.

3.2 Layer-Integrated Harmonic Motif Proximity Matrix and Layer-Specific Harmonic Motif Proximity Matrix

3.2.1 Harmonic Value

The Harmonic Value over layer p is defined as:

$$w_{\bar{I}}^{p} = \frac{\sum_{u,v \in \mathcal{V}_{\bar{I}}} True((u,v) \in \varepsilon_{\bar{I}} \cap \varepsilon^{p} \Leftrightarrow (f(u), f(v)) \in \varepsilon_{\bar{M}})}{|\varepsilon_{\bar{M}}|}$$

And a threshold value λ is used so that $w_{\bar{I}}^{p}$ is not too small. It allows a slight difference over instances of a motif thus it can tackle the second challenge. We can easily get the Harmonic Value over all layers: $w_{\bar{I}} = \sum_{p=1}^{v} w_{\bar{I}}^{p}$.

3.2.2 Layer-Integrated Harmonic Motif Proximity Matrix

Layer-integrated harmonic motif instance is a motif which satisfies the harmonic value in all layers. Every element $\bar{A}_{ij}$ in the matrix $\bar{A}$ is defined as the weighted sum of the harmonic values of all the layer-integrated 3-node and 4-node harmonic motif instances containing nodes v_i and v_j. It contains the higher-order structural information over all layers and makes a balance between the 3-node and 4-node motif. Mathematically, $\bar{A}_{ij} = \frac{1}{\bar{Z}_3 + \bar{Z}_4}(\bar{Z}_3 \sum_{\bar{I}_3 \in \{\bar{I}_3(i,j)\}} w_{\bar{I}_3} + \bar{Z}_4 \sum_{\bar{I}_4 \in \{\bar{I}_4(i,j)\}} w_{\bar{I}_4})$.

3.2.3 Layer-Specific Harmonic Motif Proximity Matrix

Layer-specific harmonic motif instance is a motif that makes the harmonic value in this layer equal to 1. Every element $\bar{A}_{ij}^{p}$ in the matrix $\bar{A}_p$ is defined as the weighted sum of the number of layer-specific 3-node and 4-node harmonic motif instances containing both nodes v_i and v_j. It can contain as many layer-specific harmonic motif instances as possible. Mathematically, $A_{ij}^{p} = \frac{1}{\bar{Z}_3 + \bar{Z}_4}(\bar{Z}_3 |\bar{I}_3^{p}(i,j)| + \bar{Z}_4 |\bar{I}_4^{p}(i,j)|)$.

3.3 Higher-Order Structural Network

The higher-order structural network has (v+1) layers. The first v layers are the coupling of the layer-integrated harmonic motif proximity matrix and the layer-specific harmonic motif proximity matrix of each layer, which are also called the auxiliary layers. The (v+1)-th layer is the layer-integrated harmonic motif proximity matrix, which is also called the primary layer.

Two kind of tensors, intra-layer tensor and inter-layer tensor are defined to characterize the intra-layer connection and the inter-layer coupling of the (v+1)-layer higher-order structural network.

3.3.1 Intra-Layer Harmonic Motif Proximity Tensor

It is defined by simply stacking the layer-specific harmonic motif proximity matrices and the layer-integrated harmonic motif proximity matrix since both parts are important in preserving layer-integrated and layer-specific harmonic motifs. $W^{p} = \begin{cases} A^{p} & \forall p = 1, \ldots v \\ \bar{A} & p = v+1 \end{cases}$.

3.3.2 Inter-Layer Harmonic Motif Coupling Tensor

It is defined as the weighted sum of the p-th layer harmonic values of all the layer-integrated

3-node and 4-node harmonic motif instances containing node v_i. The reason is that it can ensure the consistency of the corresponding higher-order connectivity pattern across the p-th layer and the q-th layer.

$$C_i^{pq} = \begin{cases} \frac{1}{\bar{Z}_3 + \bar{Z}_4}(\bar{Z}_3 \sum_{\bar{I}_3 \in \{\bar{I}_3(i)\}} w_{\bar{I}_3}^{p} + \bar{Z}_4 \sum_{\bar{I}_4 \in \{\bar{I}_4(i)\}} w_{\bar{I}_4}^{p}) & p = 1, \ldots, v, q = v+1 \\ 0 & Otherwise \end{cases}$$

3.4 The Harmonic Motif Modularity Algorithm

It is used to generate the higher-order community structure. The objective function is:

$$Q(\{g_i^p\}) = \frac{1}{2\mu} \sum_{p,q} \sum_{i,j} [(W_{ij}^p - \frac{k_i^p k_j^p}{2k^p}) \delta(p,q) + C_i^{pq} \delta(i,j)] \delta(g_i^p, g_j^q)$$

Where g_i^p is the community label of node i in the p-th layer; $\delta(x,y) = \begin{cases} 1 & x = y \\ 0 & otherwise \end{cases}$;

$k_i^p = \sum_{j=1}^{n} W_{ij}^p$; $k^p = \sum_{i=1}^{n} k_i^p$; $\mu = \frac{1}{2} \sum_{p=1}^{v+1} k^p$.

The objective function can be separated into 4 parts:

$$p = q, j \neq i : Q_1(\{g_i^p\}) = \frac{1}{2\mu} \sum_{p=1}^{v+1} \sum_{i,j} (W_{ij}^p - \frac{k_i^p k_j^p}{2k^p}) \delta(g_i^p, g_j^p)$$

It means the sum of the intra-layer harmonic motif proximities should be as large as possible.

$$p \neq q, j = i : Q_2(\{g_i^p\}) = \frac{1}{2\mu} \sum_{p,q} \sum_{i=1}^{n} C_i^{pq} \delta(g_i^p, g_i^q)$$

It means the community label of each node should be as consistent as possible across the primary layer and each auxiliary layer.

$p = q, j = i : Q_4 = const$.

The constant is independent of community labels.

$p \neq q, j \neq i : Q_3 = 0$

The generalized Louvain approach is supplied to optimize the problem. Finally, the community labels are in the primary layer $g_i^{v+1}, \forall v = 1, \ldots, n$.

4.Experiments

The experiments are conducted on 11 real-world multi-layer networks. When using the NMI (normalized mutual information) and CR (classification rate) as the measures, we can see that HM-Modularity achieves best performance in both measures on all the datasets.

5.Future Work

1)Try to design a higher-order fast motif search method.

2)Obtain the theoretical proof of constructing the harmonic motif proximity matrices.

Explainable Recommendation through Attentive Multi-View Learning

1.Main Contribution

The main contribution of the paper is that it builds a network based on an explainable deep hierarchy to improve accuracy and explainability simultaneously of personalized recommendation.

2.Challenges

The two main challenges of the model are:1) How to model multi-level explicit features from noisy and sparse data; 2) How to generate explanations that are easy for users to understand. The authors tackle those challenges by 1) develop a Deep Explicit Attentive Multi-view Learning model (DEAML), and apply an attentive multi-view learning framework on it; 2) Using dynamic programming to formulate personalized explanation generation as a constrained tree node selection problem.

3.Approach

Input: an explicit feature hierarchy Υ , user set U , item set V .

Υ is a tree indicating “IsA” relations where each node F_l is a feature from a concept graph. U is represented as $(i,\mathbf{x}_i)$, where i is the user id and x_{ij} indicates how much user i is interested in feature j . V is represented as $(i,\mathbf{y}_i)$, where i is the item id and y_{ij} indicates how well item i performs in feature j .

Output: predicted rating $\tilde{r}_{ij}$, feature level explanation $E=\{F_{l_1},\cdots,F_{l_T}\}$.

$\tilde{r}_{ij}$ is the predicted rating between user i and item j . E is a set of nodes in Υ which can give feature-level explanation of the rating.

Step1: Predection with Attentive Multi-View Learning

Step1.1 Hierarchical Propagation

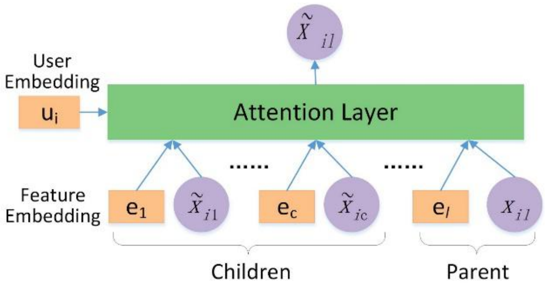
The purpose of Hierarchical Propagation is to build an accurate user-feature interest hierarchy (item-feature quality hierarchy) using $\mathbf{x}_i$ ($\mathbf{y}_i$) and Υ . Attention mechanism is applied here:

$$\alpha_{lc}^{*} = \mathbf{h}_1^T \operatorname{Re}LU(\mathbf{W}_l e_l + \mathbf{W}_c e_c + \mathbf{W}_u u_i + \mathbf{b}_1) + b_2$$

$$\alpha_{lc} = \frac{\exp(\alpha_{lc}^{*})}{\sum_{F_c \in children(F_l) \cup \{F_l\}} \exp(\alpha_{lc'}^{*})}$$

$$\tilde{x}_{il} = \alpha_{ll} x_{il} + \sum_{F_c \in children(F_l)} \alpha_{lc} \tilde{x}_{ic}$$

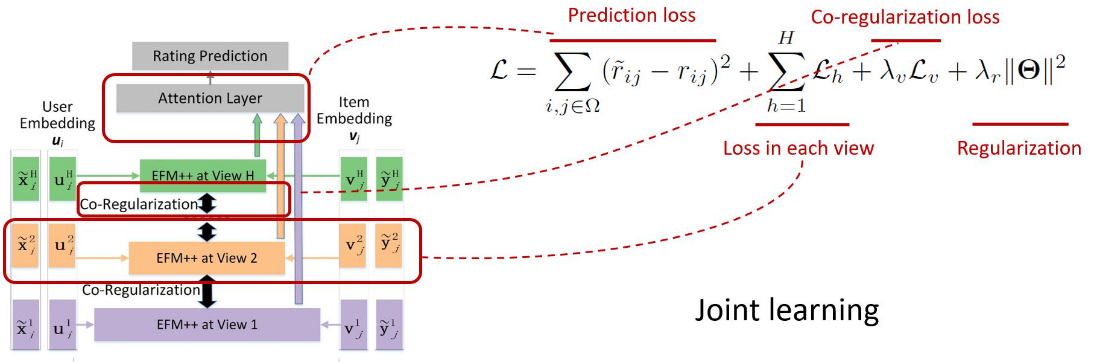
F_l is the current feature node in Υ and F_c is its child. e_i is the feature embedding of node i which is pretrained using other method. u_i^h is the user embedding of layer h .



Finally we can get $\tilde{\mathbf{x}}_i$ which is an accurate user-feature interest between user i and all the features in Υ. We ca use the same hierarchical propagation to get $\tilde{\mathbf{y}}_i$ which is an accurate item-feature quality between item i and all the features in Υ.

Step1.2 Attentive Multi-view Learning

The purpose of Attentive Multi-view Learning is to predict how much user i likes item j based on $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{y}}_j$. The authors use a multi-view learning framework and treat features at the same level as a view. So the optimization can be divided into three parts: loss in a single view, co-regularization loss between two adjacent views, and prediction loss.



Step1.2.1 loss in a single view

The authors use extended EFM by adding rating bias, user bias and item bias to calculate loss in a single view for that EFM model can enrich user and item representations. The user embedding for view h is $\mathbf{u}_i^h = \mathbf{p}_i^h \oplus \mathbf{c}_i^h$, and $\mathbf{p}_i^h$ are the explicit features. And they are used to fit $\tilde{\mathbf{x}}_i^h$ with $\mathbf{Z}_h \mathbf{p}_i^h$. The item embedding for view h is $\mathbf{v}_j^h = \mathbf{q}_j^h \oplus \mathbf{d}_j^h$, and $\mathbf{q}_j^h$ are the explicit features. And they are used to fit $\tilde{\mathbf{y}}_j^h$ with $\mathbf{Z}_h \mathbf{q}_j^h$. According to extended EFM, the prediction that rating of user i on item j is: $\tilde{r}_{ij}^h = u_i^{h^T} v_j^h + o_i + o_j + \mu$. And the loss in view h is:

$$L_h = \lambda_a \sum_{i,j\in\Omega} (\tilde{r}_{ij}^h - r_{ij}^h)^2 + \lambda_x \sum_i \| \tilde{\mathbf{x}}_i^h - \mathbf{Z}_h \mathbf{p}_i^h \|^2 + \lambda_y \sum_j \| \tilde{\mathbf{y}}_j^h - \mathbf{Z}_h \mathbf{q}_j^h \|^2 .$$

Step1.2.2 co-regularization loss

The main idea of co-regularization loss is to enforce the agreement between two adjacent views.

So the loss is: $L_v = \sum_{i,j\in\Omega} \sum_{h=1}^{H-1} (\tilde{r}_{ij}^h - \tilde{r}_{ij}^{h+1})^2$, where Ω is the set of training instances.

Step1.2.3 prediction loss

The main purpose of prediction loss is to combine multiple views using attention mechanism:

$$w_h^* = \mathbf{h}_2^T \operatorname{Re}LU(\mathbf{W}_1 u_i + \mathbf{W}_2 v_j + \mathbf{W}_3 I_h + \mathbf{b}_3) + b_4$$

$$w_h = \frac{\exp(w_h^*)}{\sum_{h'=1}^{H} \exp(w_{h'}^*)}$$

$$\tilde{r}_{ij} = \sum_{h=1}^{H} w_h \tilde{r}_{ij}^h$$

And I_h indicating a H-dimensional one-hot vector for view h.

The entire objective function is:

$$L = \sum_{i,j\in\Omega} (\tilde{r}_{ij} - r_{ij})^2 + \sum_{h=1}^{H} L_h + \lambda_v L_v + \lambda_r \,||\,\Theta\,||^2$$

Where $||\,\Theta\,||^2$ is the L2 norm of all parameters.

Step2: Personalized explanation generation

The purpose of this part is to select T features from Υ that are most useful I helping user i decide whether he will try item j.

The authors define a utility function first to judge 1) Whether user i is interested in F_l; 2) How well item j performs on F_l; 3) The weight of the view w_h that F_l belongs to. And the performance of node F_l is: $\psi(F_l) = (Z^h p_i^h)_l (Z^h q_j^h)_l w_h$.

So the problem turns out to be 1) Select T nodes from Υ; 2) argmax the sum of $\psi(F_i)$ where F_i is the selected nodes; 3)every node cannot be selected simultaneously with their ancestors in Υ.

We can use dynamic programming method to tackle the problem with time complexity of $\mathrm{O}(LMT^2)$, where L is the number of all nodes, M is the maximum number of children of a node, T is the target number of nodes.

4.Experiments

The experiment shows that this model achieves SOTA performance in terms of RMSE on three popular datasets. Besides that, the model can also provide explanations for users which are more accurate than other explainable recommendation models.