R2V-MIF: Rule-to-Vector Contrastive Learning and Multi-channel Information Fusion for Therapy Recommendation

Nengjun Zhu[1]*, **Jieyun Huang**[1], **Jian Cao** [2], **Liang Hu** [3], **Zixuan Yuan** [4], **Huanjing Gao** [1]
[1] School of Computer Engineering and Science, Shanghai University
[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[3] Tongji University
[4] Hong Kong University of Science and Technology (Guangzhou)
{zhu_nj, huang0615, shu_ghj}@shu.edu.cn, cao-jian@sjtu.edu.cn, lianghu@tongji.edu.cn,
zixuanyuan@hkust-gz.edu.cn

Abstract

Integrating data-driven and rule-based approaches is crucial for therapy recommendations since they can collaborate to achieve better performance. Medical rules, which are chains of reasoning that can infer therapies, widely exist. However, their symbolic and logical forms make integrating them with data-driven modeling technologies hard. Although rare attempts have indirectly modeled rules using data that supports them, the poor generalization of medical rules leads to inadequate supporting data and thus impairs the benefit of medical rules. To this end, we propose R2V-MIF, which fills the gap by rule-to-vector contrastive learning (R2V) and multi-channel information fusion (MIF). R2V is a data-free module and utilizes a hypergraph, including condition and result nodes, to instantiate the logic of medical rules. Each rule is reflected in the relations between nodes, and their representations are determined through contrastive learning. By taking rule representations as a bridge, MIF integrates the knowledge from medical rules, similar neighbors, and patient contents, and then recommends therapies. Extensive experiments show that R2V-MIF outperforms the baselines in several metrics using real-world medical data. Our code is available at https://github.com/vgeek-z/r2vmif.

1 Introduction

Devising a therapy recommendation system (TRS) is hot currently since it helps medical professionals make a more proper treatment plan for patients [Zhu *et al.*, 2023]. Data-driven and rule-based approaches are two typical types of TRS. The former is good at learning potential knowledge behind data using elaborate models [Zhu *et al.*, 2020; Min *et al.*, 2022] but requires high-quality data. While the latter is authoritative and explainable and can avoid outrageous recommendations, but it has poor generalization. Thus, a combination of them is wise as it can inherit the advantages from both of them and boost their collaborations.

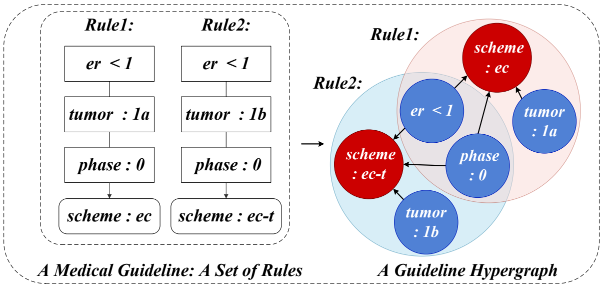


Figure 1: A snippet of medical rules and its hypergraph.

However, existing TRSs rarely integrate data-driven and rule-based approaches due to their incompatible paradigm. Specifically, the medical rules from guidelines are chains of reasoning that can infer therapies independently as shown in Figure 1. Their symbolic and logical forms make integrating them with data-driven modeling technologies hard. To overcome this issue, the work in [Zhu *et al.*, 2020] bypassed the joint modeling mechanism and proposed a heuristic collaborative strategy, i.e., filtering out the therapies that are recommended by data-driven approaches and violate medical rules. However, the strategy works only when the patient's situation matches a medical rule. Besides, DeepCtrl [Seo *et al.*, 2021] views supporting data, which matches the rules, as a substitution, and then applies a partial order modeling strategy on the substitution to learn rules indirectly. Obviously, the amount of supporting data could affect the modeling performance.

Although these efforts have a good try to integrate data-driven and rule-based approaches, their performance still suffers the following challenges: *1) Poor generalization of medical rules.* As aforementioned, the existing approaches require the target patient to match at least one medical rule or the target rule to have sufficient supporting data. However, the medical rules can only apply to minimal cases in practice, i.e., most patients can't obtain a recommended therapy through medical rules, and most medical rules have very limited supporting data. *2) Complicated collaboration over medical rules.* Existing approaches usually consider medical rules independently. However, the complicated collaboration over different medical rules exists and it can help promote

[*] Corresponding author.

medical rules' generalization. For instance, the most common preconditions of medical rules are sometimes enough to infer a result. By contrast, the missing of some insignificant preconditions might have a tiny effect on rule inference.

To this end, we propose R2V-MIF for therapy recommendations via ***r**ule-**to**-**v**ector* contrastive learning (R2V for short) and ***m**ulti-channel **i**nformation **f**usion* (MIF for short). R2V utilizes a hypergraph shown as the right part of Figure 1 to organize medical rules since it can well instantiate the logic of medical rules. For instance, the hypergraph has two types of nodes, where the *condition* node represents an attribute-relation-value tuple such as "$er < 1$", and the *result* node represents a therapy item such as "*ec-t*". Then, each hyperplane is associated with a rule, and the complicated collaboration over medical rules is reflected in the relations between nodes and hyperplanes. By doing so, contrastive learning can be applied to learn node representations, aiming at pulling the nodes inside a hyperplane close and pushing the nodes from different hyperplanes away. Besides, a graph convolutional network (GCN) and a path-based attention mechanism are applied to aggregate neighbor information. Since R2V learns the logic of medical rules directly using the hypergraph, it is a sample-free module and thus can avoid the influence of data limitations.

By taking rule representations from R2V as a bridge, MIF is proposed to integrate the knowledge from three channels: medical guidelines, similar neighbors, and patient contents. Specifically, in the guideline channel (GC), a representation is modeled for a patient by summarizing their matched *condition* nodes. It thus promotes the generalization of medical rules by extending a binary relation between patients and rules (i.e., matching or not) to continuous scales (i.e., the similarity between their representations). In the neighbor channel (NC), the patient representation from GC is further utilized to query similar patients, and then their therapy information is used for reference. In the content channel (CC), instead of transforming the content of patients into the rule representation space, MIF handles them directly to inherit the pure data-driven ability. Finally, the information from the three channels is adaptively fused for therapy recommendations.

Our contributions are summarized: 1) We propose a novel R2V-MIF, which integrates data-driven and rule-based approaches by simultaneously modeling the information from medical guidelines, similar neighbors, and patient contents for therapy recommendations. 2) We devise a new rule encoding method, i.e., R2V, which promotes the generalization of medical rules and supports data-driven approaches more flexibly. 3) We conduct extensive experiments to evaluate our approach, and the performance is superior to baselines.

2 Related Work

Incorporating rules for a data-driven approach is popular in many domains, such as retail and healthcare [Seo *et al.*, 2021; Wu *et al.*, 2023]. Recent works frequently considered two types of rules: 1) the ones that can only provide some indirect knowledge assisting the data-driven approach in finishing a task, i.e., classification, and 2) the other ones that can infer a result independently without a data-driven approach.

In the first branch, models [Yang *et al.*, 2021; Shang *et al.*, 2019] usually view the rules as auxiliary or constraint information. The contribution of these rules is limited for the final task compared to their partner, i.e., a data-driven approach. For example, MKHAN [Zhang *et al.*, 2019] uses a knowledge graph (KG) containing external regulations to enhance the interpretability of medical question answering (QA). But KG can't conduct QA tasks independently. Besides, the already-known drug interaction is considered as constraint rules in [Ren *et al.*, 2022]. Although it can restrain the model from recommending drugs having a negative interaction, it cannot determine a drug for patients directly. Similarly, the work in [Zheng *et al.*, 2020] treated the position constraints as descriptive rules, e.g., "a lesion must be attached to tissues", and proposed a mathematical encoding to transform descriptive knowledge. However, rules are treated as a regularized item, which can only influence the final task indirectly. Most current research falls into this branch since medical domains have abundant such scattered prior knowledge.

In the second branch, the decision rules are more informative and can act as an independent agent. They significantly influence the task if combined with a data-driven approach. For example, the authors [Zhu *et al.*, 2020] used the inferred results from an authoritative medical guideline which is a set of decision rules, to a data-driven approach for recommendation correction. It is a filtering strategy and ensures the recommended therapies from a data-driven approach do not violate the authoritative guidelines. The cooperation is rough and requires vast rules to cover as many cases as possible. Besides, DeepCtrl [Seo *et al.*, 2021] is a fusion framework that encodes decision rules and data simultaneously. It assumes a partial order "$\succ$" between a sample x^+, which completely matches a rule, and a sample x^-, which partially matches the rule. If the rule infers a result y, the former is more likely to be categorized to y than the latter, i.e., $y_{x^+} \succ y_{x^-}$, where x^- is generated for x^+ using a perturbation technology. Then, a partial order loss and a conventional classification loss are combined for training. Since "$\succ$" is defined on sample relations, a rule can only be considered when at least one sample matches it perfectly.

Our work follows the second branch, and the main difference is that our approach incorporates a sample-free rule encoding module. The learned rule embeddings can be freely applied to any other data-driven approach to boost their collaborations. Besides, based on rule embeddings, we easily fuse multi-channel information, including a data channel, to enhance therapy recommendations.

3 R2V-MIF Framework

In this section, we first describe the framework of our proposed R2V-MIF, i.e., a therapy recommendation approach based on *r*ule-*to*-*v*ector contrastive learning and *m*ulti-channel *i*nformation *f*usion, and then introduce its two main modules, i.e., R2V and MIF, in detail.

3.1 Overview

Our R2V-MIF comprises two main modules, i.e., R2V and MIF, shown as Figure 2. R2V first constructs a medical

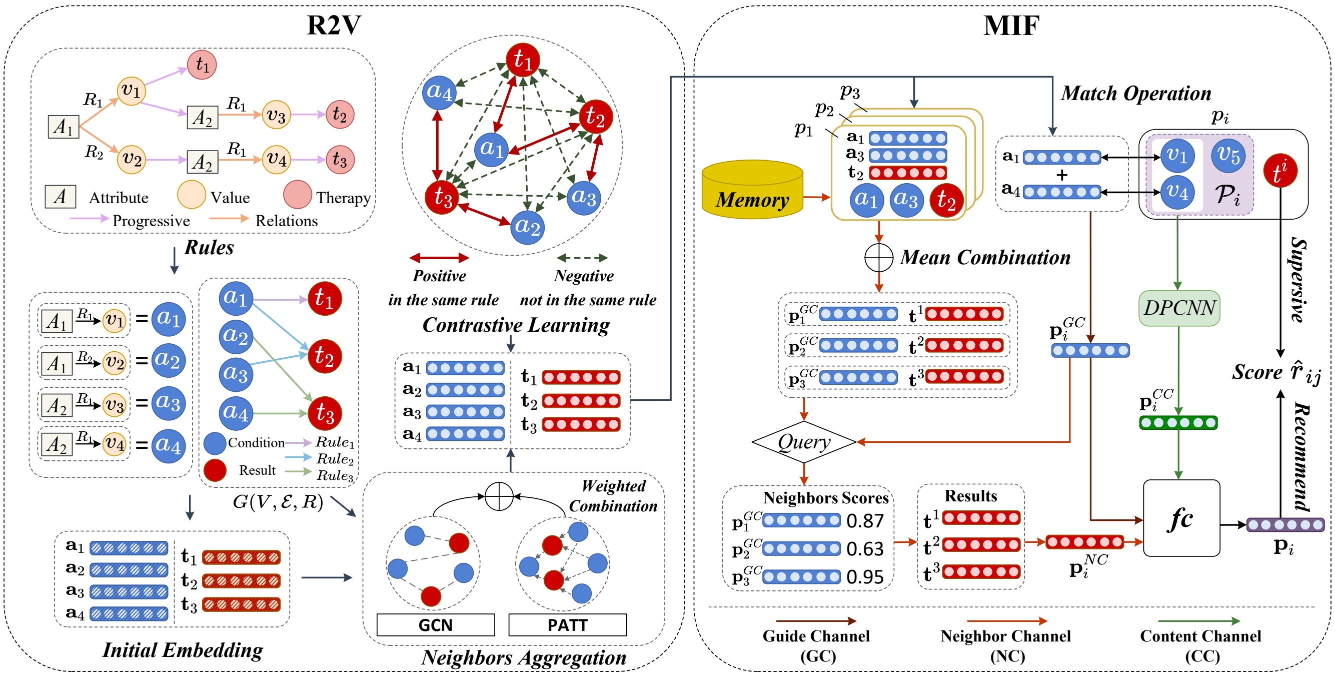


Figure 2: The overall framework of our proposed R2V-MIF.

guideline hypergraph according to medical rules. Based on the hypergraph, neighbor aggregating and contrastive learning are applied to learn the representation of condition nodes, result nodes, and rules. Then, building upon the representations, MIF models patient information from guidelines, neighbors, and contents. The guideline channel evaluates the coincidence degree between patient information and medical rules; The neighbor channel queries the similar patients in the rule representation space and refers to neighbors' results; The content channel models patient information directly without considering medical rules. Finally, the three channels are fused to make therapy recommendations. Next, we introduce each module in detail.

3.2 R2V Module

Medical Guideline Hypergraph Construction

The first step of R2V is to construct a hypergraph for rules from medical guidelines. The hypergraph can better describe medical rules including their conditions and results, and model their relationship. Next, we declare some definitions first.

Definition 1 (Condition and Result Node) *Several conditions and one result constitute a medical rule. For example, $a_1 \wedge a_2 \xrightarrow{infer} t_i$ is a rule, where a_1 and a_2 are two conditions, which are attribute-relation-value tuples, e.g., "$er \geq 1$" and "$tumor = 1a$", and t_i is a result, which is a recommended therapy item, e.g., "ec-t". Accordingly, we define two types of nodes, i.e.,* ***condition nodes*** *and* ***result nodes****, for all medical rules.*

Definition 2 (Guideline Hypergraph) *The guideline hypergraph is defined by $G = (\mathcal{V}, \mathcal{E}, \mathcal{R})$: $\mathcal{V}$ is a set of all condition and result nodes introduced in Definition 1; $\mathcal{E}$ is a set of weighted edges, and we construct them according to the following strategy: if a condition and a result exist in the same rule, there is an edge between their corresponding nodes in G, and the weight is one as default; $\mathcal{R} = \{\mathcal{V}_1, \mathcal{V}_2, \cdots, \mathcal{V}_r, \cdots\}$ is a partition of nodes according to rules, where $\mathcal{V}_r$ is a node set of a sub-graph/hyperplane, marked by rule r, i.e., the condition and result nodes of rule r are recorded in $\mathcal{V}_r$.*

All medical rules are transformed into a guideline hypergraph according to Definition 2. Then, the relations between sub-graphs, i.e., rules, and between nodes, i.e., conditions and results, can be learned by hypergraph learning.

Condition and Result Embedding

Each node x_i, which can be a_i for a condition node and t_i for a result node, is initialized by a randomized and trainable embedding $\boldsymbol{x}_i \in \mathbb{R}^K$, which is a column vector. Specially, $\boldsymbol{a}_i \in \mathbb{R}^K$ and $\boldsymbol{t}_j \in \mathbb{R}^K$ are an embedding of a_i and t_j, respectively, where K is the dimensionality. Then, we utilize a GCN [Kipf and Welling, 2016] and a path-based attention mechanism to capture and enhance node connections.

Regarding **GCN**, the representation $\boldsymbol{x}_i^{GCN}$, which incorporates the structural information of a graph to node x_i, is defined as follows:

$$\boldsymbol{x}_i^{GCN} = \boldsymbol{\Theta}^\top \sum_{x_k \in \mathcal{N}_i \cup \{x_i\}} \frac{w_{ik}}{\sqrt{d_i d_k}} \boldsymbol{x}_k \qquad (1)$$

where $\mathcal{N}_i$ is a set of 1-hop neighbors of node x_i, w_{ik} represents the edge weight between node x_i and x_k, and its value is 1 in our settings, d_i and d_k are the vertex degrees, and $\boldsymbol{\Theta} \in \mathbb{R}^{K \times K}$ is a trainable parameter of GCN. Thus, there is an extra condition embedding $\boldsymbol{a}_i^{GCN}$ and result embedding $\boldsymbol{t}_j^{GCN}$ according to Equ. (1).

The path-based attention mechanism (PATT) considers the path "$a_i \leftrightarrow t_j \leftrightarrow a_k$", in which t_j is a bridge, to enhance the relations between condition nodes. Inspired by a self-attentive mechanism [Vaswani *et al.*, 2017; Niu *et al.*, 2021], we treat $\boldsymbol{t}_j$ as *Query*, and construct *Key* $=$ *Value* $= \boldsymbol{N}_j$ as follows:

$$\boldsymbol{N}_j = [\cdots, \boldsymbol{a}_k, \cdots], a_k \in \mathcal{N}_j \tag{2}$$

Then, a fused representation $\boldsymbol{t}'_j$ is generated as follows:

$$\boldsymbol{t}'_j = softmax\left(\boldsymbol{t}_j^\top \frac{\boldsymbol{N}_j}{\sqrt{K}}\right)\boldsymbol{N}_j^\top \tag{3}$$

By fusing the bridge $\boldsymbol{t}'_j$, the connection between a_i and a_k in the path "$a_i \leftrightarrow t_j \leftrightarrow a_k$" is established as follows:

$$\boldsymbol{a}_i^{PATT} = mean([\cdots, \boldsymbol{t}'_j, \cdots]), t_j \in \mathcal{N}_i \tag{4}$$

Using the same way, we can obatin $\boldsymbol{t}_i^{PATT}$ by considering the path "$t_i \leftrightarrow a_j \leftrightarrow t_k$". Finally, the initial representation of a condition and result node is updated by the following weighted combination:

$$\boldsymbol{x}_i = \alpha \boldsymbol{x}_i + \beta \boldsymbol{x}_i^{GCN} + (1 - \alpha - \beta)\boldsymbol{x}_i^{PATT} \tag{5}$$

where α and β are two hyper-parameters balancing the contributions of different embeddings.

Contrastive Learning

Contrastive learning [Hassani and Khasahmadi, 2020; Zhang *et al.*, 2023] can be utilized to model the position of nodes in a hidden high-dimensional space. The distance between positions in the space reveals the relationship between the nodes in their graph. We thus use contrastive learning to narrow the distance between $\boldsymbol{a}_i$ and $\boldsymbol{t}_j$ if there exists a rule r subject to $a_i \in \mathcal{V}_r \wedge t_j \in \mathcal{V}_r$, i.e., a_i and t_j belong to the same sub-graph regarding rule r. By contrast, if a_i and t_j have no such a rule, we should push $\boldsymbol{a}_i$ and $\boldsymbol{t}_j$ away from each other. If two conditions are not necessary to be considered jointly in the same rule, the difference between their nodes should be increased. Besides, we also magnify the distance between result nodes. Accordingly, we construct a set of positive pairs and a set of negative pairs for contrastive learning as follows:

$$\begin{aligned} \mathcal{O}^+ &= \{(a_i^+, t_j^+) \mid \exists \mathcal{V}_r \in \mathcal{R}, (a_i, t_j \in \mathcal{V}_r)\} \\ \mathcal{O}^- &= \{(x_i^-, x_j^-) \mid \nexists \mathcal{V}_r \in \mathcal{R}, (x_i, x_j \in \mathcal{V}_r)\} \end{aligned} \tag{6}$$

We calculate the distance between nodes as a dot product, followed by a Sigmoid transform:

$$\hat{y} = \sigma(\boldsymbol{x}_i^\top \boldsymbol{x}_j) \tag{7}$$

where $\sigma(\cdot)$ is a Sigmoid function mapping a scalar to a $(0, 1)$ scale. We then label a positive pair $e_i \in \mathcal{O}^+$ as $y_i = 1$, and a negative pair $e_i \in \mathcal{O}^-$ as $y_i = 0$. Finally, we minimize a cross-entropy loss as follows to train node representations.

$$L_{rule} = -\sum_{e_i \in \mathcal{O}^+ \cup \mathcal{O}^-} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i) \tag{8}$$

Besides, we can have a rule representation $\boldsymbol{r}$ by summarizing the embedding of its condition notes.

$$\boldsymbol{r} = \frac{1}{|\mathcal{V}_r| - 1} \sum_{a_i \in \mathcal{V}_r - \{t^r\}} \boldsymbol{a}_i \tag{9}$$

where t^r is the result node of rule r. The distance between $\boldsymbol{r}$ and $\boldsymbol{t}^r$ should be narrowed. Thus, we update L_{rule} as follows:

$$L_{rule} = L_{rule} - \sum_r \log(\sigma(\boldsymbol{r}^\top \boldsymbol{t}^r)) \tag{10}$$

3.3 MIF Module

Based on the node representations, i.e., $\boldsymbol{a}_i$ and $\boldsymbol{t}_j$ in Equ. (5), MIF models a patient information from the guideline channel, neighbor channel, and content channel, simultaneously.

Guideline Channel Modeling

Regarding the target patient p_i, they have a set of sparse attribute-value pairs (AVPs) denoted by

$$\mathcal{P}_i = \{\cdots, v_j, \cdots\} \tag{11}$$

where v_j is an AVP, e.g., "*er value* $= 0.1$", and can be viewed as a particular case of an attribute-relation-value tuple in which "relation" is "$=$". Thus, v_j can be utilized to match condition nodes in the guideline hypergraph. Note that the AVP "*er value* $= 1$" can also match the condition node "*er value* ≥ 0". Then, patient p_i has a set of matched condition nodes denoted by

$$\mathcal{M}_i = \{\cdots, a_j, \cdots\} \tag{12}$$

where a_j is matched by v_j. The size $|\mathcal{M}_i|$ is usually less than $|\mathcal{P}_i|$ since not all v_j can be matched. However, as we discussed before, the matched v_j usually has a higher importance as it appears in medical guidelines. Thus, we generate a representation of patient p_i guided by medical guidelines.

$$\boldsymbol{p}_i^{GC} = \frac{1}{|\mathcal{M}_i|} \sum_{a_j \in \mathcal{M}_i} \boldsymbol{a}_j \tag{13}$$

By doing so, even if the patient can't match a rule completely, medical guidelines can also contribute to the decision of therapies as long as $\mathcal{M}_i$ is not empty.

Neighbor Channel Modeling

The determined therapies of similar patients have great reference significance to the therapy decision of the target patient [Zhu *et al.*, 2020]. Therefore, we devise a neighbor channel to incorporate this information. However, unlike traditional approaches, we query similar patients using condition nodes as a bridge. Specifically, for historical cases, each patient p_m has a pair

$$(\boldsymbol{p}_m^{GC}, \boldsymbol{t}^m)|p_m \tag{14}$$

where $\boldsymbol{t}^m$ is equal to $\boldsymbol{t}_j^m$ and its j is omitted for convenience. It is a result node embedding and matched by therapy item t^m determined for patient p_m.

Then, we query historical cases p_m by evaluating their similarities to the target patient p_i:

$$s_{i,m} = \frac{\exp(s'_{i,m})}{\sum_{p_{m'} \in \mathcal{S}_i} \exp(s'_{i,m'})}$$

$$s'_{i,m} = \frac{\boldsymbol{p}_i^{GC\top} \boldsymbol{p}_m^{GC}}{\|\boldsymbol{p}_i^{GC}\| \|\boldsymbol{p}_m^{GC}\|} \tag{15}$$

where $\mathcal{S}_i$ is the set of top-K patients, and K is tuned according to datasets; $\| \cdot \|$ is an L2-normalization. The therapy items of top-K most similar patients are combined to obtain a neighbor-guided representation:

$$\boldsymbol{p}_i^{NC} = \sum_{p_m \in \mathcal{S}_i} s_{i,m} \boldsymbol{t}^m \tag{16}$$

Content Channel Modeling

The $\boldsymbol{p}_i^{GC}$ in Equ. (13) is formed by selecting AVPs according to medical guidelines; thus, the AVPs that do not appear in medical guidelines but are valuable are neglected. To address this issue, we incorporate a traditional content-based DPCNN [Johnson and Zhang, 2017] to model all possible patient information. DPCNN is a widely efficient deep classification method that is pyramid-shaped, and is selected as it performs well compared to others.

The AVPs of a patient are preprocessed according to attribute types. One-hot encoding is employed for a discrete attribute; continuous attributes are encoded to a denser representation by a fully connected layer network (FC). The missing value of attributes is set to 0 for alignment. Then, all outputs are concatenated, which are further fed to DPCNN to obtain a content-based representation as follows:

$$\boldsymbol{p}_i^{CC} = DPCNN(fc_1(concat(preproc(\mathcal{P}_i)))) \tag{17}$$

where $fc_1(\cdot)$, $concat(\cdot)$, and $preproc(\cdot)$ represent FC, concatenation, and preprocessing, respectively.

For an extreme situation that there is no AVP of patient p_i matched by a condition node, i.e., $|\mathcal{M}_i| = 0$, $\boldsymbol{p}_i^{CC}$ ensures that our approach still can provide therapy recommendations based on a pure data-driven strategy.

Three-channel Fusion

The final representation of a patient is generated by fusing the information from the three channels.

$$\boldsymbol{p}_i = fc_2(concat(\boldsymbol{p}_i^{GC}, \boldsymbol{p}_i^{NC}, \boldsymbol{p}_i^{CC})) \tag{18}$$

where $fc_2(\cdot)$ is another FC.

3.4 Model Inference and Recommendation

The fused representation $\boldsymbol{p}_i$ is utilized to generate a probability vector $\boldsymbol{y}_i \in \mathbb{R}^N$ for therapy recommendations as follows:

$$\hat{\boldsymbol{y}}_i = \sigma(fc_3(\boldsymbol{p}_i)) \tag{19}$$

where N is the number of therapy items, and each dimension $\hat{\boldsymbol{y}}_{i,j}$ represents the recommendation score of therapy item t_j.

In the training phase, each patient has a one-hot label vector $\boldsymbol{y}_i$, in which $\boldsymbol{y}_{i,j} = 1$ if t_j is the ground truth, otherwise $\boldsymbol{y}_{i,j} = 0$. We then minimize the following cross-entropy loss to train model parameters.

$$L_{rec} = -\sum_{(p_i, t_j)} \boldsymbol{y}_{i,j} \log \hat{\boldsymbol{y}}_{i,j} \tag{20}$$

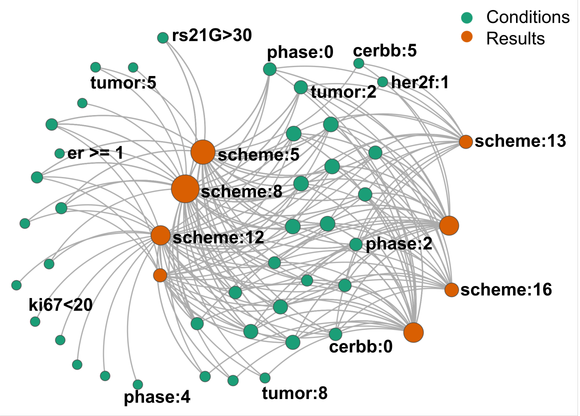


Figure 3: The hypergraph whose hyperplane is omitted.

4 Experiment

4.1 Decision Rule and Dataset

We extract 23 rules for breast cancer from NCCN guidelines[1] [Gradishar *et al.*, 2018]. Then, there are 41 unique condition nodes and 8 result nodes for constructing a medical guideline hypergraph shown as Figure 3. Each rule involves 16.43 condition nodes in average and one result node.

BCDB [Zhu *et al.*, 2020] is a real-world medical dataset recording the physical indicators and chemotherapy items (therapies for short) of patients with breast cancer. To better analyze the effect of medical rules, we select the records with at least one attribute appearing in the rules to form our dataset. Then, the final dataset contains 2,648 records, 12 attributes, and 8 therapies. About 18.35% of records can match at least one of the above medical rules completely.

4.2 Evaluation Metric

We employ commonly-used Hit Rate (HR), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG) to assess R2V-MIF and the baselines' performance.

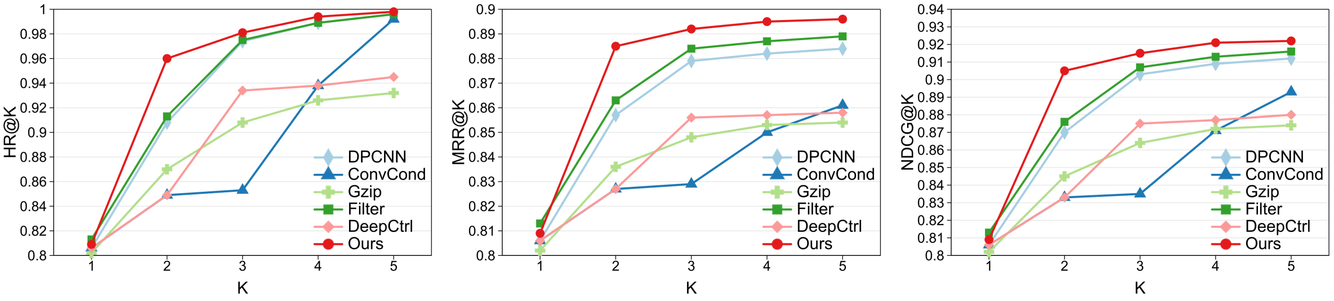
HR is calculated as the proportion of the cases that the actual therapies are in the top-K recommendations. **MRR** evaluates the position of actual therapies in the recommendation list. The higher the rank of the actual therapies in the recommendation list, the higher the value of MRR. **NDCG** is similar to MRR but simultaneously considers the position and predicted relevance of actual therapies.

4.3 Baseline

We compare R2V-MIF with three typical data-driven approaches without considering medical rules, i.e., Gzip, ConvCond, and DPCNN, and two recent neural networks considering medical rules, i.e., Filter and DeepCtrl. All data-driven parts of baselines have the same data preprocessing as that in our content channel modeling.

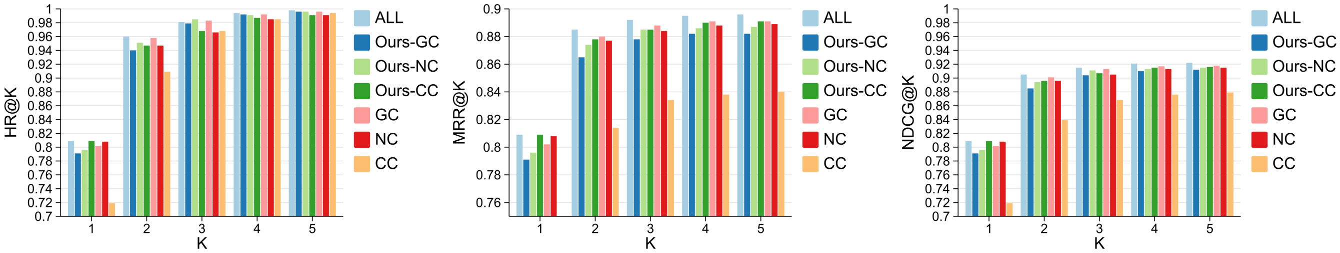
Gzip [Jiang *et al.*, 2023] is a simple, lightweight, and universal alternative to NNs. It comprises a lossless compressor, a compressor-based distance metric, normalized compression distance, and a neighbor classifier.

[1] https://www.nccn.org/



(a) HR Performance
(b) MRR Performance
(c) NDCG Performance

Figure 4: Performance comparison of different approaches.



(a) HR Performance
(b) MRR Performance
(c) NDCG Performance

Figure 5: Performance of variants of the proposed R2V-MIF.

ConvCond [Yang *et al.*, 2019] uses the idea of dynamic convolution to learn different convolution kernel parameters for different samples, thereby increasing its modeling ability.

DPCNN [Johnson and Zhang, 2017] is a deep pyramid neural network that emphasizes long-range associations in the text while maintaining a lower computational complexity.

Filter adopts the filtering strategy in [Zhu *et al.*, 2020] to handle medical rules. But the difference is that we select DPCNN as a data-driven approach as that in R2V-MIF for a better comparison.

DeepCtrl [Seo *et al.*, 2021] is a fusion framework that encodes medical rules and data simultaneously. A partial order loss and a conventional classification loss are combined for training based on a set of positive-negative sample pairs.

4.4 Performance Comparison

We first introduce some default settings in our experiments. For all approaches, the number of epochs and the value of batch size are tuned to 12 and 256, respectively. Since the sample size is limited and the models converge quickly, we set the learning rate to a fixed value of 0.01. Then, for our approach, we query 10 most similar cases in the neighbor channel modeling and the parameters α and β in Equ. (5) are both set to 0.33 for the best performance. Then, all approaches' performance on the BCDB dataset is shown in Figure 4. We can have three main conclusions:

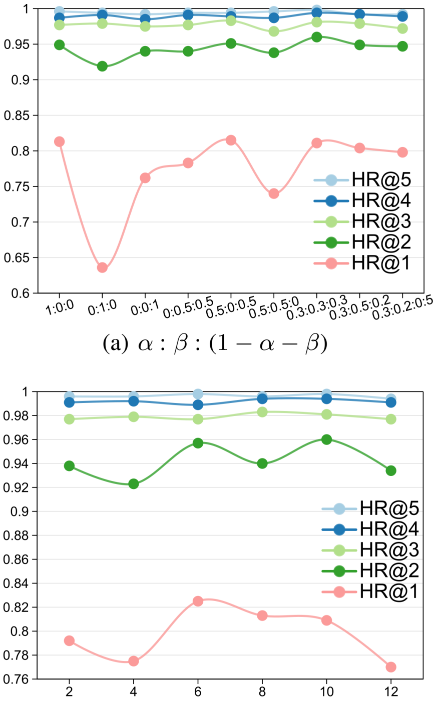
1) Over the three pure data-driven approaches, i.e., Gzip, ConvCond, and DPCNN, DPCNN achieves the best performance in all situations. The improvement might come from their various network structures, which is outside the scope of our discussion. However, we thus select DPCNN as the data-driven part of our model.

2) Although on the top-1 recommendations, the performance difference of all approaches is insignificant, our performance line is on top of all baselines. Especially it greatly improves the top-2 recommendations, e.g., increasing HR@2 by about 5.15% compared to the second-best performance, i.e., Filter, and about 13.1% compared to the worst performance, i.e., ConvCond. The observation has proven the effectiveness of our framework.

3) The performance of Filter is better than that of DPCNN. The only difference between them is that the former considers medical rules, showing that it is crucial for a data-driven approach to incorporate medical rules. However, the improvement is limited since only 18.35% samples can completely match a decision rule. Besides, DeepCtrl, another approach combining rules and data, has results that are not as good as expected. As mentioned, sufficiently matched samples must be required for DeepCtrl to learn rules. By contrast, our R2V-MIF is unaffected by data limitation, and the knowledge from medical rules is modeled well.

4.5 Ablation Study

We propose six variants to verify the channel influence on R2V-MIF by combining or eliminating different channels. The variants' performance is shown as Figure 5, where **ALL**



(a) $\alpha : \beta : (1 - \alpha - \beta)$

(b) K

Figure 6: Parameter analysis.

represents R2V-MIF in which three channels are considered; the notation "**-**" represents the removal of a channel, e.g., **Ours-GC** indicates the guideline channel modeling is removed from R2V-MIF. Besides, **GC**, **NC**, and **CC** represent the variants that only keep the corresponding channel and eliminate others, e.g., **GC** only keeps the guideline channel.

We then have the following observations: 1) **ALL** outperforms others, showing the effectiveness of its structure. 2) Except **ALL**, **GC** achieves a better performance than others in most situations. That means R2V has a powerful ability to encode rules. As a result, medical rules can also infer a suitable result for the case that partially matches the rules.

4.6 Parameter Analysis

We analyze the parameter α and β in Equ. (5) and the number of neighbors queried in the neighbor channel, i.e., K. Others are fixed when exploring one of them. To save space, we only exhibit HR performance in different settings.

The influence of α and β is shown in Figure 6(a). The setting $\alpha : \beta : 1{-}\alpha{-}\beta = 0.33 : 0.33 : 0.33$ achieves the best performance. It is equal to combining all node embeddings by a mean operation and indicates that GCN and PATT should be involved for better embedding. The setting $\alpha : \beta : 1{-}\alpha{-}\beta = 0.5 : 0.5 : 0$ significantly decreases the HR value, indicating that PATT is important.

The impact of parameter K is shown in Figure 6(b). The best result is obtained when K is set to 6 on HR@1 and is 0.825. However, we set the value of K to 10 as default since the other metrics, e.g., HR@2, HR@3, and HR@4, achieve a great performance. That is why the top-1 performance of our approach is similar to others.

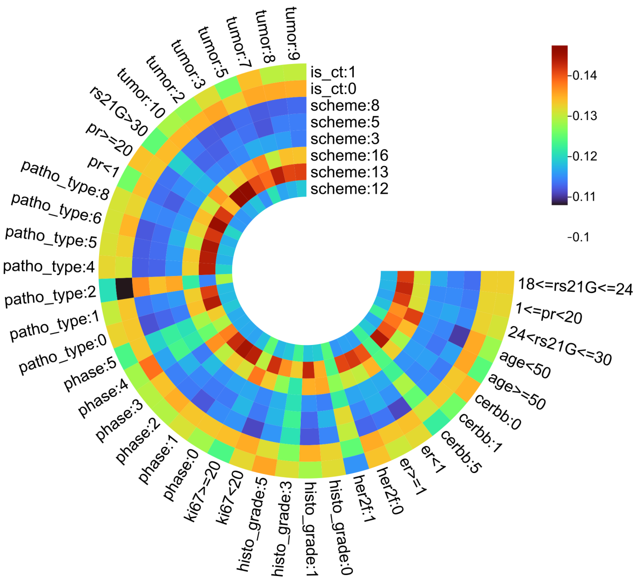


Figure 7: Cosine similarity between pre-trained nodes.

4.7 Case Study

Figure 7 is a heat map[2] and shows the cosine similarity between condition and result nodes in the learned embedding space. We can observe that a condition node has various similarities to different result nodes, i.e., the condition has different importances when inferring different results. There are some other findings. For example, the condition "patho_type:2" has a tiny influence on determining the result "is_ct:0", but should be carefully considered when evaluating whether "scheme:8" is appropriate for a patient. Besides, the result node "scheme:13" is close to almost all condition nodes because most of them appear in the same rule.

5 Conclusion

In this work, we proposed R2V-MIF, which integrates data-driven and rule-based approaches for therapy recommendations. Specifically, based on the constructed medical guideline hypergraph, the relations between rules are modeled as distances in a rule representation space. Then, the information from guidelines, neighbors, and contents is fused by taking the rule representation, i.e., condition and result embeddings, as a bridge. R2V-MIF achieved excellent performance on a real-world medical dataset, proving the effectiveness of rule modeling and channel fusion. There are several directions for future work. For example, the relations between conditions and results can be elaborated further by defining more positive or negative sample types for contrastive learning. Besides, more technologies and strategies [Xu *et al.*, 2021] to fuse multi-aspect information can be explored.

[2] Generated using ChiPlot (https://www.chiplot.online/)

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62202282) and the Shanghai Youth Science and Technology Talents Sailing Program (No. 22YF1413700). It is also partially sponsored by the Interdisciplinary Program of Shanghai Jiao Tong University (No. YG2024QNB05), and the Guangzhou-HKUST (GZ) Joint Funding Program (No. 2024A03J0630).

References

[Gradishar *et al.*, 2018] William J Gradishar, Benjamin Anderson, Ron Balassanian, et al. Breast cancer, nccn clinical practice guidelines in oncology. *Journal of the National Comprehensive Cancer Network*, 16(3):310–320, 2018.

[Hassani and Khasahmadi, 2020] Kaveh Hassani and Amir Hosein Khasahmadi. Contrastive multi-view representation learning on graphs. In *Proceedings of the International Conference on Machine Learning*, pages 4116–4126, 2020.

[Jiang *et al.*, 2023] Zhiying Jiang, Matthew Yang, Mikhail Tsirlin, Raphael Tang, Yiqin Dai, and Jimmy Lin. "low-resource" text classification: A parameter-free classification method with compressors. In *Findings of the Association for Computational Linguistics*, pages 6810–6828, 2023.

[Johnson and Zhang, 2017] Rie Johnson and Tong Zhang. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 562–570, 2017.

[Kipf and Welling, 2016] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[Min *et al.*, 2022] Xin Min, Wei Li, Jinzhao Yang, Weidong Xie, and Dazhe Zhao. Dual-level diagnostic feature learning with recurrent neural networks for treatment sequence recommendation. In *Journal of Biomedical Informatics*, volume 134, page 104165, 2022.

[Niu *et al.*, 2021] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[Ren *et al.*, 2022] Yongjian Ren, Yuliang Shi, Kun Zhang, Xinjun Wang, Zhiyong Chen, and Hui Li. A drug recommendation model based on message propagation and ddi gating mechanism. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3478–3485, 2022.

[Seo *et al.*, 2021] Sungyong Seo, Sercan Arik, Jinsung Yoon, Xiang Zhang, Kihyuk Sohn, and Tomas Pfister. Controlling neural networks with rule representations. *Advances in Neural Information Processing Systems*, 34:11196–11207, 2021.

[Shang *et al.*, 2019] Junyuan Shang, Cao Xiao, Tengfei Ma, Hongyan Li, and Jimeng Sun. Gamenet: Graph augmented memory networks for recommending medication combination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1126–1133, 2019.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Wu *et al.*, 2023] Likang Wu, Junji Jiang, Hongke Zhao, Hao Wang, Defu Lian, Mengdi Zhang, and Enhong Chen. Kmf: knowledge-aware multi-faceted representation learning for zero-shot node classification. *arXiv preprint arXiv:2308.08563*, 2023.

[Xu *et al.*, 2021] Yuanbo Xu, En Wang, Yongjian Yang, and Yi Chang. A unified collaborative representation learning for neural-network based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(11):5126–5139, 2021.

[Yang *et al.*, 2019] Brandon Yang, Gabriel Bender, Quoc V Le, and Jiquan Ngiam. Condconv: Conditionally parameterized convolutions for efficient inference. *Advances in Neural Information Processing Systems*, 32, 2019.

[Yang *et al.*, 2021] Chaoqi Yang, Cao Xiao, Fenglong Ma, Lucas Glass, and Jimeng Sun. Safedrug: Dual molecular graph encoders for recommending effective and safe drug combinations. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 3735–3741, 2021.

[Zhang *et al.*, 2019] Yingying Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering. In *Proceedings of the ACM International Conference on Multimedia*, pages 1089–1097, 2019.

[Zhang *et al.*, 2023] Yifei Zhang, Hao Zhu, Zixing Song, Piotr Koniusz, and Irwin King. Spectral feature augmentation for graph contrastive learning and beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11289–11297, 2023.

[Zheng *et al.*, 2020] Zhiyang Zheng, Hao Yan, Frank C Setzer, Katherine J Shi, Mel Mupparapu, and Jing Li. Anatomically constrained deep learning for automating dental cbct segmentation and lesion detection. *IEEE Transactions on Automation Science and Engineering*, 18(2):603–614, 2020.

[Zhu *et al.*, 2020] Nengjun Zhu, Jian Cao, Kunwei Shen, Xiaosong Chen, and Siji Zhu. A decision support system with intelligent recommendation for multi-disciplinary medical treatment. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(1s):1–23, 2020.

[Zhu *et al.*, 2023] Nengjun Zhu, Jieyun Huang, Jian Cao, Xinjiang Lu, Hao Liu, and Hui Xiong. Mtirec: A medical test recommender system based on the analysis of treatment programs. In *Proceedings of the IEEE International Conference on Data Mining*, pages 898–907, 2023.