TabGNN  
**摘要（Abstract）翻譯與重點歸納**

**翻譯：**

表格數據預測（Tabular Data Prediction，TDP）是最受工業界關注的應用之一，許多方法被設計來提升預測表現。然而，現有的工作主要專注於特徵交互（Feature Interactions），卻忽略了樣本之間的關係，例如：具有相同教育水平的用戶可能有類似的債務償還能力。

在本研究中，我們透過顯式且系統性地建模樣本關係，提出了一種新穎的框架 **TabGNN**，該方法基於近年流行的 **圖神經網絡（Graph Neural Networks, GNN）**。具體來說，我們首先構建 **多重圖（Multiplex Graph）** 來建模多方面的樣本關係，然後設計 **多重圖神經網絡（Multiplex Graph Neural Network）** 來學習每個樣本的增強表示（Enhanced Representation）。

為了將 TabGNN 與公司內部的表格數據解決方案整合，我們將學習到的嵌入（Embeddings）與原始特徵進行拼接，然後將其輸入到預測模型中。實驗結果顯示，在來自多個領域的 11 個表格數據集（包括分類和回歸任務）上，TabGNN **能夠持續提升預測性能，相較於 4Paradigm 公司的 AutoFE 解決方案有穩定的改進**。

**重點歸納**

1. **問題背景：**
   * 表格數據預測（TDP）是工業界的熱門應用，如詐欺檢測、銷售預測、推薦系統等。
   * 傳統方法主要聚焦於 **特徵交互**，但忽略了 **樣本間的關係**。
2. **研究貢獻：**
   * 提出 **TabGNN**，一個基於圖神經網絡的 TDP 方法。
   * **建模多重樣本關係**，透過 **多重圖（Multiplex Graph）** 表示不同類型的樣本關係。
   * 設計 **多重圖神經網絡（Multiplex GNN）** 來學習樣本的增強特徵表示。
   * **與 AutoFE 整合**，通過拼接特徵提升預測能力。
3. **實驗結果：**
   * 在 11 個數據集（分類與回歸任務）上進行測試。
   * **TabGNN 相較於 AutoFE 持續提升預測性能**。

**1. 引言（Introduction）翻譯與重點歸納**

**翻譯：**

近年來，機器學習在各種應用場景中取得了巨大成功，其中 **表格數據預測（Tabular Data Prediction, TDP）** 是最受工業界關注的應用之一，例如：

* **詐欺檢測（Fraud Detection）**【銀行業】
* **銷售預測（Sales Prediction）**【零售業】
* **線上廣告（Online Advertisement）**【行銷科技】
* **推薦系統（Recommender Systems）**【電子商務】

在這些應用中，數據通常以表格形式存儲，其中 **每一行代表一個樣本，每一列代表一個特徵**。TDP 的主要目標是 **根據一行的數據（樣本）來預測特定目標欄位的值**，例如：

* **預測商品的銷售量**（回歸任務）
* **預測銀行交易是否異常**（分類任務）

圖 **Figure 1** 展示了一個 **金融貸款應用的表格數據示例**：

* 每一行代表一個貸款用戶。
* 每一列是相關的特徵，例如 **教育水平（Education）**、\*\*所在城市（City）\*\*等。
* 目標欄位為 **是否準時還款（Overdue）**，即 **我們的預測目標**。

根據 Kaggle 最新的調查報告【16】，TDP **是當前最常見的數據科學任務之一**。此外，**4Paradigm（本研究所屬公司）** 作為 AI 技術服務商，在 TDP 領域擁有廣泛的商業應用，涵蓋 **銀行、醫療、電商等行業**，並且 **TDP 相關業務貢獻了公司大部分收入**。因此，設計高效的 TDP 解決方案極為重要。

**現有挑戰**

然而，在實踐中，TDP **仍然面臨重大挑戰**：

1. 直接使用 **原始特徵** 來進行預測通常效果不佳，因為：
   * **單獨使用某個特徵的資訊有限**
   * **特徵之間的交互** 可能攜帶更豐富的信息。
2. **現有方法的局限性**：
   * 許多方法專注於 **特徵交互（Feature Interactions）**，例如：
     + **線性模型（LR, Logistic Regression）**
     + **樹模型（GBM, Gradient Boosting Machines）**
     + **深度學習模型（DeepFM）**
     + **自動特徵工程（AutoCross）**
   * 目前，**4Paradigm 公司的 TDP 解決方案 AutoFE** 已經整合了上述技術，並在許多真實場景中成功部署，**但仍然存在改進空間**。
3. **忽略樣本之間的關係（Sample Relations）**
   * 以往的 TDP 方法 **主要關注特徵的組合**，但 **沒有考慮樣本之間的關聯性**。
   * 例如：
     + **擁有相同教育水平的用戶** 可能有類似的還款能力。
     + **年齡相近的用戶** 可能有相似的消費習慣。
   * 在 **流形學習（Manifold Learning）** 和 **幾何深度學習（Geometric Deep Learning）** 研究中，樣本關係已被證實對預測結果有幫助。
4. **樣本關係具有多種形式（多重關係）**
   * 樣本之間可能具有 **多種不同的關聯**，例如：
     + **用戶 A 和用戶 B 來自相同城市（地理關係）**
     + **用戶 A 和用戶 C 擁有相同職業（社會關係）**
   * 這些 **多重關係（Multiplex Relations）** 需要被合理建模，但現有方法無法捕捉這些資訊。

**本文貢獻**

為了解決這些問題，我們提出了一個基於 **多重圖（Multiplex Graph）** 的解決方案——**TabGNN**：

1. **樣本關係建模：**
   * 透過構造 **多重圖（Multiplex Graph）** 來顯式建模樣本間的多種關係。
   * **多重圖允許同一個節點（樣本）之間有多種不同類型的邊（關係）**，例如 **教育相似性、年齡相似性、職業相似性**。
2. **設計多重圖神經網絡（Multiplex GNN）：**
   * 設計一種 **多重圖神經網絡（Multiplex Graph Neural Network）**，讓樣本能夠從其 **關聯樣本中學習到更豐富的特徵信息**。
3. **與 AutoFE 整合**
   * **TabGNN 的輸出特徵將與原始特徵進行拼接**，然後再輸入 AutoFE 進行最終預測。
   * **這種方法既能夠保留特徵交互信息，也能充分利用樣本關係**。
4. **實驗驗證**
   * **在 11 個來自不同領域的表格數據集上進行測試（分類與回歸）**。
   * **結果顯示 TabGNN 能夠顯著提升預測效果，相較於 AutoFE 在所有數據集上都能取得進步**。

**重點歸納**

1. **背景介紹：**
   * **表格數據預測（TDP）** 是工業界最常見的應用之一，涵蓋詐欺檢測、銷售預測、推薦系統等領域。
   * **現有方法（LR, GBM, DeepFM, AutoFE）主要關注特徵交互，忽略了樣本之間的關係。**
2. **關鍵挑戰：**
   * **特徵交互不足以捕捉樣本之間的潛在聯繫。**
   * **樣本關係可能涉及多種維度（教育、地理、社交）。**
   * **現有方法無法有效建模這些樣本關係。**
3. **本文貢獻（TabGNN）：**
   * **透過構造「多重圖（Multiplex Graph）」來建模多種樣本關係。**
   * **設計「多重圖神經網絡（Multiplex GNN）」來學習更好的樣本表示。**
   * **與 AutoFE 整合，結合特徵交互與樣本關係，提升預測表現。**
   * **在 11 個數據集上的實驗結果表明 TabGNN 具有穩定的增益效果。**

**2. 相關工作（Related Works）翻譯與重點歸納**

**2.1 現有的表格數據預測方法（Existing Methods for Tabular Data Prediction）**

**翻譯：**

在工業界，**表格數據學習（TDP）** 是最受歡迎的機器學習應用之一【16】。為了提升預測效能，許多研究提出了各種 **特徵交互建模（Feature Interaction Modeling）** 方法，這些方法大致可以分為以下幾類：

1. **線性模型（Linear Models）**
   * 代表方法：
     + **邏輯回歸（Logistic Regression, LR）【4】**
     + **Wide&Deep【7】**
   * 這些模型依賴於手工設計的特徵（Hand-crafted Features）。
   * 優勢：**簡單易解釋**。
   * 缺點：**高度依賴領域知識，泛化能力有限**。
2. **自動特徵交互方法（Automated Feature Interaction）**
   * 代表方法：
     + **DFS（Deep Feature Synthesis）【17】**
     + **AutoCross【23】**
     + **AutoFIS【21】**
     + **AutoFeature【19】**
   * 這類方法 **自動生成交互特徵**，避免人工設計特徵的困難。
   * 主要策略：
     + **DFS**：自動組合基礎特徵。
     + **AutoCross、AutoFIS**：根據數據學習最優特徵交互方式。
   * 優勢：**能夠適應不同領域，減少人工干預**。
   * 缺點：**仍然忽略了樣本之間的關係**。
3. **基於因子分解（Factorization-Based Methods）**
   * 代表方法：
     + **因子分解機（Factorization Machine, FM）【27】**
     + **DeepFM【13】**
     + **PIN【26】**
   * 這些方法學習特徵之間的交互關係，並能夠捕捉 **高階交互特徵**。
   * 優勢：**比傳統線性模型更強大，能夠自動發現有用的特徵組合**。
   * 缺點：**仍然沒有考慮樣本之間的潛在關聯**。
4. **基於梯度提升（Gradient Boosting-Based Methods）**
   * 代表方法：
     + **GBDT（Gradient Boosting Decision Tree）【10】**
     + **XGBoost【6】**
   * 這些方法能夠從原始數據中學習有效特徵，並廣泛應用於各種表格數據預測任務。
   * 優勢：**在工業界表現極為優秀，是目前最受歡迎的 TDP 方法之一**。
   * 缺點：**仍然局限於特徵交互，沒有考慮樣本之間的關係**。

**AutoFE**

作為 AI 技術提供商，**4Paradigm 公司** 為來自各個領域的企業客戶提供 TDP 解決方案，其中 **AutoFE（Automatic Feature Engine）** 是核心產品之一：

* **AutoFE 主要依賴 LR 和 GBM** 進行預測。
* **特徵增強模組（Feature Engineering Modules）** 主要基於特徵交互。
* **雖然 AutoFE 已在許多場景下取得良好效果，但仍然缺乏樣本關係建模。**

**重點歸納**

1. **現有方法大多關注特徵交互（Feature Interactions），但忽略了樣本關係（Sample Relations）。**
2. **AutoFE 代表了工業界的主流 TDP 解決方案，但仍然可以進一步改進。**

**2.2 多重圖神經網絡（Multiplex Graph Neural Networks）**

**翻譯：**

**多重圖（Multiplex Graph）** 最初被設計用於 **社會學（Sociology）** 研究，以建模人與人之間的多層關係【29】。其核心概念是：

* **一個節點可以同時具有多種類型的邊（關係）**，例如：
  + **朋友關係（Friendship）**
  + **工作關係（Workplace Relationship）**
  + **共同興趣（Shared Interests）**

**多重圖的應用** 近年來，多重圖被廣泛應用於各種領域，包括：

* **網絡嵌入（Network Embedding）【24, 32】**
* **推薦系統（Recommendation Systems）【3, 9, 34】**
* **分子結構分析（Molecular Analysis）【33】**
* **金融詐欺檢測（Financial Anti-Fraud Detection）【15】**
* **邏輯推理（Diagrammatic Reasoning）【30】**

這些方法大多基於 **圖神經網絡（Graph Neural Networks, GNN）**，例如：

* **圖卷積網絡（Graph Convolutional Networks, GCN）【20】**
* **GraphSAGE【14】**
* **圖注意力網絡（Graph Attention Networks, GAT）【28】**
* **圖同構網絡（Graph Isomorphism Networks, GIN）【31】**

然而，這些研究通常要求樣本之間 **存在明確的關係（顯式關係）**，例如 **朋友關係、交易關係**。但在許多表格數據場景中，**樣本關係並不明顯，無法直接獲取**。

**TabGNN 的創新**

與以往的方法不同，本研究提出 **TabGNN**，它的核心創新點包括：

1. **適用於表格數據的多重圖建模**
   * 設計了一種基於啟發式（Heuristic-based）的方式來構建多重圖，而不需要顯式關係。
2. **多重圖神經網絡（Multiplex GNN）**
   * 設計了一種新的 **多重 GNN**，能夠從不同類型的樣本關係中學習更好的特徵表示。
3. **與現有 TDP 方法整合**
   * TabGNN 不會取代 AutoFE，而是將其學習到的嵌入與 AutoFE 提取的特徵結合，以提升最終預測效果。

**重點歸納**

1. **多重圖最初來自社會學研究，但近年來被應用於推薦系統、反詐欺、分子分析等領域。**
2. **傳統多重圖方法通常依賴顯式關係，而表格數據中的樣本關係可能是隱式的，需要新的建模方法。**
3. **TabGNN 提出了一種適用於表格數據的新穎多重圖建模方式，並設計了相應的 GNN 架構。**
4. **TabGNN 能夠與 AutoFE 互補，提升最終預測性能。**

**3. 框架（Framework）翻譯與重點歸納**

本章節主要介紹 **TabGNN** 的技術架構，包括：

1. **問題定義（Problem Formulation）**
2. **多重圖構建（Multiplex Graph Construction）**
3. **多重圖神經網絡（Multiplex Graph Neural Network）**
4. **與 AutoFE 的整合（Integration with AutoFE）**
5. **討論（Discussion）**

**3.1 問題定義（Problem Formulation）**

**翻譯**

對於一個 **表格數據集**，假設其包含 **NNN 行（樣本）** 和 **ddd 列（特徵）**：

* 每一行 **xxx** 代表一個樣本。
* 每一列代表該樣本的某個屬性（數值或類別特徵）。
* 每個樣本 **xxx** 的特徵向量表示為 **x∈Rdx \in \mathbb{R}^dx∈Rd**。
* 目標變量（標籤）集合為 **YYY**，其中 **yx∈Yy\_x \in Yyx​∈Y** 表示樣本 **xxx** 的標籤。

在 TDP（Tabular Data Prediction）中，我們主要關心兩類任務：

1. **分類（Classification）：** 預測某個樣本的類別，例如：
   * **金融風險評估：** 預測用戶是否會違約（是否能夠準時還款）。
   * **推薦系統：** 預測用戶是否會點擊某則廣告。
2. **回歸（Regression）：** 預測某個樣本的數值，例如：
   * **銷售預測：** 預測某商品的銷售額。

**多重圖（Multiplex Graph）的定義**

我們引入 **多重圖（Multiplex Graph）** 的概念，以更好地建模樣本之間的多種關聯。

**定義 1. 多重圖（Multiplex Graph）**

* 一個 **多重圖** 定義為： G=(V,E1,E2,...,ER)G = (V, E\_1, E\_2, ..., E\_R)G=(V,E1​,E2​,...,ER​) 其中：
  + **VVV**：節點集合（對應於數據集中的樣本）。
  + **ErE\_rEr​**：第 rrr 種關係類型的邊集合。
  + **Gr=(V,Er)G\_r = (V, E\_r)Gr​=(V,Er​)**：第 rrr 層的圖（也稱為 Plex）。

換句話說，**一個多重圖由多個不同類型的關係圖 G1,G2,...,GRG\_1, G\_2, ..., G\_RG1​,G2​,...,GR​ 組成**，可以描述樣本之間的多重關聯，例如：

* **相同年齡的樣本之間的關聯（Age-based edges）**
* **相同城市的樣本之間的關聯（City-based edges）**

**問題 1. 基於多重圖的表格數據預測**

\*\*給定一個表格數據集，對於每個樣本 xxx：

* 構造一個多重圖 Gx=(Vx,E1x,...,ERx)G\_x = (V\_x, E\_1^x, ..., E\_R^x)Gx​=(Vx​,E1x​,...,ERx​) 來表示與 xxx 相關的樣本關係。
* 使用 xxx 及其多重圖 GxG\_xGx​ 來預測其標籤 yxy\_xyx​。\*\*

**3.2 多重圖構建（Multiplex Graph Construction）**

**翻譯**

在這一部分，我們探討如何構造 **多重圖（Multiplex Graph）**，以捕捉表格數據中的 **多層樣本關係**。主要挑戰有兩個：

1. **如何提取不同類型的關係（Relation Extraction）**
2. **如何處理時間約束（Temporal Constraint）**

**(1) 關係提取（Relation Extraction）**

與以往研究不同，**在表格數據中，樣本關係並不是直接給定的**，我們需要基於數據特徵來構造關係。根據我們在 **4Paradigm 公司的實踐經驗**，我們歸納出幾種常見的啟發式規則：

1. **相同 ID 特徵（Shared ID Features）**
   * 例如，在 **廣告點擊預測（CTR Prediction）** 中，**相同用戶 ID 的樣本應該被連接**。
2. **高重要性的類別特徵（Important Categorical Features）**
   * 例如，在 **貸款風險評估（Loan Default Analysis）** 中，**相同教育水平的用戶可能具有相似的還款能力**，因此應該連接。
3. **數值特徵的離散化（Discretized Numerical Features）**
   * 例如，在貸款應用中，**將年齡分組（如 25-30 歲）後，具有相似年齡的用戶可以建立連接**。
4. **類別特徵的交叉（Cross Product of Categorical Features）**
   * 例如，在貸款場景中，**同時擁有相似年齡和相同教育水平的用戶，應該建立更強的關聯**。

這些方法幫助我們為表格數據構造多種樣本關係，並為後續的 **多重 GNN 訓練提供基礎**。

**(2) 時間約束（Temporal Constraint）**

在許多真實世界應用中，數據具有時間性，例如：

* **貸款申請數據** 會有「申請時間」。
* **銷售預測數據** 會有「交易時間」。

為了避免 **數據洩漏（Data Leakage）**，我們需要確保：

* **先前的樣本不能「看到」未來樣本的信息**。
* 我們使用 **有向圖（Directed Graph）** 來連接樣本：
  + **早期的樣本指向晚期的樣本**，而不是相反。

如果數據 **沒有時間約束（如靜態數據）**，則可以使用 **無向圖（Undirected Graph）** 來表示樣本關係。

**3.3 多重圖神經網絡（Multiplex Graph Neural Network）**

我們設計了一種新的 **多重 GNN（Multiplex GNN）**，主要包括四個步驟：

1. **特徵編碼（Feature Encoding）**
2. **層內聚合（Intra-Layer Aggregation）**
3. **層間聚合（Inter-Layer Aggregation）**
4. **模型訓練（Model Training）**

這部分涉及數學公式，下一章節會詳細解釋。

**3.4 與 AutoFE 的整合（Integration with AutoFE）**

我們的方法不會取代 AutoFE，而是 **增強 AutoFE**：

1. **TabGNN 負責學習樣本之間的關係嵌入（Graph Embeddings）。**
2. **將 TabGNN 的輸出與 AutoFE 提取的特徵拼接（Concatenate）。**
3. **最終使用 AutoFE 的 LR/GBM 模型進行預測。**

這種方法簡單靈活，且能夠 **兼容傳統的 TDP 工作流**。

**3.5 討論（Discussion）**

**與 CTR 預測模型的關聯**

* 在 CTR（Click-Through Rate）預測中，已有一些方法考慮了樣本關係，例如：
  + **Deep Interest Network（DIN）【35】**
  + **Behavior Sequence Transformer（BST）【5】**

然而：

* **這些方法主要基於時間序列（Sequential Behavior）。**
* **TabGNN 更通用，適用於所有表格數據場景。**

**重點歸納**

1. **透過構造「多重圖（Multiplex Graph）」來捕捉樣本之間的多種關係。**
2. **設計「多重 GNN（Multiplex GNN）」來學習更豐富的樣本表示。**
3. **TabGNN 可以無縫整合到 AutoFE，提升最終預測效果。**

**4. 實驗（Experiments）翻譯與重點歸納**

本章節主要探討 **TabGNN 在不同數據集上的表現**，包括：

1. **實驗設置（Experimental Settings）**
2. **實驗結果（Experimental Results）**
3. **消融研究（Ablation Study）**
4. **計算開銷（Computational Overhead）**

**4.1 實驗設置（Experimental Settings）**

**(1) 任務與數據集（Tasks and Datasets）**

本研究針對 **分類（Classification）** 和 **回歸（Regression）** 兩類 TDP 任務，使用了 **9 個私有數據集** 和 **2 個公開數據集**（共 11 個數據集）。

**數據集統計**

| **任務** | **數據集** | **訓練樣本** | **測試樣本** | **數值特徵數** | **類別特徵數** | **領域** | **是否有時間約束** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **分類** | Data1 | 35,581 | 8,895 | 16 | 17 | 貸款 | ✅ |
|  | Data2 | 1,888,366 | 1,119,778 | 8 | 23 | 新聞點擊 | ✅ |
|  | Data3 | 108,801 | 27,201 | 19 | 9 | 貸款 | ❌ |
|  | Data4 | 226,091 | 34,867 | 14 | 26 | 電商 | ✅ |
|  | Data5 | 435,329 | 31,076 | 8 | 34 | 電商 | ✅ |
| **回歸** | Data6 | 1,638,193 | 702,016 | 43 | 16 | 直播觀看時長 | ✅ |
|  | Data7 | 3,923,406 | 694,194 | 0 | 25 | 零售 | ✅ |
|  | Data8 | 10,512,133 | 29,879 | 4 | 17 | 零售 | ✅ |
|  | Data9 | 179,893 | 43,236 | 5 | 2 | 政府 | ✅ |
| **公開分類** | Home Credit | 307,511 | 48,744 | 175 | 51 | 貸款 | ✅ |
|  | JD | 4,992,910 | 446,763 | 6 | 17 | 電商 | ✅ |

* **數據集領域涵蓋貸款、電子商務、新聞點擊、直播觀看、零售、政府公共數據等。**
* **部分數據集（如貸款、電商）具有時間約束，因此圖結構為有向圖。**

**(2) 評估指標（Evaluation Metrics）**

* **分類任務：AUC（Area Under Curve）**
* **回歸任務：MSE（Mean Squared Error）**
  + **AUC 越高越好**
  + **MSE 越低越好**

**(3) 模型實現（Implementation Details）**

* **基線模型（Baselines）：**
  + **AutoFE（LR & GBM）**：現有的 TDP 解決方案，未考慮樣本關係。
  + **DeepFM**：流行的深度學習 TDP 方法。
* **TabGNN 的實現：**
  + **GNN 框架：DGL 0.5.2 & PyTorch 1.5.0**
  + **計算環境：**
    - **CPU：Intel Xeon Silver 4214 @ 2.20GHz**
    - **GPU：8 × NVIDIA GeForce RTX 2080TI-11GB**
    - **RAM：512GB**
  + **使用 AutoFE 進行特徵處理，再結合 TabGNN 輸出的圖嵌入進行最終預測。**

**4.2 實驗結果（Experimental Results）**

**TabGNN vs. AutoFE**

| **數據集** | **AutoFE** | **AutoFE + TabGNN** | **提升幅度** |
| --- | --- | --- | --- |
| Data1 | 0.6021 | 0.6139 | **+1.9%** |
| Data2 | 0.8662 | 0.8929 | **+3.1%** |
| Data3 | 0.9019 | 0.9139 | **+1.3%** |
| Data4 | 0.9787 | 0.9857 | **+0.7%** |
| Data5 | 0.8310 | 0.8754 | **+5.3%** |
| Data6 (MSE) | 15726.31 | 14392.52 | **-5.8%** |
| Data7 (MSE) | 10.94 | 10.05 | **-8.1%** |
| Data8 (MSE) | 20.47 | 19.93 | **-2.6%** |
| Data9 (MSE) | 196.47 | 191.24 | **-2.7%** |
| Home Credit | 0.7298 | 0.7408 | **+1.5%** |
| JD | 0.7159 | 0.7537 | **+5.3%** |

* **TabGNN 在所有 11 個數據集上都帶來顯著提升。**
* **尤其是在回歸任務（Data6, Data7）上的提升最大（約 5-8%）。**
* **在 JD 數據集上的提升也很顯著（+5.3%）。**

**TabGNN vs. DeepFM**

TabGNN 也能夠提升 **深度學習模型（DeepFM）** 的表現：

* 在 **Data7 數據集**，TabGNN **將 DeepFM 的 MSE 降低了 28.3%**！
* 在 **JD 數據集**，TabGNN **提升 DeepFM 的 AUC 7.3%**。

這表明：

* **樣本關係對於傳統 ML 方法（AutoFE）和深度學習方法（DeepFM）均有幫助。**
* **TabGNN 的方法具有通用性，能夠適用於多種預測場景。**

**4.3 消融研究（Ablation Study）**

**(1) GNN 模型比較**

| **GNN 模型** | **Data3 (AUC)** | **Home Credit (AUC)** | **Data6 (MSE)** |
| --- | --- | --- | --- |
| **GCN** | 0.8911 | 0.7372 | 14563.4 |
| **GAT** | 0.8965 | 0.7387 | 14404.2 |
| **TabGNN（最終模型）** | **0.9139** | **0.7408** | **14392.5** |

* **TabGNN 優於標準 GCN 和 GAT，證明其多重層聚合的有效性。**

**(2) 不同關係類型的影響**

| **關係類型** | **Data3 (AUC)** | **Home Credit (AUC)** | **Data6 (MSE)** |
| --- | --- | --- | --- |
| **單一關係（Relation 1）** | 0.8975 | 0.7388 | 14434.8 |
| **單一關係（Relation 2）** | 0.8979 | 0.7394 | 14411.9 |
| **兩者結合** | **0.9139** | **0.7408** | **14392.5** |

* **同時考慮多種樣本關係能夠顯著提升模型性能。**

**4.4 計算開銷（Computational Overhead）**

| 模型 | JD 訓練時間 | Data8 訓練時間 | |---|---| | **AutoFE** | ~4h | ~7h | | **AutoFE + TabGNN** | ~11h | ~20h |

* TabGNN **計算開銷約為 AutoFE 的 2-3 倍**。
* **考慮到預測性能的提升，這一額外的計算成本是可以接受的。**

**5. 結論與未來工作（Conclusion & Future Work）翻譯與重點歸納**

**5.1 結論（Conclusion）**

**翻譯**

在本研究中，我們提出了一個新穎的框架 **TabGNN**，專門用於 **表格數據預測（TDP）**，這是一個在工業界廣泛應用的機器學習任務，例如：

* **商業銀行的詐欺檢測**
* **電子商務的用戶行為預測**
* **線上銷售的業績預測**

目前的 **TDP 方法主要關注特徵交互（Feature Interactions）**，但忽略了 **樣本之間的關係（Sample Relations）**。我們的研究表明：

* **樣本關係對於 TDP 具有重要價值**，但過去的研究並未充分利用這一信息。
* **TabGNN 採用了多重圖神經網絡（Multiplex Graph Neural Networks, MGNN）來顯式建模樣本關係**，這在 TDP 領域是全新的嘗試。
* **TabGNN 的創新點包括：**
  1. **提出一種新的多重圖建模方法**，適用於表格數據場景，即使樣本關係沒有顯式給出。
  2. **設計了一個多重圖神經網絡（Multiplex GNN）來學習樣本之間的關聯性，提高預測準確度。**
  3. **與 AutoFE 無縫整合，將 GNN 提取的特徵與 AutoFE 生成的特徵結合，進一步提升性能。**

**實驗結果**

我們在 **11 個數據集（分類與回歸）** 上進行了廣泛的實驗，結果表明：

* **TabGNN 在所有數據集上都顯著提升了預測效果**，相比 AutoFE，AUC **平均提升 1.5% - 5.3%**，MSE **平均降低 2.6% - 8.1%**。
* **TabGNN 也能增強深度學習方法（如 DeepFM）**，在 JD 數據集上提升 AUC **7.3%**。
* **樣本關係越複雜，TabGNN 的提升幅度越大**，例如 Data7 數據集的 MSE 降低了 **28.3%**。

**總結**

TabGNN **不僅適用於傳統的 LR / GBM 模型，也能增強深度學習方法**，這證明了：

1. **樣本關係對於 TDP 的重要性**。
2. **多重圖神經網絡能夠有效捕捉樣本間的隱式關聯**。
3. **TabGNN 能夠輕鬆集成到現有的 TDP 工作流中，具有廣泛的應用價值。**

**5.2 未來工作（Future Work）**

**翻譯**

本研究為 TDP 任務提供了一種新穎的解決方案，但仍然有許多值得探索的方向：

1. **擴展到更多應用場景**
   * 本研究主要關注 **金融、電商、新聞點擊預測** 等場景。
   * 未來，我們希望在 **醫療數據、供應鏈管理、智能製造等領域** 進一步測試 TabGNN 的效果。
2. **更好的多重圖構建方法**
   * 目前的 **多重圖是基於啟發式規則構建的（Heuristic-based）**，這可能不適用於所有場景。
   * 我們計劃探索 **基於數據驅動的方法（Data-driven Methods）** 來自動學習樣本關係，例如：
     + **基於相似性度量（Similarity Measures）**
     + **自監督學習（Self-supervised Learning）**
     + **圖自動構建（Graph Auto-Construction）**
3. **增強 GNN 模型**
   * 目前，我們使用 **多重 GNN（Multiplex GNN）** 來建模不同類型的關係。
   * 但在某些場景下，不同的樣本關係可能具有不同的重要性，因此我們希望：
     + **探索自適應關係權重學習（Adaptive Relation Weighting）**
     + **使用 Transformer 結構來進行關係建模**
4. **降低計算成本**
   * 目前，TabGNN 的計算成本約為 **AutoFE 的 2-3 倍**。
   * 未來，我們希望：
     + **優化 GNN 計算，加快訓練與推理速度**
     + **探索輕量級圖神經網絡（Lightweight GNNs）**
     + **使用知識蒸餾（Knowledge Distillation）來減少模型大小**
5. **自動化超參數調優**
   * 目前的模型需要手動調整超參數（例如 GNN 層數、鄰居聚合策略）。
   * 未來，我們希望：
     + **使用貝葉斯優化（Bayesian Optimization）**
     + **基於強化學習（Reinforcement Learning）來自動搜索最佳超參數**。
6. **開源與部署**
   * 我們計劃將 **TabGNN 開源**，並提供：
     + **與 AutoFE 的完整集成版本**
     + **Docker 化的部署方案**
     + **更簡單的 API 以方便開發者使用**

**5.3 重點歸納**

1. **TabGNN 是一種基於多重圖神經網絡（Multiplex GNN）的 TDP 方法，能夠顯著提升預測性能。**
2. **在 11 個數據集上，TabGNN 相比 AutoFE 顯著提升 AUC（+1.5% 至 +5.3%），降低 MSE（-2.6% 至 -8.1%）。**
3. **TabGNN 可以增強傳統 ML 方法（LR, GBM）和深度學習方法（DeepFM），證明了樣本關係對於 TDP 的重要性。**
4. **未來工作重點包括：**
   * **探索更智能的圖構建方法（自監督學習、自動構圖）。**
   * **提升 GNN 模型（自適應關係建模、輕量級 GNN）。**
   * **優化計算成本（加速訓練、知識蒸餾）。**
   * **開源與實際部署（提供 API & Docker 支持）。**

**總結**

這篇論文的核心思想是：  
**「TDP 任務不僅需要考慮特徵交互，還應該關注樣本之間的關係，並利用 GNN 來提升預測效果。」**  
TabGNN 提供了一種新穎的方法來解決這個問題，並在各種應用場景中證明了其有效性。