

TabNet 論文詳細解析

介紹部分

TabNet 是一種專為表格數據設計的創新深度學習架構。介紹部分主要闡述了以下幾個重要觀點：

表格數據學習的差距

深度神經網路在圖像、文本和音頻數據上已經取得了顯著成功，但表格數據卻缺乏一種能有效將原始數據編碼為有意義表示的標準化架構。儘管表格數據在現實世界中極為常見，但針對表格數據的深度學習研究仍然不足，決策樹基礎的方法在大多數應用中仍然占主導地位。

現有方法的問題

介紹解釋了為什麼基於樹的方法目前主導表格數據應用：

1. 決策樹對於具有近似超平面邊界的決策流形表現高效（表格數據中常見）
2. 決策樹在基本形式上具有高度可解釋性
3. 決策樹訓練速度快

而傳統的深度神經網路不適合表格數據，因為它們過度參數化，缺乏為表格決策流形找到最優解所需的適當歸納偏置。

表格數據深度學習的動機

作者指出使用深度學習處理表格數據的幾個潛在優勢：

1. 性能提升，特別是對於大型數據集
2. 端到端學習，使得以下操作成為可能：
 - 高效編碼多種數據類型（如圖像與表格數據）
 - 減少特徵工程的需求
 - 從流數據中學習
 - 表示學習，支持數據高效的領域適應、生成建模和半監督學習等應用

主要貢獻

論文概述了四項主要貢獻：

1. **TabNet** 輸入原始表格數據無需預處理，並使用基於梯度下降的優化進行訓練，實現靈活地融入端到端學習
2. **TabNet** 使用序列注意力機制選擇在每個決策步驟中推理的特徵，實現可解釋性和更好的學習，因為學習能力用於最顯著的特徵
3. **TabNet** 的設計導致兩個有價值的特性：(i) 在各種分類和回歸問題數據集上優於其他表格學習模型，(ii) 在局部和全局層面的可解釋性
4. 論文還通過無監督預訓練預測掩蔽特徵，展示了顯著的性能提升

相關工作

相關工作部分將 **TabNet** 置於現有表格數據學習研究的背景下，分為幾個關鍵領域：

特徵選擇

特徵選擇是指基於特徵對預測的用處，精心挑選特徵子集。傳統方法如前向選擇和 **Lasso** 正則化是基於整個訓練數據（全局方法）賦予特徵重要性。

相反，實例級特徵選擇是指為每個輸入單獨挑選特徵。之前的工作使用單獨的解釋器模型或 **actor-critic** 框架探索這一點。不同於這些方法，**TabNet** 在端到端學習中採用具有可控稀疏性的軟特徵選擇 - 單一模型同時執行特徵選擇和輸出映射，以更緊湊的表示實現更優的性能。

基於樹的學習

決策樹常用於表格數據，因為它們能高效選擇具有最大統計信息增益的全局特徵。為提高性能，集成方法如隨機森林使用數據的隨機子集和隨機選擇的特徵來生成多棵樹。**XGBoost** 和 **LightGBM** 是最近主導大多數數據科學競賽的集成決策樹方法。

深度神經網絡與決策樹的整合

幾種方法試圖將神經網絡與決策樹結合：

- 用 **DNN** 構建模塊表示 **DT** 會導致表示冗餘和學習效率低

- 軟（神經）決策樹使用可微分決策函數取代不可微分的軸對齊分割，但往往失去自動特徵選擇
- 其他方法提出軟分箱函數，利用表現力豐富的特徵組合，或從原始模塊自適應地增長網絡

TabNet 與這些方法的不同之處在於通過序列注意力嵌入了具有可控稀疏性的軟特徵選擇。

自監督學習

無監督表示學習改進了監督學習，特別是在數據有限的情況下。近期針對文本和圖像數據的工作顯示了基於無監督學習目標（掩蔽輸入預測）和基於注意力的深度學習所帶來的顯著進步。

TabNet 表格學習方法

本節詳細解釋了 **TabNet** 的架構和功能組件：

設計理念

作者指出決策樹在學習真實世界表格數據集上表現優異。通過特定設計，傳統 **DNN** 構建模塊可用於實現類似 **DT** 的輸出流形。在這種設計中，單個特徵選擇是獲得超平面形式決策邊界的關鍵，這可以推廣為特徵的線性組合，其中係數確定每個特徵的比例。

TabNet 基於這一洞見，通過以下精心設計超越決策樹同時保持其優勢：

1. 使用從數據學習的稀疏實例級特徵選擇
2. 構建序列多步架構，每一步基於所選特徵貢獻決策的一部分
3. 通過所選特徵的非線性處理提升學習能力
4. 通過更高維度和更多步驟模擬集成

架構概述

TabNet 使用原始數值特徵，使用可訓練的嵌入映射分類特徵，不進行任何全局特徵標準化，僅應用批量標準化。相同的 **D** 維特徵傳遞給每個決策步驟，其中 **B** 是批量大小。

TabNet 的編碼基於具有 **Nsteps** 決策步驟的序列多步處理。每個步驟輸入來自

前一步驟的處理信息來決定使用哪些特徵，並輸出處理後的特徵表示以聚合到整體決策中。

特徵選擇

TabNet 的一項關鍵創新是其特徵選擇機制：

TabNet 採用可學習掩碼 $M[i]$ 對顯著特徵進行軟選擇。通過對最重要特徵的稀疏選擇，決策步驟的學習能力不會浪費在無關特徵上，使模型更加參數高效。

掩碼是乘法式的 $(M[i] \cdot f)$ ，並使用注意力轉換器基於前一步驟的處理特徵獲取掩碼：

$$M[i] = \text{sparsemax}(P[i - 1] \cdot h_i(a[i - 1]))$$

Sparsemax 標準化通過將歐幾里得投影映射到概率單純形上來鼓勵稀疏性，支持稀疏特徵選擇的可解釋性目標。

$P[i]$ 是一個先驗尺度項，表示特徵之前被使用了多少： $P[i] = \prod_{j=1}^i (\gamma - M[j])$ ，其中 γ 是鬆弛參數 - 當 $\gamma = 1$ 時，一個特徵被強制只在一個決策步驟中使用，隨著 γ 增加，提供了更多在多個決策步驟中使用特徵的靈活性。

特徵處理

TabNet 使用特徵轉換器處理過濾後的特徵，然後將輸出分為決策步驟輸出和後續步驟的信息。

為了參數高效、魯棒且高容量的學習：

- 特徵轉換器應該有在所有決策步驟間共享的層（因為相同的特徵在不同步驟間輸入）
- 它也應該包括決策步驟相關的層

每個全連接層後跟批量標準化和門控線性單元(GLU)非線性，連接到具有標準化的歸一化殘差連接，以穩定學習。

對於整體決策嵌入，受決策樹聚合的啟發，TabNet 構建： $d_{out} = \sum_{i=1}^{Nsteps} \text{ReLU}(d[i])$

可解釋性

TabNet 的特徵選擇掩碼可以揭示在每個步驟中選擇了哪些特徵。如果 $M_{b,j}[i] =$

0，則第 b 個樣本的第 j 個特徵對決策沒有貢獻。

作者提出了一個聚合特徵重要性掩碼，根據對決策的相對重要性加權不同步驟的掩碼：

$$\text{Magg-b,j} = \sum_{i=1 \text{ 到 } \text{Nsteps}} \eta_{b[i]} \text{Mb,j}[i] / \sum_{j=1 \text{ 到 } D} \sum_{i=1 \text{ 到 } \text{Nsteps}} \eta_{b[i]} \text{Mb,j}[i]$$

其中 $\eta_{b[i]}$ 表示第 b 個樣本在第 i 個決策步驟的聚合決策貢獻。

表格自監督學習

TabNet 包括一個解碼器架構，從編碼表示重建表格特徵。解碼器由特徵轉換器和每個決策步驟的全連接層組成。輸出求和得到重建的特徵。

自監督學習任務是從其他特徵預測缺失的特徵列。給定二進制掩碼 $S \in \{0, 1\}^{(B \times D)}$ ，編碼器輸入 $(1 - S) \cdot \hat{f}$ ，解碼器輸出重建的特徵 $S \cdot \hat{f}$ 。

這允許有效的預訓練，在標記數據有限時能顯著提高性能。

實驗

實驗部分在各種數據集上評估 TabNet 並與其他方法進行比較：

實例級特徵選擇

論文首先在 6 個合成數據集上評估 TabNet，這些數據集中只有特徵的子集決定輸出。對於數據集 Syn1-Syn3，顯著特徵對所有實例都相同，而對於 Syn4-Syn6，顯著特徵是實例相關的。

結果顯示：

- TabNet 優於樹集成、LASSO 正則化和 L2X，與 INVASE 相當
- 對於 Syn1-Syn3，TabNet 性能接近全局特徵選擇 - 它能識別全局重要特徵
- 對於 Syn4-Syn6，通過消除實例級冗餘特徵，TabNet 改進了全局特徵選擇
- TabNet 以更少的參數(26k-31k)實現了這一點，相比其他方法(43k-101k)

真實世界數據集上的性能

論文在幾個真實世界數據集上評估了 TabNet：

1. Forest Cover Type（森林覆蓋類型）：

- TabNet 達到 96.99%測試精度，優於 XGBoost(89.34%)、LightGBM(89.28%)、CatBoost(85.14%)，甚至 AutoML Tables(94.95%)

2. Poker Hand（撲克牌手牌）：

- 在常規 DNN、DT 和混合變體難以處理的不平衡數據和學習排序/排名操作的任務上，TabNet 達到 99.2%的測試精度
- 這接近手工規則的 100%精度，顯著優於 XGBoost(71.1%)、LightGBM(70.0%)等方法

3. Sarcos（機器人逆動力學）：

- 在這個人類體型機器人臂的逆動力學回歸任務上，當模型大小不受限制時，TabNet 的測試 MSE(0.14)比其他方法低近一個數量級
- 即使是小模型大小，TabNet 的性能也與具有 100 倍參數的最佳模型相當

4. Higgs Boson（希格斯玻色子）：

- 在這個大型數據集(10.5M 實例)上，TabNet 以更緊湊的表示優於 MLP，並達到與相同參數數量的最先進稀疏進化訓練相似的性能

5. Rossmann Store Sales（羅斯曼商店銷售）：

- 對於這個預測任務，TabNet 優於常用方法，包括 XGBoost、LightGBM、CatBoost 和 MLP

可解釋性

實驗展示了 TabNet 的可解釋性能力：

1. 合成數據集：

- 論文展示了 TabNet 的聚合特徵重要性掩碼如何清晰識別合成數據集的相關特徵
- 對於 Syn2，無關特徵的掩碼幾乎為零
- 對於 Syn4，TabNet 準確執行實例級特徵選擇，專注於指示器特

徵和相關特徵組，同時為無關特徵分配接近零的權重

2. 真實世界數據集:

- 對於蘑菇可食性預測，TabNet 正確識別"氣味"作為最具辨別力的特徵，賦予它 43%的重要性，而其他可解釋性方法不到 30%
- 對於 Adult Census Income（成人人口收入），TabNet 的特徵重要性排名與眾所周知的排名一致，正確識別"年齡"為最重要特徵

自監督學習

論文證明了無監督預訓練顯著提高了監督分類任務的性能，特別是當未標記數據集遠大於已標記數據集時。

例如：

- 僅使用 Higgs 數據集的 1k 訓練示例，使用預訓練將測試精度從 57.47% 提高到 61.37%
- 使用 10k 示例，精度從 66.66%提高到 68.06%
- 即使是 100k 示例，仍有從 72.92%提高到 73.19%的改進

此外，使用無監督預訓練的模型收斂速度更快，這對於持續學習和領域適應非常有價值。

結論

結論總結了 TabNet 作為一種新穎的表格學習深度學習架構，該架構使用序列注意力在每個決策步驟選擇一個語義上有意義的特徵子集。

關鍵要點包括：

1. 實例級特徵選擇使得學習高效，因為模型容量完全用於最顯著的特徵
2. 這種方法也通過選擇掩碼的可視化產生更可解釋的決策
3. TabNet 在不同領域的表格數據集上優於之前的工作
4. 論文展示了無監督預訓練對快速適應和性能提升的顯著益處

總體而言，TabNet 解決了傳統神經網絡和決策樹模型在表格數據上的關鍵限制，提供了結合兩種範式優勢的平衡方法，同時引入了專為表格數據量身定制

的序列特徵注意力和自監督學習等新元素。