**詳解 TabTransformer: 使用上下文嵌入的表格數據建模**

我將針對論文中的主要部分進行詳細解釋，並提供例子幫助理解。

**1. Introduction**

論文的引言部分主要介紹了以下幾個關鍵點：

**1.1 表格數據的重要性和現有模型限制**

表格數據是現實世界中最常見的數據類型，廣泛應用於推薦系統、在線廣告和投資組合優化等領域。目前處理表格數據的最先進方法是基於樹的集成方法，如梯度提升決策樹(GBDT)，這些模型具有較高的預測準確性、訓練速度快且易於解釋。

然而，基於樹的模型存在幾個局限性：

1. 不適合從流數據中持續訓練
2. 不允許在存在多模態數據的情況下進行端到端學習
3. 基本形式不適合最先進的半監督學習方法
4. 無法應用最先進的深度學習方法來處理缺失和噪聲數據

**1.2 多層感知機(MLP)及其缺點**

MLP是一種使用梯度下降訓練的經典模型，允許端到端學習。然而，由於其淺層架構和上下文無關的嵌入，MLP存在以下限制：

* 模型和學習的嵌入不可解釋
* 對缺失和噪聲數據不夠穩健
* 半監督學習性能不佳
* 性能無法匹配GBDT等基於樹的模型

**1.3 TabTransformer的提出**

作者提出TabTransformer來解決MLP和現有深度學習模型的局限性，同時縮小MLP和GBDT之間的性能差距。這是基於Transformer在NLP領域取得的巨大成功。

在NLP領域，從早期的Word2Vec的上下文無關詞嵌入到BERT提供的上下文詞元嵌入，嵌入方法被廣泛研究和應用。相比上下文無關嵌入，基於上下文嵌入的模型在NLP中取得了巨大成功。特別是基於自注意力的Transformer已成為NLP模型的標準組件，以達到最先進的性能。

**1.4 主要貢獻**

作者在論文中提出了四個主要貢獻：

1. 提出TabTransformer架構，該架構提供並利用分類特徵的上下文嵌入。通過大量實驗證明TabTransformer優於基線MLP和最近的表格數據深度網絡，同時匹配基於樹的集成模型(GBDT)的性能。
2. 研究了生成的上下文嵌入，並強調其可解釋性，與現有藝術達到的參數化上下文無關嵌入形成對比。
3. 證明了TabTransformer對噪聲和缺失數據的穩健性。
4. 提供並廣泛研究了表格數據的兩階段預訓練和微調程序，超越了半監督學習方法的最先進性能。

**2. The TabTransformer**

TabTransformer的核心架構包含三個主要組件：

**2.1 整體架構**

TabTransformer架構包括：

1. 列嵌入層(column embedding layer)
2. N個Transformer層的堆疊
3. 多層感知機(MLP)

如論文圖1所示，整體架構將分類特徵轉換為上下文嵌入，然後與連續特徵一起輸入到MLP中進行最終預測。

**2.2 數學公式和模型處理流程**

假設(x,y)表示特徵-目標對，其中x≡{xcat, xcont}。xcat表示所有分類特徵，xcont∈Rc表示所有c個連續特徵。令xcat≡{x1, x2, ..., xm}，每個xi是一個分類特徵。

每個xi分類特徵通過列嵌入轉換為維度為d的參數化嵌入。令eφi(xi)∈Rd表示xi特徵的嵌入，Eφ(xcat)={eφ1(x1), ..., eφm(xm)}表示所有分類特徵的嵌入集合。

這些參數化嵌入Eφ(xcat)被輸入到第一個Transformer層。第一個Transformer層的輸出被輸入到第二個Transformer層，依此類推。每個參數化嵌入通過從其他嵌入中連續聚合上下文，在輸出自頂層Transformer時被轉換為上下文嵌入。

我們將Transformer層序列表示為函數fθ。函數fθ對參數化嵌入{eφ1(x1), ..., eφm(xm)}進行操作，並返回相應的上下文嵌入{h1, ..., hm}，其中hi∈Rd。

上下文嵌入{h1, ..., hm}與連續特徵xcont連接，形成維度為(d×m+c)的向量。這個向量被輸入到MLP(表示為gψ)以預測目標y。

損失函數L(x,y)定義為： L(x,y) ≡ H(gψ(fθ(Eφ(xcat)), xcont), y)

其中H為分類任務的交叉熵或回歸任務的均方誤差。通過一階梯度方法進行端到端學習，優化TabTransformer的所有參數，包括列嵌入的φ、Transformer層的θ和頂層MLP的ψ。

**2.3 Transformer層**

Transformer由多頭自注意力層和位置前饋層組成，每層之後進行元素級加法和層正規化。

自注意力層包括三個參數矩陣：Key、Query和Value。每個輸入嵌入都投影到這些矩陣上，生成它們的key、query和value向量。

形式上，設K∈Rm×k, Q∈Rm×k和V∈Rm×v分別為所有嵌入的key、query和value向量組成的矩陣，m為輸入到Transformer的嵌入數量，k和v分別為key和value向量的維度。每個輸入嵌入通過注意力頭關注所有其他嵌入：

Attention(K, Q, V) = A · V

其中A = softmax((QKT)/√k)。

對於每個嵌入，注意力矩陣A∈Rm×m計算它對其他嵌入的關注程度，從而將嵌入轉換為上下文嵌入。注意力頭的輸出(維度為v)通過全連接層投影回嵌入維度d，然後通過兩個位置前饋層傳遞。第一層將嵌入擴展到其大小的四倍，第二層將其投影回其原始大小。

**2.4 列嵌入(Column Embedding)**

對於每個分類特徵(列)i，有一個嵌入查找表eφi(.)，i∈{1,2,...,m}。對於具有di個類的第i個特徵，嵌入表eφi(.)有(di+1)個嵌入，其中額外的嵌入對應於缺失值。

編碼值xi = j∈[0,1,2,..,di]的嵌入是eφi(j) = [cφi, wφij]，其中cφi∈Rℓ，wφij∈Rd-ℓ。cφi的維度ℓ是一個超參數。

唯一標識符cφi∈Rℓ將列i中的類與其他列中的類區分開。這種唯一標識符的使用是新的，特別針對表格數據設計。在語言建模中，嵌入是與句子中單詞的位置編碼元素級相加。由於在表格數據中沒有特徵的排序，我們不使用位置編碼。

**2.5 嵌入的預訓練**

上述的上下文嵌入是在使用標記樣本的端到端監督訓練中學習的。對於只有少量標記樣本和大量未標記樣本的場景，作者引入了一個預訓練程序，使用未標記數據訓練Transformer層。

預訓練之後，使用標記數據對預訓練的Transformer層和頂層MLP層進行微調。對於微調，使用方程式(1)中定義的監督損失。

作者探索了兩種不同的預訓練程序：

1. 掩碼語言建模(MLM)：給定輸入xcat={x1,x2,...,xm}，MLM隨機選擇從索引1到m的k%特徵並將其掩碼為缺失。Transformer層與列嵌入一起通過最小化多類分類器的交叉熵損失進行訓練，該分類器嘗試從頂層Transformer輸出的上下文嵌入中預測掩碼特徵的原始特徵。
2. 替換令牌檢測(RTD)：不是掩碼特徵，RTD將原始特徵替換為該特徵的隨機值。這裡，損失是為二元分類器最小化的，該分類器嘗試預測特徵是否已被替換。

兩種預訓練方法分別命名為TabTransformer-MLM和TabTransformer-RTD。

**3. 實驗**

作者通過一系列實驗證明了TabTransformer的效果：

**3.1 數據和設置**

實驗使用了15個公開可用的二元分類數據集，來自UCI存儲庫、AutoML挑戰賽和Kaggle，用於監督和半監督學習。

每個數據集分為五個交叉驗證拆分。每個拆分的訓練/驗證/測試數據比例為65/15/20%。數據集中的分類特徵數量範圍從2到136不等。

半監督實驗中，對於每個數據集和拆分，訓練數據中的前p個觀察被標記為標記數據，剩餘的訓練數據作為未標記集。p值選擇為50、200和500，對應於3種不同的場景。

模型超參數設置為：對於TabTransformer，隱藏(嵌入)維度、層數和注意力頭數分別固定為32、6和8。MLP層大小設置為{4×l, 2×l}，其中l是其輸入的大小。

**3.2 Transformer層的有效性**

作者首先比較了TabTransformer和沒有Transformer層的基線MLP：

從架構中移除Transformer層fθ，固定其餘組件，並與原始TabTransformer進行比較。沒有基於注意力的Transformer層的模型等效於MLP。兩個模型的分類特徵嵌入維度d都設置為32。

結果顯示，具有Transformer層的TabTransformer在15個數據集中的14個上優於基線MLP，平均AUC提升1.0%。

作者還使用t-SNE可視化了不同層Transformer產生的上下文嵌入：

對於bank marketing數據集，可以看到語義相似的類在嵌入空間中彼此接近，形成集群。例如，所有基於客戶的特徵(如職業、教育水平和婚姻狀況)位於中心附近，而非客戶基於的特徵(如月份、星期幾)位於中心區域外；在底部集群中，擁有住房貸款的嵌入與違約的嵌入接近；在左側集群中，學生的嵌入、單身的婚姻狀況、沒有住房貸款和高等教育水平聚集在一起；而在右側集群中，教育水平與職業類型密切相關。

對比來看，MLP中的上下文無關嵌入不顯示這種模式，許多語義上不相似的分類特徵被分組在一起。

**3.3 TabTransformer的穩健性**

作者通過對噪聲數據和缺失值數據的實驗證明了TabTransformer的穩健性：

**噪聲數據**

在測試樣本上，首先通過用從相應列(特徵)隨機生成的值替換一定數量的值來污染數據。然後將噪聲數據輸入到訓練好的TabTransformer以計算預測AUC分數。

結果顯示，隨著噪聲率的增加，TabTransformer在預測準確性方面表現更好，因此比MLP更穩健。特別是在Blastchar數據集中，沒有噪聲時性能幾乎相同，但隨著噪聲的增加，TabTransformer的性能顯著優於基線。

作者推測，這種穩健性來自嵌入的上下文特性。儘管一個特徵有噪聲，它從正確的特徵中獲取信息，允許一定程度的糾正。

**缺失值數據**

類似地，在測試數據上人為選擇一定數量的值為缺失，並將具有缺失值的數據發送到訓練好的TabTransformer以計算預測分數。處理缺失值嵌入有兩個選項：

1. 使用相應列中所有類的平均學習嵌入
2. 缺失值類的嵌入，即每列提到的額外嵌入

由於基準數據集中沒有足夠的缺失值來有效訓練選項(2)中的嵌入，使用選項(1)中的平均嵌入進行插補。結果表明，TabTransformer在處理缺失值方面表現出比MLP更好的穩定性。

**3.4 監督學習**

作者將TabTransformer與四類方法進行了比較：

* 邏輯回歸和GBDT
* MLP和稀疏MLP
* TabNet模型
* 變分信息瓶頸模型(VIB)

結果顯示TabTransformer、MLP和GBDT是表現最好的前3名。TabTransformer比基線MLP平均提高1.0%，性能與GBDT相當。此外，TabTransformer明顯優於TabNet和VIB這些最近的表格數據深度網絡。

**3.5 半監督學習**

作者評估了TabTransformer在半監督學習場景中的表現：

具體來說，作者將預訓練後微調的TabTransformer-RTD/MLM與以下半監督模型進行比較：

* 熵正則化(ER)與MLP和TabTransformer相結合
* 偽標記(PL)與MLP、TabTransformer和GBDT相結合
* MLP(DAE)：為表格數據上的深度模型設計的無監督預訓練方法：交換噪聲去噪自編碼器

結果分為兩組數據集呈現：超過30K數據點的6個數據集和剩餘的9個數據集。

當未標記數據量較大時，TabTransformer-RTD和TabTransformer-MLM明顯優於所有其他競爭對手。特別是，TabTransformer-RTD/MLM在50、200和500個標記數據點的情況下，平均AUC分別至少提高了1.2%、2.0%和2.1%。

基於Transformer的半監督學習方法TabTransformer(ER)和TabTransformer(PL)以及基於樹的半監督學習方法GBDT(PL)的表現比所有模型的平均水平差。

當未標記數據量變小時，TabTransformer-RTD仍然優於大多數競爭對手，但改進幅度較小。

此外，當未標記數據量較小時，TabTransformer-RTD的表現優於TabTransformer-MLM，這要歸功於其更簡單的預訓練任務(二元分類)，而不是MLM的任務(多類分類)。這與ELECTRA論文的發現一致。

**4. 相關工作**

論文的相關工作部分主要討論了兩類現有研究：監督學習和半監督學習。

**4.1 監督學習**

標準MLP已經應用於表格數據多年。針對表格數據專門設計的深度模型包括深度版本的因子分解機、基於Transformer的方法和基於決策樹的算法的深度版本。

特別是，

* Song等人(2019)將一層多頭注意力應用於嵌入以學習高階特徵
* Li等人(2020)使用自注意力層並跟踪注意力分數以獲得特徵重要性分數
* Sun等人(2019)將因子分解機模型與transformer機制結合

這三篇論文都專注於推薦系統，使得與本文進行明確比較變得困難。其他模型是圍繞表格數據的假定特性設計的，如低階和稀疏特徵交互。這些包括Deep & Cross Networks、Wide & Deep Networks、TabNets和AdaNet。

**4.2 半監督學習**

Izmailov等人(2019)提出了一種基於密度估計的半監督方法，並在表格數據上評估了他們的方法。偽標記(Lee 2013)是一種簡單、高效且流行的基線方法。

偽標記使用當前網絡推斷未標記樣本的偽標記，通過選擇最有信心的類。這些偽標記在交叉熵損失中被視為人類提供的標記。

標記傳播(Zhu and Ghahramani 2002; Iscen等人 2019)是一種類似的方法，其中節點的標記根據它們的接近程度傳播到所有節點，並被訓練模型用作真實標記。

半監督學習的另一種標準方法是熵正則化(Grandvalet and Bengio 2005; Sajjadi, Javanmardi, and Tasdizen 2016)。它將未標記樣本的平均每樣本熵添加到標記樣本的原始損失函數中。

半監督學習的另一種經典方法是協同訓練(Nigam and Ghani 2000)。然而，最近的方法——熵正則化和偽標記——通常更好且更流行。

**5. 結論**

作者提出了TabTransformer，一種用於監督和半監督學習的新型深度表格數據建模架構。作者提供了廣泛的實證證據，表明TabTransformer明顯優於MLP和最近的表格數據深度網絡，同時與基於樹的集成模型(GBDT)的性能相匹配。

作者提供並廣泛研究了表格數據的兩階段預訓練和微調程序，超越了半監督學習方法的最先進性能。TabTransformer對噪聲和缺失數據表現出良好的穩健性，上下文嵌入的可解釋性也很有前景。

**實例解釋**

為了更好地理解TabTransformer，我將用一個簡單的例子來解釋：

假設我們有一個銀行客戶數據表，包含以下特徵：

* 分類特徵：職業(engineer, teacher, doctor等)、教育水平(high school, college, graduate)、婚姻狀況(single, married)
* 連續特徵：年齡、收入、資產

傳統的MLP會將每個分類特徵編碼為獨立的嵌入向量，這些嵌入之間沒有交互。例如，"engineer"和"doctor"的嵌入是獨立學習的。

而TabTransformer會做以下處理：

1. 首先通過列嵌入將每個分類值轉換為初始嵌入向量
2. 這些嵌入向量通過Transformer層處理，每個嵌入通過自注意力機制"看到"其他所有嵌入
3. 例如，"engineer"的嵌入會受到同一行中其他特徵值的影響，如教育水平和婚姻狀況
4. 這創建了上下文敏感的嵌入，捕捉特徵間的關係

在實際應用中，這意味著模型能夠理解"engineering + graduate degree"和"doctor + graduate degree"之間的相關性，因為這些特徵組合在數據中可能表現出類似的模式。

當面對噪聲數據時(例如某些職業值被錯誤輸入)，TabTransformer可以利用其他特徵的上下文來"糾正"或減輕這種噪聲的影響，因為真實的模式通常會在多個特徵中表現出來。

這就是為什麼TabTransformer在實驗中表現出比標準MLP更好的穩健性，並且在半監督場景中特別有效，因為它能夠從未標記數據中學習特徵間的關係。