Attention 機構を有する FT-Transformer の精度向上と効率化に関する研究

1X19C015-7　磯村時将
指導教員　後藤正幸

1. 研究背景と目的

近年，表形式データに対する機械学習モデルとしては，ディープラーニング (以下，DL) モデルよりも勾配ブースティング (以下，GBDT) モデルが主流となりつつある．これは，多くの事例から GBDT モデルの方が表形式データに対して安定して高い精度を示すことが示されてきたためである．一方, DL モデルは画像分類や自然言語処理などの非構造化データに対して高い性能を示すことで知られ, 様々なデータに対する DL モデルの適用に関する研究が盛んに行われている．特に，DL モデルの一つである Transformer[1] は自然言語や画像の分野で高い性能を示すモデルとして注目されている．Transformer は，全特徴量（単語やパッチ画像など）と共に，Attention 機構によって全特徴量の間の高度な関係性を考慮することにより高い性能を示す．

一方，自然言語や画像の範囲を超えて，様々なタスクに対して Transformer を適用する研究が進められている．特に，表形式データに対しては FT-Transformer[2](以下，FTT) が提案されている．FTT は単純に，Transformer の構造に表形式データを当てはめるための工夫がされた手法であるが，GBDT モデルよりも高い計算精度を示す可能性がある．しかし，FTT が表形式データに対しても文書や画像データと同様に全ての特徴量同士の関係性を考慮している点は改善の余地があると考えられる．実際，文書や画像データと比較して，表形式データには特徴量同士の関係性にそれほど重要な意味は含まれていない場合が多いと想定できる．

そこで，本研究では Transformer の Attention 機構において，表形式データの特徴量同士の関係性を過度に考慮しないことで精度や計算効率を改善するような，表形式データにより特化した改良型の FTT を提案する．機械学習における代表的なタスクである回帰・二値分類・多値分類のそれぞれに関する評価実験を実施し，精度・計算量・解釈性の観点から提案手法の有効性を示す．

2. 準備

2.1. Transformer の概要

Transformer[1] は，Multi-Head Self-Attention(以下，MHSA) 層と Feed Forward(以下，FFN) 層から構成される．MHSA 層では，それぞれが d 次元のベクトルで表現される全 k 個の特徴量ベクトルの集合を，h 個のヘッドに分割し，式 (1) によって特徴量間の Attention 値 $\boldsymbol{H}_i \in \mathbb{R}^{k\times\frac{d_v}{h}}$ を算出する $(i = 1, 2, \ldots, h)$.

$$\boldsymbol{H}_i = softmax(\boldsymbol{Q}_i\boldsymbol{K}_i^T/\sqrt{d_k})\boldsymbol{V}_i \tag{1}$$

ここで，$softmax$ はベクトルの要素の合計値が 1 になるように基準化する関数を表す．また，$\boldsymbol{Q} \in \mathbb{R}^{k\times d_k}$，$\boldsymbol{K} \in \mathbb{R}^{k\times d_k}$，$\boldsymbol{V} \in \mathbb{R}^{k\times d_v}$ は全体のクエリ・キー・バリューを表し，特徴量を線形変換することによって得られる．そして，i 番目のヘッドに対応するクエリ・キー・バリューを $\boldsymbol{Q}_i \in \mathbb{R}^{k\times\frac{d_k}{h}}$，$\boldsymbol{K}_i \in \mathbb{R}^{k\times\frac{d_k}{h}}$，$\boldsymbol{V}_i \in \mathbb{R}^{k\times\frac{d_v}{h}}$ と表す．クエリとキーの内積によって類似度を計算し，その値を元にバリューに重みづけすることにより，Attention を考慮した各特徴量の i 番目のヘッドに対する潜在表現が求まる．

そして，式 (2) によって h 個のヘッドを一つの行列に連結 $(concat)$ して全体の潜在表現を算出し，$\boldsymbol{W} \in \mathbb{R}^{d_v\times d}$ により線形変換して，元の入力と同じ形の出力を得る．

$$Multihead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = concat(\boldsymbol{H}_1, \ldots, \boldsymbol{H}_h)\boldsymbol{W} \tag{2}$$

また，FFN 層はディープニューラルネットワーク構造を有しており，順伝播の役割を担っている．MHSA 層と FFN 層を多層に積み重ねることにより，Transformer の出力が得られる．このような Attention 機構により，Transformer は文書や画像のデータにおいて，特徴量同士の高度な関係性を捉えることができ，高い予測性能を示す．

2.2. FT-Transformer

FTT[2] とは表形式データに対して Transformer を適用したモデルである．図 1 に FTT の概要図を示す．

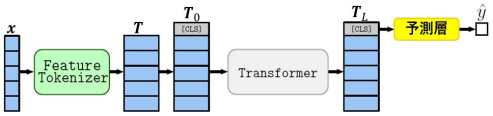

図 1: FT-Transformer の概要図

FTT では，入力データに含まれる量的変数と質的変数を Feature Tokenzier によって特徴量 $\boldsymbol{T}$ に変形する．さらに，式 (3) により，予測層への入力のベースとなる Classification(以下，CLS) トークン $[CLS]$ を $\boldsymbol{T}$ に付加 $(stack)$ し，Transformer への入力 $\boldsymbol{T}_0$ を得る．

$$\boldsymbol{T}_0 = stack[[CLS], \boldsymbol{T}] \tag{3}$$

Transformer では，データに含まれる各特徴量を，CLS トークンをベースとした一つのベクトルに反映することによって予測に用いる．Transformer を L 層積み重ね，最終的な出力 $\boldsymbol{T}_L$ を得る．そして，$\boldsymbol{T}_L$ のうち，CLS トークンに対応するベクトルのみが予測層へ入力される．予測層では変換された CLS トークンに対応するベクトルを線形変換し，回帰タスクなら数値，分類タスクなら各クラスの確率を予測値として得る．

3. 提案手法

Transformer が提案された言語翻訳タスクにおいては，各単語間の高度な関係性，すなわち，一つ一つの単語間の結びつきが重要であり，これらを詳細に計算・考慮する必要がある．そのため，従来の Transformer モデルは図 2 左の①のように CLS トークンおよび文章中の全ての単語間におい

て Attention を (しかもヘッドごとに) 計算している．しかし，表形式データにおいては，文書や画像とは異なり，全ての特徴量間の高度な関係性はそこまで予測に大きな影響を与えないと考えられる．

そこで，Attention の計算を行う特徴量を重要な箇所のみに限定することを考える．これにより，重要性の低い情報を考慮しなくて済み，推定精度の向上と計算の効率化が期待できる．また，FTT では CLS トークンをベースに予測が行われるため，CLS トークンと各特徴量との関係性が特に重要であると考えられる．

ここまでを踏まえ，提案手法では, FTT に含まれる Transformer 内部の Attention 計算において, CLS トークンとその他の特徴量との Attention のみを計算する．提案手法に含まれる Attention 機構のイメージを図 2 右の③に示す．

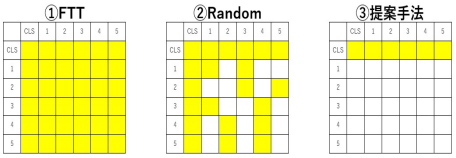

図 2: 各モデルにおける Attention の計算箇所の例

4. 実験

4.1. 実験条件

提案手法の有効性を明らかにするために，異なる 3 つのタスク (多値分類・二値分類・回帰) に関する実験を行う．多値分類には ALOI データセット (データ数:108,000, 特徴量の数:128，ラベル数:1,000)，二値分類には Higgs Small(Higgs) データセット (データ数:98,050，特徴量の数:28)，回帰には California Housing(CA) データセット (データ数:20,640，特徴量の数:8) を用いる．

比較モデルとして, 1)FTT，2)FTT における Attention 計算で CLS トークンはすべての特徴量と，その他の特徴量はランダムに選択した特徴量間でのみ Attention を計算するモデル (以下，Random)，3)GBDT モデルの一つである LightGBM を用いる．Random では，選択する特徴量の数を分類タスクでは 5,10,20, 回帰タスクでは 1,3,5 の 3 パターン設定し，それぞれ Random1,2,3 と表現する. 図 3 に特徴量の数が 5 個の場合の提案手法と比較手法の Attention 計算箇所の例を示す．さらに，評価指標としては，分類タスクには正解率，回帰タスクには RMSE を用いる．

4.2. 実験結果と考察

表 1 に精度の比較結果を示す．なお，表 1 において*, **はそれぞれ提案手法と FTT の結果を比較した t 検定の結果，それぞれ 5 %，10 %の有意差があることを表す．

表 1: 各データセットにおける性能比較 (10 回の平均)

モデル	ALOI	Higgs	CA
FTT	0.9513	0.7251	0.5717
Random1	0.9525	0.7244	0.6246
Random2	0.9524	0.7245	0.5732*
Random3	0.9529	0.7196	0.5754*
LightGBM	0.9342	0.7233	0.5678
提案手法	**0.9570****	**0.7260****	**0.5627***

表 1 より, 全てのタスクに対して，提案手法が最も良い性能を示していることがわかる．このことから，表形式データにおいては Attention の計算箇所を絞ることが精度に良い影響を与えたといえる．また, GBDT モデルの一つである LightGBM と比較した際も，すべてのタスクにおいて提案手法が優れており，今後，表形式データに対しても DL モデルが有効な分析手法となる可能性が示唆される．

次に，実行時間に関する実験結果を表 2 に示す．

表 2: 計算量に関する結果比較 (秒)

モデル	ALOI	Higgs	CA
提案手法	**2643 ± 472**	**185 ± 19**	**50 ± 8**
FTT	3880 ± 752	229 ± 31	54 ± 14

表 2 より, 全てのデータセットにおいて提案手法が従来手法よりも実行時間が大幅に短縮されていることがわかる．特に, 特徴量の数が 128 と多い ALOI では, 計算時間が 30 % ほど短縮されている．加えて, 提案手法は従来手法よりも実行時間のばらつきが小さくなっている．

さらに，図 3 に提案手法において, CA における全 20 個の入力データ (サンプル) に対して算出された (CLS トークンとの)Attention(各特徴量が出力にどれだけ影響を与えているか) の結果に関するヒートマップを示す．

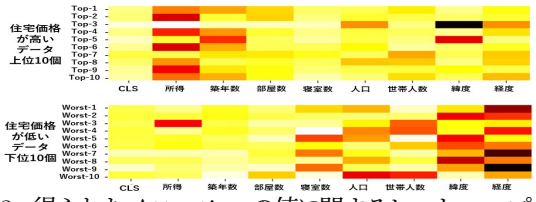

図 3: 得られた Attention の値に関するヒートマップ (上: 住宅価格が高いデータ，下: 住宅価格が低いデータ)

図 3 より, 住宅価格が高いデータにはその地域の平均所得や住宅の築年数が寄与しており, 住宅価格の低いデータには住宅の場所 (緯度や経度) が寄与していることがわかる．

以上の実験結果より提案手法は様々なタスクにおいて，推定精度を向上させつつ，効率化を実現できたといえる．また，得られた Attention を観察することにより, 予測結果に対して各特徴量が与えた影響の大きさを確認することができる点から，解釈性の面でも有用なモデルであるといえる．

5. まとめと今後の課題

本研究では, FTT を表形式データの特性に合うように改良した手法を提案した．また，機械学習における代表的な 3 つのタスクに関する評価実験により，推定精度・計算量・解釈性の観点から提案手法の有効性を示した．今後の課題として，解釈性をより向上させることなどが挙げられる．

参考文献

[1] A Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 30, 2017.

[2] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko. Revisiting deep learning models for tabular data. In *Advances in Neural Information Processing Systems*, Vol. 34, pp. 18932–18943, 2021.