TABULAR DATA: DEEP LEARNING IS NOT ALL YOU NEED

论文介绍

Tabular data: Deep learning is not all you need

R Shwartz-Ziv, A Armon - Information Fusion, 2022 - Elsevier

A key element in solving real-life data science problems is selecting the types of models to use. Tree ensemble models (such as XGBoost) are usually recommended for classification and regression problems with tabular data. However, several deep learning models for tabular data have recently been proposed, claiming to outperform XGBoost for some use cases. This paper explores whether these deep models should be a recommended option for tabular data by rigorously comparing the new deep models to XGBoost on various …

☆ 保存　99 引用　被引用次数：188　相关文章　所有 5 个版本

Information Fusion

Volume 81, May 2022, Pages 84-90

分数不高也不太低；
引用不高也不太低；
但对我研究也许有帮助，所以尝试一下

Information Fusion

Supports *open access*

28.4	17.564
CiteScore	Impact Factor

背景与目的

- 目前有不少基于深度学习的表格数据学习模型，且在各自论文中表现得与集成学习（如：RF,XGBoost）不相上下，甚至优于这些模型；
- 作者思考，真的如此吗？
- 于是作者做了基于不同数据集（11个数据集）下的实验：
  - 单个表格深度学习模型（TabNet, NODE, DNF-Net, 1D-CNN）
  - 单个集成学习模型：XGBoost
  - 表格深度学习模型与XGBoost集成
  - 表格深度学习模型与非XGBoost模型（如：SVM）集成

内容摘要

Dataset	Features	Classes	Samples	Source	Paper
Gesture Phase	32	5	9.8k	OpenML	DNF-Net
Gas Concentrations	129	6	13.9k	OpenML	DNF-Net
Eye Movements	26	3	10.9k	OpenML	DNF-Net
Epsilon	2000	2	500k	PASCAL Challenge 2008	NODE
YearPrediction	90	1	515k	Million Song Dataset	NODE
Microsoft (MSLR)	136	5	964k	MSLR-WEB10K	NODE
Rossmann Store Sales	10	1	1018K	Kaggle	TabNet
Forest Cover Type	54	7	580k	Kaggle	TabNet
Higgs Boson	30	2	800k	Kaggle	TabNet
Shrutime	11	2	10k	Kaggle	New dataset
Blastchar	20	2	7k	Kaggle	New dataset

- 共11个数据集，其中9个来自表格深度学习模型（TabNet, NODE, DNF-Net）源论文使用的数据集（3个模型分别3个数据集，共9个），2个来自论文外作者自己找的数据集，对比这些模型对论文外数据集的泛化能力；
- 得到XGBoost在11个数据集上综合结果优于表格深度学习模型；
- 得到表格深度学习模型与XGBoost集成结果最优，而表格深度学习模型与非XGBoost模型集成结果不理想；
- 在运行速度、超参数搜索上，XGBoost均优于表格深度学习模型；

实验结果

Name	Average Relative Performance (%)
XGBoost	3.34
NODE	14.21
DNF-Net	11.96
TabNet	10.51
1D-CNN	7.56
Simple Ensemble	3.15
Deep Ensemble w/o XGBoost	6.91
Deep Ensemble w XGBoost	**2.32**

Model Name	Rossman	CoverType	Higgs	Gas	Eye	Gesture
XGBoost	490.18 ± 1.19	3.13 ± 0.09	21.62 ± 0.33	2.18 ± 0.20	**56.07**±0.65	80.64 ± 0.80
NODE	488.59 ± 1.24	4.15 ± 0.13	21.19 ± 0.69	2.17 ± 0.18	68.35 ± 0.66	92.12 ± 0.82
DNF-Net	503.83 ± 1.41	3.96 ± 0.11	23.68 ± 0.83	**1.44** ±0.09	68.38 ± 0.65	86.98 ± 0.74
TabNet	**485.12**±1.93	3.01 ± 0.08	**21.14**±0.20	1.92 ± 0.14	67.13 ± 0.69	96.42 ± 0.87
1D-CNN	493.81 ± 2.23	3.51 ± 0.13	22.33 ± 0.73	1.79 ± 0.19	67.9 ± 0.64	97.89 ± 0.82
Simple Ensemble	488.57 ± 2.14	3.19 ± 0.18	22.46 ± 0.38	2.36 ± 0.13	58.72 ± 0.67	89.45 ± 0.89
Deep Ensemble w/o XGBoost	489.94 ± 2.09	3.52 ± 0.10	22.41 ± 0.54	1.98 ± 0.13	69.28 ± 0.62	93.50 ± 0.75
Deep Ensemble w XGBoost	485.33 ± 1.29	**2.99** ±0.08	22.34 ± 0.81	1.69 ± 0.10	59.43 ± 0.60	**78.93** ±0.73

TabNet　　DNF-Net

Model Name	YearPrediction	MSLR	Epsilon	Shrutime	Blastchar
XGBoost	77.98 ± 0.11	55.43±2e-2	11.12±3e-2	13.82 ± 0.19	20.39 ± 0.21
NODE	76.39 ± 0.13	55.72±3e-2	**10.39**±1e-2	14.61 ± 0.10	21.40 ± 0.25
DNF-Net	81.21 ± 0.18	56.83±3e-2	12.23±4e-2	16.8 ± 0.09	27.91 ± 0.17
TabNet	83.19 ± 0.19	56.04±1e-2	11.92±3e-2	14.94±, 0.13	23.72 ± 0.19
1D-CNN	78.94 ± 0.14	55.97±4e-2	11.08±6e-2	15.31 ± 0.16	24.68 ± 0.22
Simple Ensemble	78.01 ± 0.17	55.46±4e-2	11.07±4e-2	13.61±, 0.14	21.18 ± 0.17
Deep Ensemble w/o XGBoost	78.99 ± 0.11	55.59±3e-2	10.95±1e-2	14.69 ± 0.11	24.25 ± 0.22
Deep Ensemble w XGBoost	**76.19** ±0.21	**55.38**±1e-2	11.18±1e-2	**13.10**±0.15	**20.18**±0.16

NODE　　New datasets

回归问题使用均方误差；
分类问题使用交叉熵损失；

初始超参数为源论文设定；
每个数据集每个算法分别进行1000步超参数搜索；

两种集成方式：

$$p(y|x) = \sum_{k=1}^{K} p_{\theta_m}(y|x, \theta_m)$$

$$p(y|x) = \sum_{k=1}^{K} l_k^{\text{val}} p_{\theta_m}(y|x, \theta_m)$$

实验方法

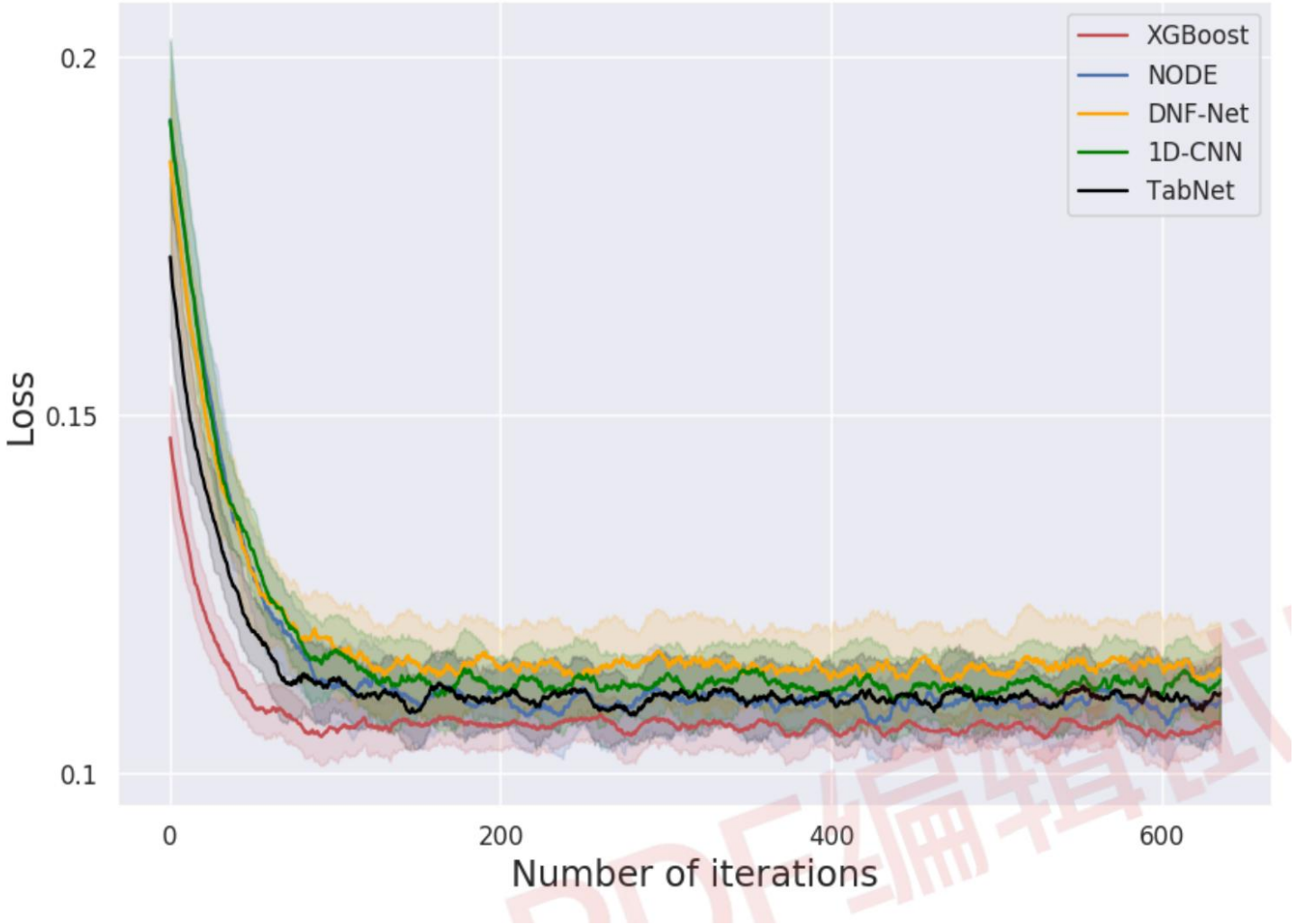
Q1: 不同数据集下，如何做不同算法对比（统计学意义检测）：

A1：Friedman test（对比不同数据集下，不同算法得到的结果，检测这些算法是否有显著差别）

Q2: 除了评价平均标准误差之外，还有哪些指标可评价模型好坏?

A2: 每秒计算浮点数（但超参数不同，其值不同）、运算总时长（但应用广的模型优化较好，因此该指标对于新模型不公平）、超参数优化的迭代次数（该指标体现寻找超参数难易程度，体现业界需求）

超参数优化的迭代次数对比



为什么会有这种结果?

Q1: 这些表格深度学习模型为何仅在自己论文数据集下表现好?

A1: （1）选择偏误，仅选择最好的数据集写论文

（2）超参数优化，每一篇论文都可以基于对文中提供的数据集进行更广泛的超参数搜索来设置模型的超参数，从而获得更好的性能;

Q2: 为什么XGBoost能表现得好?

A2: （1）XGBoost初始化超参数可能更好，可能因为XGBoost是基于更多数据集而提出;

（2）XGBoost可能存在内置特性使其更易优化;

讨论与展望

- 未来表格数据的研究，需要基于更多数据集；
- 开发新的深度学习模型，使其更易于优化，与XGBoost这类模型竞争；

论文评价

- 该论文带着问题的方式写作，提高读者的阅读兴趣，值得学习；
- 作者思考问题很仔细，思路很严谨；

3.2 Results

Do the deep models generalize well to other datasets?

Do we need both XGBoost and deep networks?

How difficult is the optimization?