 ISBN: 9789073592421

IPS Paper

Deep Learning on Administrative Tabular Data: A Comparative Study

Author: Dr Chiung Ching Ho

Coauthors: Chi-Ken Shum

Submission ID: 1304

Reference Number: 1304

Presentation File

abstracts/ottawa-2023_b829446887bf89eb027ce79caf068722.pdf

Brief Description

Deep learning has traditionally been applied to perform analytics on large unstructured data such as videos, audio, images, and text.

Initially, deep learning research on tabular data was not performed much, as the fixed structure inherent in tabular data was seen to negate the ability of deep learning techniques to elicit useful representation of tabular data.

Recent advances in tabular deep learning have seen applications of the self-attention and transformer architecture as well as transfer learning which has improved the performance of deep learning on tabular data.

Some of these approaches has improved on the results attained using traditional machine learning models such as gradient boosted trees for classification and regression tasks.
While these results are promising, the datasets used in these works were datasets which were typically used for bench-marking machine learning algorithms.

This raises the question on how extensible the results would be if it were to be applied to administrative tabular data.

To answer this question, we will curate a selection of administrative tabular datasets from open-sourced Malaysian administrative data.

The curated dataset will be used to perform classification tasks using both traditional machine learning approaches as well as deep learning approaches.

Classification is a machine learning technique which has been used to support policy decisions .

As an evaluation, we will evaluate the results of the classification tasks as a measure of feature representation.
Feature representation is an important measure, as feature selection of administrative data by subject-matter-experts is a manual and laborious task.

It is believed that automatic feature elicitation via deep learning approaches will reduce this dependency.

The results of experiments conducted in this study have shown that deep learning tabular algorithms can achieve comparable results with optimised traditional machine learning when applied on open administrative data, without the need for extensive feature selection and feature engineering.

Abstract

Deep learning has traditionally been applied to perform analytics on large unstructured data such as videos, audio, images, and text. Initially, deep learning research on tabular data was not performed much, as the fixed structure inherent in tabular data was seen to negate the ability of deep learning techniques to elicit useful representation of tabular data. Recent advances in tabular deep learning have seen applications of the self-attention and transformer architecture as well as transfer learning which has improved the performance of deep learning on tabular data.
Some of these approaches has improved on the results attained using traditional machine learning models such as gradient boosted trees for classification and regression tasks.
While these results are promising, the datasets used in these works were datasets which were typically used for bench-marking machine learning algorithms. This raises the question on how extensible the results would be if it were to be applied to administrative tabular data. To answer this question, we will curate a selection of administrative tabular datasets from open-sourced Malaysian administrative data. The curated dataset will be used to perform classification tasks using both traditional machine learning approaches as well as deep learning approaches.
Classification is a machine learning technique which has been used to support policy decisions . As an evaluation, we will evaluate the results of the classification tasks as a measure of feature representation.

Feature representation is an important measure, as feature selection of administrative data by subject-matter-experts is a manual and laborious task. It is believed that automatic feature elicitation via deep learning approaches will reduce this dependency. The results of experiments conducted in this study have shown that deep learning tabular algorithms can achieve comparable results with optimised traditional machine learning when applied on open administrative data, without the need for extensive feature selection and feature engineering.