Data Science Schneider Electric

Javier Grandío González

Data acquisition

Csv read data using Pandas.

Get Json data from website.

Parse PDF files.

Concatenate everything into a single pandas dataframe. With this, the raw data available for training is joined together.

Modify attributes format. To make sure that a single attribute is not mixing strings with numeric data.

Data Preprocessing

Remove useless attributes. There are several attributes that are not relevant for the task. E.g. reporter. Also feature attributes that have thousands of values are discarded because they do not provide useful information. Finally, some attributes encode the same information, so duplicate information is removed.

Encode feature inputs in one-hot. Some labels need to be encoded, but they do not have any ordinal meaning. In consequence, the feature attributes that do not have too many values are encoded as one-hot.

Finally, some attributes are encoded using tensorflow embedding layers. Some attributes have to many possible values to encode them in one-hot vectors, so this approach is taken.

Models evaluation

Compare different models to evaluate performance:

- Random Forest
- Adaboost classifier
- Feed Forward Neural Network → Best results

	precision	recall	f1-score	support
0	0.64	0.62	0.63	5169
1	0.60	0.60	0.60	4618
2	0.90	0.91	0.90	3339
accuracy			0.69	13126
macro avg	0.71	0.71	0.71	13126
weighted avg	0.69	0.69	0.69	13126