Recognising innovative companies by using a diversified stacked generalisation method for website classification – the raw results

1. Introduction

The classification models were trained out by using the Classification and Regression Training package (caret)[1]. The models' parameters were fine-tuned by the 10-fold cross-validation procedure[2].

2. Cluster parameters

Most computations were carried out on a cluster having the following parameters:

- GPU: NVIDIA Tesla P100;
- CPU: 2.0 GHz Intel® Xeon® Platinum 8167M;
- The number of GPUs: 2;
- The number of CPU cores: 28;
- The number of CPU threads: 56;
- RAM: 192 GB;
- Storage: 3 TB.

Only one model (k-nn) was calculated on a cluster having the following parameters:

- Processor: Intel(R) Core(TM) i7-4770 CPU @ 3.40GHz 3.40 GHz;
- RAM: 16 GB;
- Windows 64 bit.

3. Performance statistics

All performance statistics are stored in cvs files. Each file corresponds to a particular machine learning method such as a file, "methodName-stat.csv" contains all data regarding a method, "methodName." All files cover the following columns:

- *dataSetName* – a name of a data set on which evaluation was carried out; there are three possible values: (i) *firstPages* refers to the first data set (L_D) that contains textual description of a company; (ii) *firstPageLabels* refers to the second data set (L_L) that involves link labels that were extracted from an index page; (iii)

1 https://cran.r-project.org/web/packages/caret/

2 https://topepo.github.io/caret/model-training-and-tuning.html

aggregateDocument refers to the third data set (L_B) that consists of a so-called big document;

- *featureNo* – the number of features that were taken into account during evaluation;
- *method* – the name of function in the caret package;
- *parameters* – the values of parameters received from a tuning phase of a given classification method;
- *precision* – the value of method's precision;
- *recall* – the value of method's recall;
- *fmeasure* – the value of method's F-measure;
- *error* – the value of method's error;
- *acc* – the value of method's.

4. Time processing statistics

All time processing statistics, like the performance statistics, are stored in cvs files. Each file corresponds to a particular machine learning method such as a file, "methodName-time.csv". All files cover the following columns:

- *dataSetName* – a name of a data set on which evaluation was carried out; there are three possible values: (i) *firstPages* refers to the first data set (L_D) that contains textual description of a company; (ii) *firstPageLabels* refers to the second data set (L_L) that involves link labels that were extracted from an index page; (iii) *aggregateDocument* refers to the third data set (L_B) that consists of a so-called big document;
- *featureNo* – the number of features that were taken into account during evaluation;
- *method* – the name of function in the caret package;
- *user* – user time elapsed for executing a *method* as an R process;
- *system* – system time elapsed for executing a *method* as an R process;
- *elapsed* – total time elapsed for executing a *method* as an R process;

For more information about user, system and total elapsed time, please see documentation[3].

[3] https://stat.ethz.ch/R-manual/R-devel/library/base/html/proc.time.htm