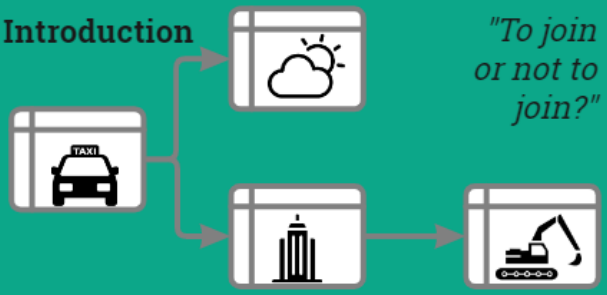
From Feature Selection to Data Augmentation:
The ADA Algorithm

Introduction



"To join or not to join?"

Figure 1. Potential tables to join

Let's predict how much will a taxi ride take! We have data on the driver, the locations and the time. What other data can be useful? The weather? the construction sides on the way?

After exploiting all the signal in our parent table, how can we improve the performance of our model by improving the data?

Research Question

With the existece of thousands of publicly available repositories containing thousands of tables, can we augment our data in an efficient and automatic way?

- What characteristics make a feature desirable?
- Can we devise an efficient algorithm for automatic data augmentation using sample joins?
- How does this approach work for different datasets and other models?

TU Delft

Methodology

Using Support Vector Machines (SVMs), a myriad of feature selection techniques and nine different datasets, we examine the desirability of the different 115 features analysed.

Univariate. The algorithm computes table desirability using variance, frequency, datatypes and number of variables. Only if a threshold is reached, then the algorithm performs a sample join.

Multivariate. After the sample join is performed, the algorithm computes the desirability of each variable. It computes the correlation with the one-dimensional projection of the linear discriminant analysis. It also takes into account the number of categories and their frequency for the categorical variables.

The ADA Algorithm

Algorithm 3 ADA

Input: Base table t_0, other tables and path pair (t_i, p_i) in T.
$\rho \leftarrow proportion(target\_feature)$
for $(t_i, p_i) \in T$ using BSF **do**
  $d \leftarrow univariateDesirability(t_i, \rho)$
  if $d_i \geq threshold_0$ **then**
    $s_i \leftarrow sampleJoin(t_i)$
    $A \leftarrow multivariateDesirability(s_i)$
    $Join(A, t_0)$
return t_0

Figure 2. Pseudo-code of the ADA Algorithm

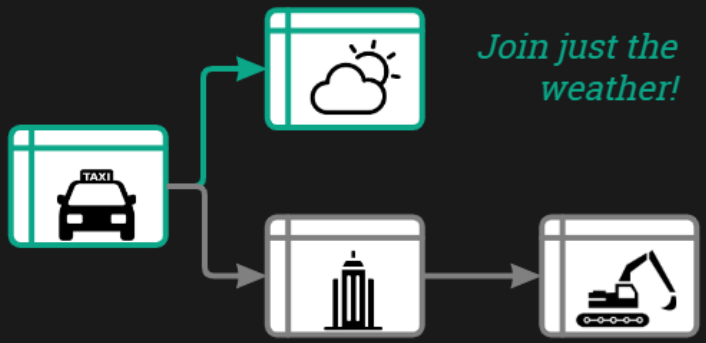


Figure 3. Tables that are beneficial to join

Results

Approach	Support Vector Machine - Linear		
	Train	Test	Time
Base Table	0,729	0,675	33ms
Naive Join-All	0,979	0,983	17,2s
Feature Selection	0,993	0,992	3,1s
ADA Algorithm	0,925	0,908	41,2s

Figure 4. Results for SVM and the Kidney Disease dataset

The ADA algorithm manages to capture most of the signal contained in the entire dataset: on average, 96,45% of the accuracy obtained by performing feature selection on the join-all table.

Nevertheless, the sample joins are very costly computationaly-wise. This is espcecially the case in the databases used for evaluation where the signal was roughly evenly spread across tables. The univariate analysis was not enough to disregard some of the tables, and sample joins are frequently performed.

Limitations & Future Work

The ADA Algorithm might be helpful in some particular cases. When the signal is contained in a few tables and the data is distributed in many small tables, the ADA Algorithm might be able to capture most of the signal at a reduced computational cost. Neverthless, the evaluation highlights some limitations:

- Using other datasets for the evaluation.
- Sample joins can drastically slow down the algorithm.
- Incorporating the residual variance.
- Stringent data format.

Eduard Cruset Pla
ecrusetpla@student.tudelft.nl

Rihan Hai
Andra Ionescu
{r.hai, a.ionescu-3}@tudelft.nl