Improving css-KNN Classification Performance by Shifts in Training Data

Karol Draszawka[1], Julian Szymański[1] , Francesco Guerra[2]

[1] Gdańsk University of Technology, Poland

[2] Universita' di Modena e Reggio Emilia, Italy

1st International KEYSTONE Conference, Coimbra Portugal, 8-9 September 2015

GDAŃSK UNIVERSITY OF TECHNOLOGY

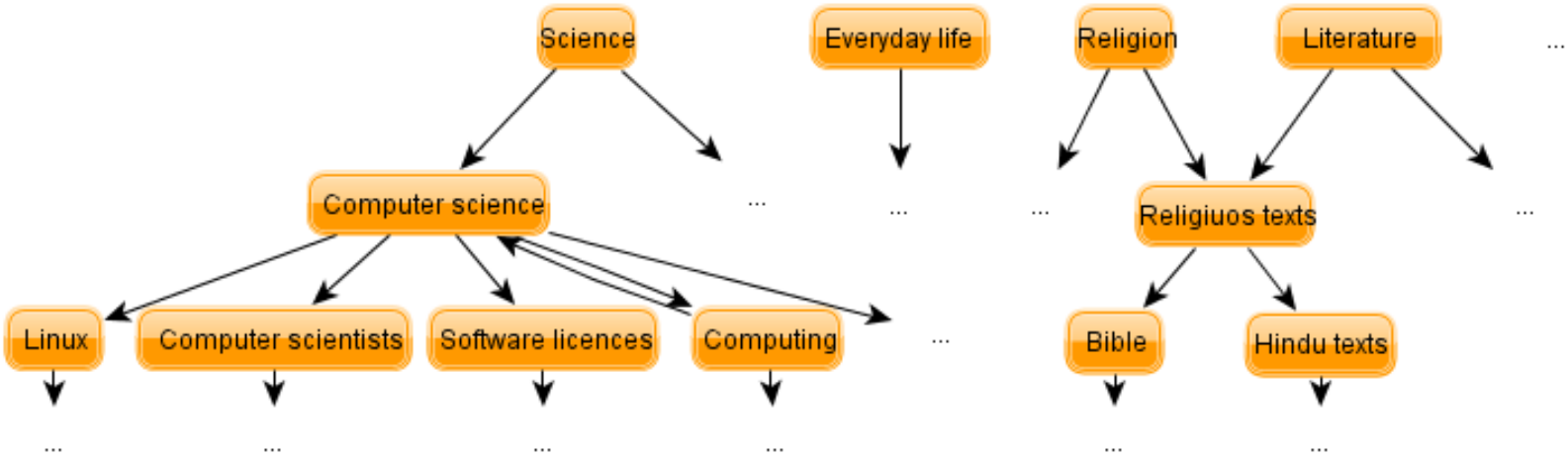
Agenda

1. Classification
2. Shifts in Training Data
   - Common form
   - Shifts towards globally defined destinations
   - Shifts towards locally defined destinations
3. Results

GDAŃSK UNIVERSITY OF TECHNOLOGY

Problem description

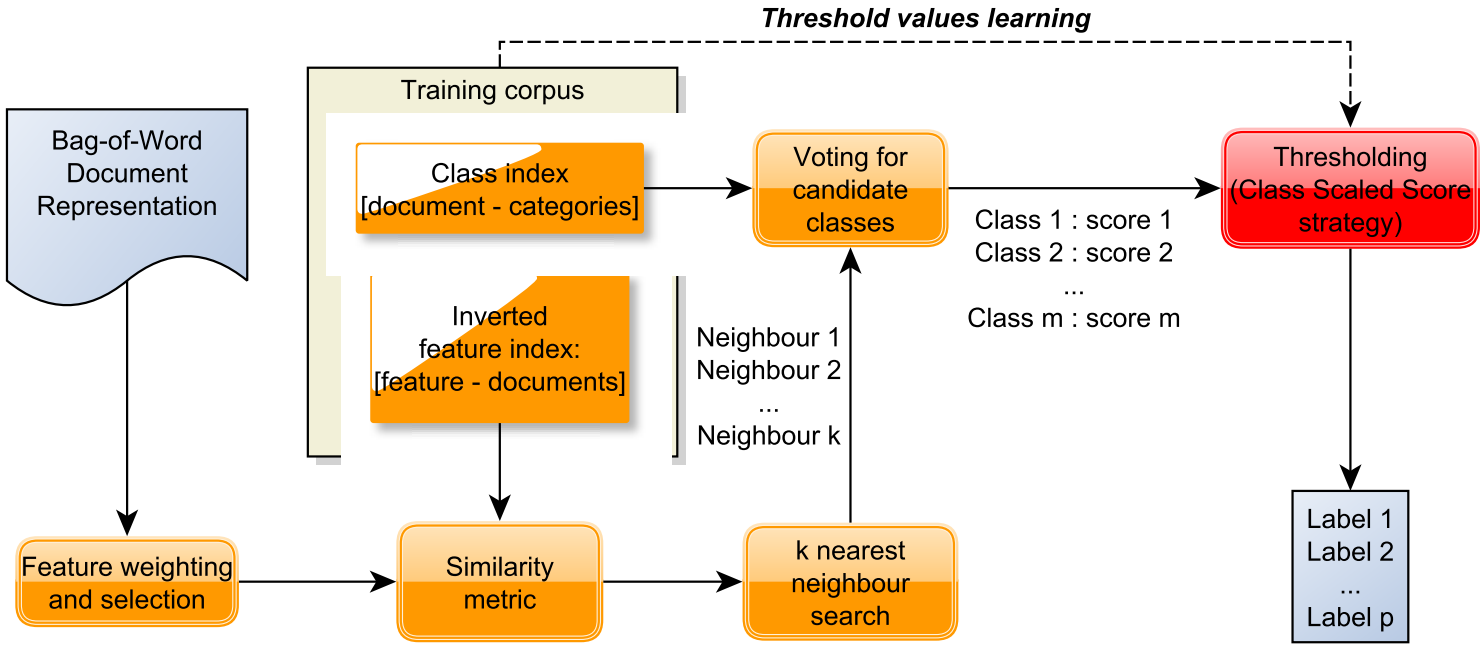
- The need: structured, well organized information
- The problem: *Text Classification*
  - Large Scale (thousands of categories)
  - Multi-Label
  - (pseudo)Hierarchical



GDANSK UNIVERSITY OF TECHNOLOGY

A small snippet of Wikipedia "hierarchy" of categories.

Classifier architecture



This is our version of the classifier proposed by Xiao-Lin Wang et al., the winner of *The Second Pascal Challenge on Large Scale Hierarchical Text classification*.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Thresholding strategies

- *S-cut* – (Score-cut) returns a set of labels which scores exceed or equal some constant threshold value t:

$$h(\boldsymbol{x}) = \{l : \text{score}(l) >= t\}, \quad \forall_{l \in \boldsymbol{L}}$$

- *DS-cut* – (Distinctive Score-cut) uses a few threshold values t_i, each one corresponding to the position that a label occupies according to its score in the list of sorted scores:

$$h(\boldsymbol{x}) = \{l : \text{score}(l) >= t_{\text{rank}(l)}\}, \quad \forall_{l \in \boldsymbol{L}}$$

- *CS-cut* – (Class-specific Score-cut) instead of one global constant t needs one threshold value per each class:

$$h(\boldsymbol{x}) = \{l : \text{score}(l) >= t_l\}, \quad \forall_{l \in \boldsymbol{L}}$$

- *CDS-cut* – (Class-specific Distinctive Score-cut) accordingly:

$$h(\boldsymbol{x}) = \{l : \text{score}(l) >= t_{l,\text{rank}(l)}\}, \quad \forall_{l \in \boldsymbol{L}},$$

- we use scaled versions of these strategies: *DSS-cut* (original EKNN) and our *CSS-cut*.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Training Data Amendments

- feature reweighting (choosing proper weighting scheme)
- outlier detection & removal
- re-labeling of training examples
- shifting/moving/translating examples

GDAŃSK UNIVERSITY OF TECHNOLOGY

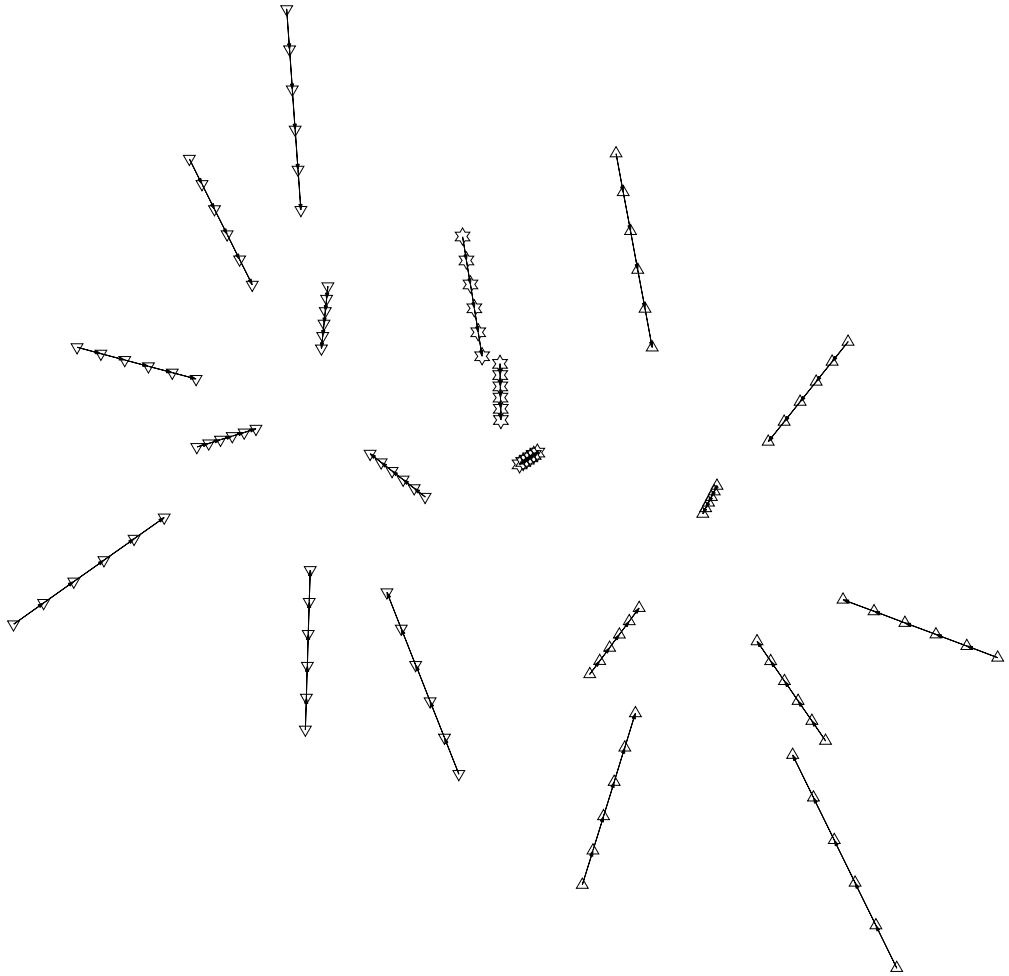
- Data points are modified using vector translation (shift):

$$\boldsymbol{x}'^{(i)} = \boldsymbol{x}^{(i)} + \alpha \cdot \left(\boldsymbol{d}^{(i)} - \boldsymbol{x}^{(i)} \right) = (1 - \alpha) \cdot \boldsymbol{x}^{(i)} + \alpha \cdot \boldsymbol{d}^{(i)}$$

- i-th data point $\boldsymbol{x}^{(i)}$ is translated towards some destination point $\boldsymbol{d}^{(i)}$ by some *step size* $\alpha \in [0, 1]$
- the knowledge of $\mathcal{Y}^{(i)}$ as well as the rest of D_{train} is necessary to compute $\boldsymbol{d}^{(i)}$

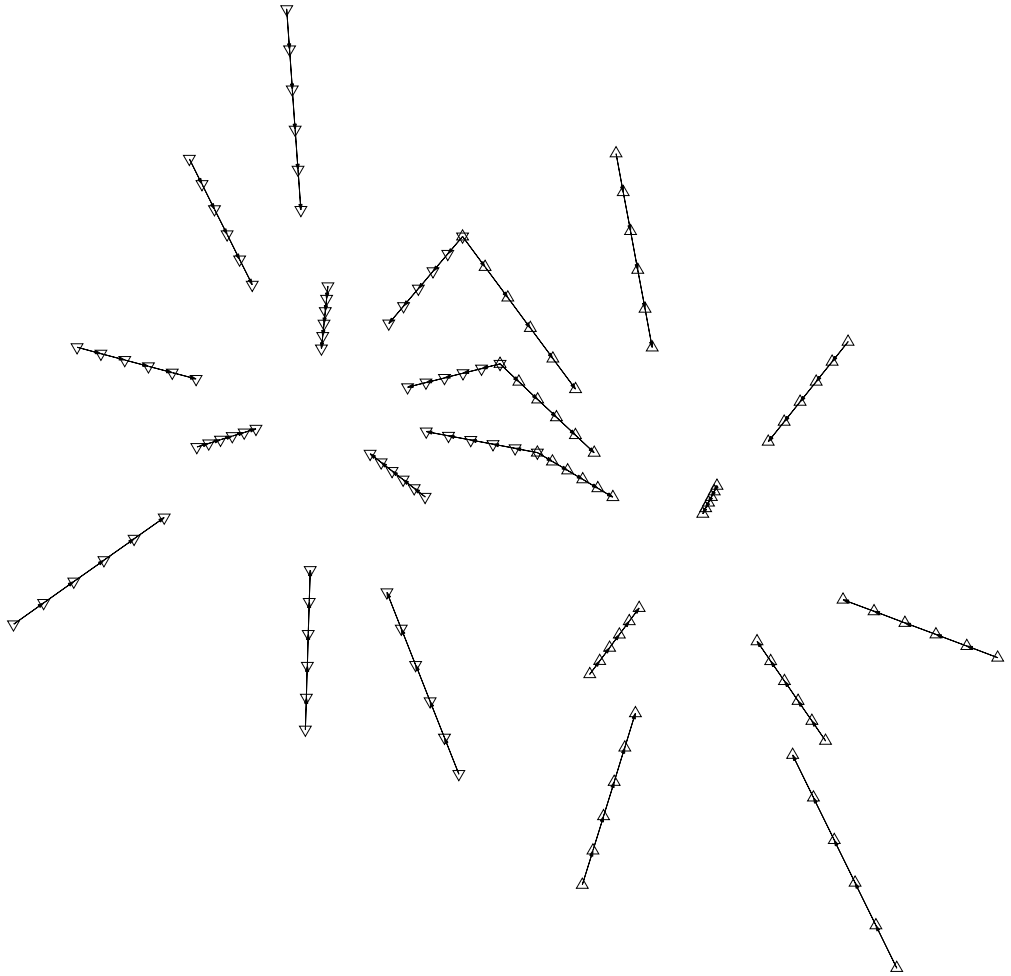
GDAŃSK UNIVERSITY OF TECHNOLOGY

T2CC – shift to centroid of centroids



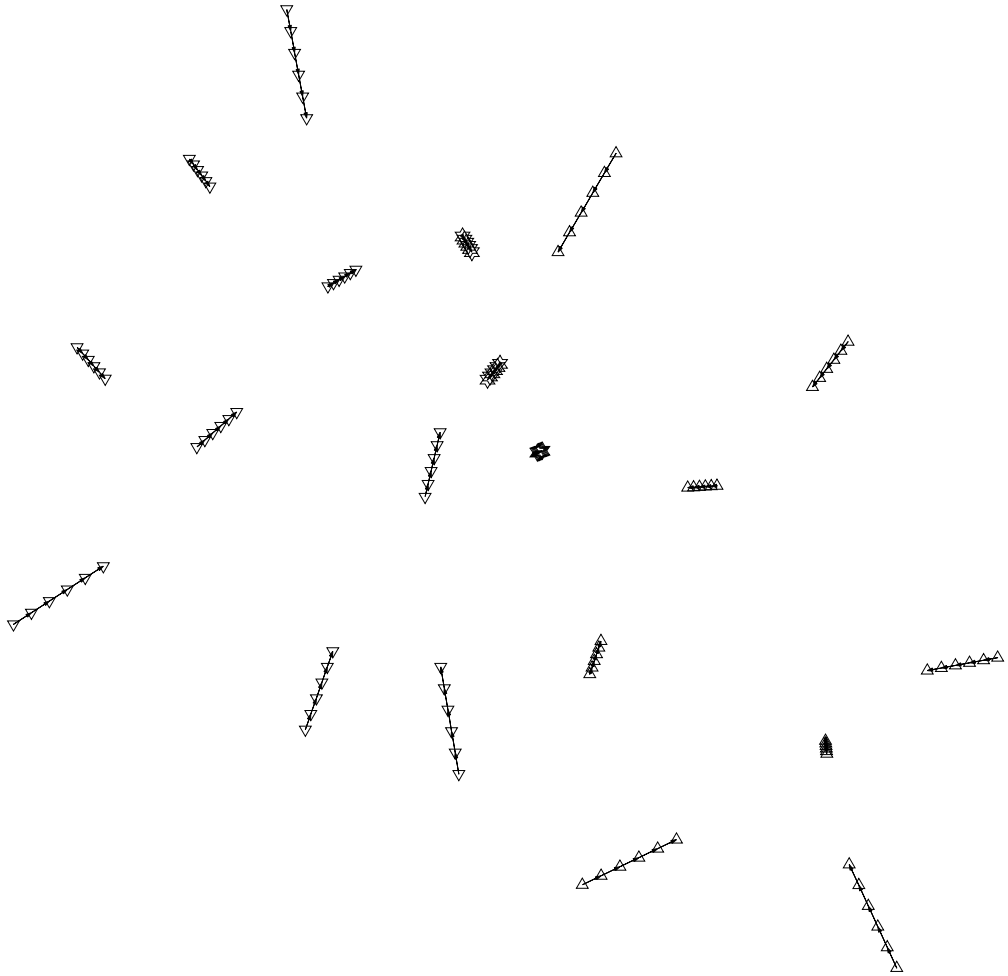
GDAŃSK UNIVERSITY OF TECHNOLOGY

C&T2C – copy & shift to centroid



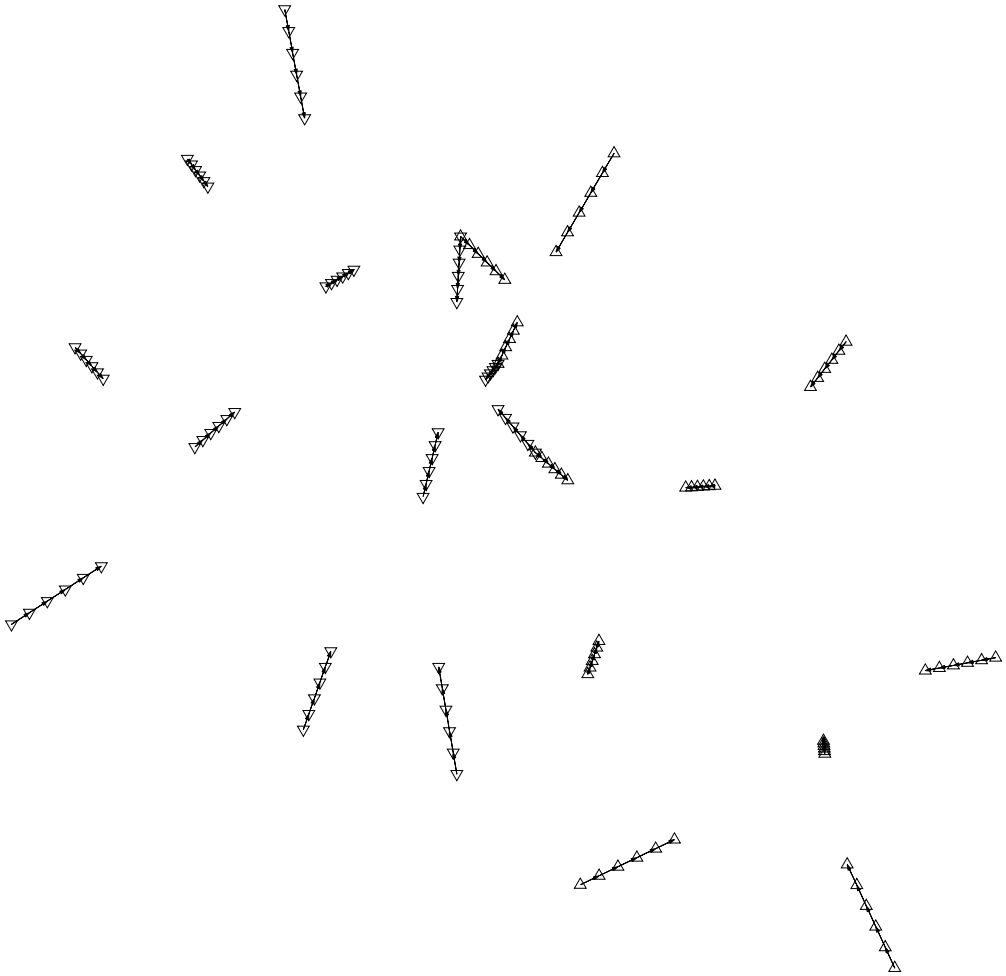
GDAŃSK UNIVERSITY OF TECHNOLOGY

T2CNIN – shift to centroid of nearest in-classes neighbors



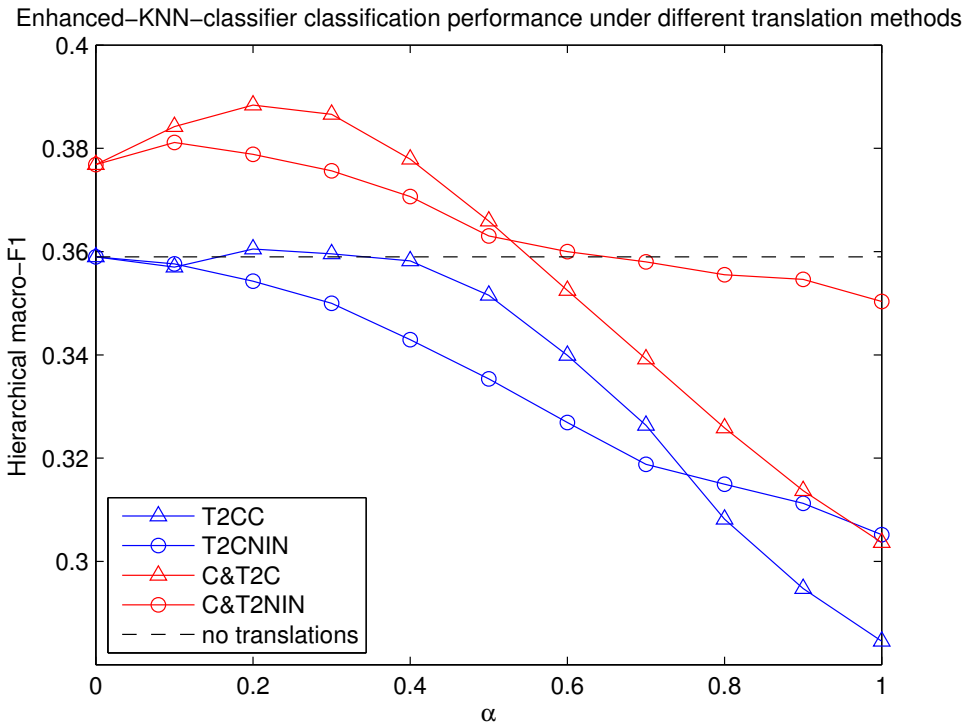
GDAŃSK UNIVERSITY OF TECHNOLOGY

C&T2NIN – copy&shift to centroids of n. in-class neighbors



GDAŃSK UNIVERSITY OF TECHNOLOGY

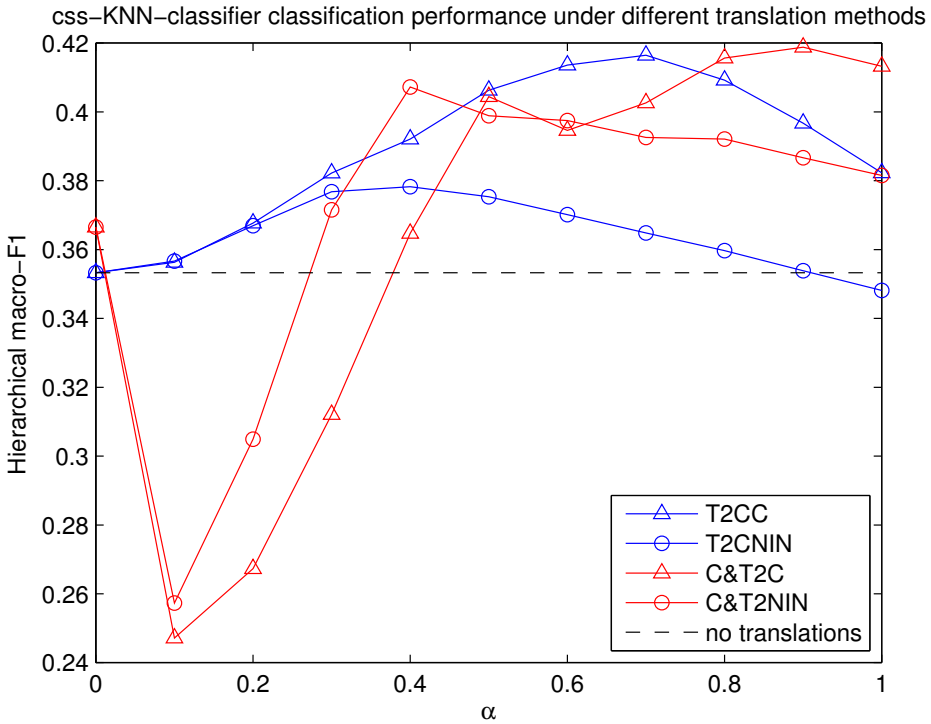
Results



The results of Enhanced-KNN classifier performance on a held out test set under train data translations using proposed methods. Simple English Wikipedia dataset, centroid pruning $th_{rel} = 0.05$, shifted data pruning $th_{rel} = 0.05$, neighborhood size for local methods: $m = 3$.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Results



The results of css-KNN classifier performance on a held out test set under train data translations using proposed methods. Simple English Wikipedia dataset, centroid pruning $th_{rel} = 0.05$, shifted data pruning $th_{rel} = 0.05$, neighborhood size for local methods: $m = 3$.

GDAŃSK UNIVERSITY OF TECHNOLOGY

Importance of pruning

Table: The impact of pre- and post-translation data pruning on the size of the modified training dataset (in MB). Simple English Wikipedia train data. Original train data size is 25.

		post-shift pruning				
		no pruning	relative 0.01	absolute 0.01	relative 0.05	absolute 0.05
pre-shift	no pruning	-	-	-	-	34
	relative 0.01	1060	149	78	40	34
	absolute 0.01	430	143	78	40	34
	relative 0.05	304	127	75	40	34
	absolute 0.05	90	74	60	40	34

GDAŃSK UNIVERSITY OF TECHNOLOGY

Conclusions

- We proposed four variants of shifts in the training data that can improve classification performance of kNN-based classifiers, if the variant and step-factor chosen carefully
- This can be seen as a form of data smoothing or missing value imputation
- In the future we plan to *learn* a transformation in the feature space, so that testing examples can undergo the same processing as training ones

GDAŃSK UNIVERSITY OF TECHNOLOGY