Hub Co-occurrence Modeling for Robust High-dimensional *k*NN Classification

Artificial Intelligence Laboratory
Jožef Stefan Institute

Nenad Tomašev, Dunja Mladenić

SUMMARY

- High-dimensional data poses significant challenges for most standard *k*NN methods.
- It becomes harder to distinguish between relevant and irrelevant points.
- Hubs have been shown to emerge as centers of influence within the data. They dominate most *k*-nearest neighbor sets and have been shown to be quite detrimental.
- We propose a new *k*NN classification method, the Augmented Naïve Hubness-Bayesian *k*NN (ANHBNN), as an extension of the existing hubness-aware NHBNN method. It learns from hub co-occurrences, by utilizing the Hidden Naïve Bayes approach.
- We show that it is possible to extract useful information from past co-occurrences of neighbor points, as the proposed method significantly outperforms the original NHBNN classifier.
- We also investigate the distribution of neighbor co-occurrences in intrinsically high-dimensional data. Our initial experiments reveal many interesting and previously unknown properties of high-dimensional neighbor co-occurrences.

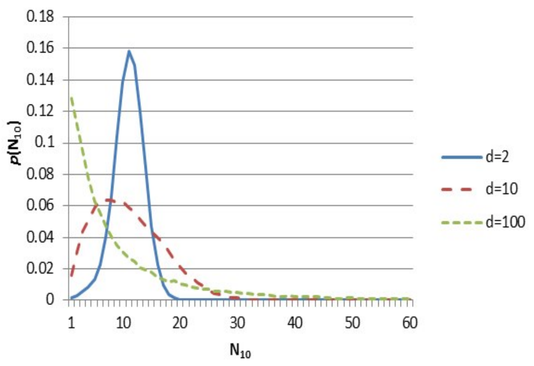
HUBS, ORPHANS AND THE DISTRIBUTION OF INFLUENCE

Hubness is a recently described aspect of the well known curse of dimensionality.

Most points never occur as neighbors (**orphans**) and a small number of points (**hubs**) occur in most *k*NN sets.

This behavior has been documented in many intrinsically high-dimensional data types, such as text, images, audio and time series.

Some hubness-aware methods have recently been proposed for clustering, classification, instance selection and re-ranking in intrinsically high-dimensional data.



Distribution of neighbor *k*-occurrence frequency as dimensionality increases

Consequences of hubness:

- **Reduced classification performance**, as hubs often act as semantic singularities.
- **Information loss**, as most points are never retrieved in *k*NN queries

NOTATION

$D = (X, Y)$: the dataset. X are the feature vectors. Y are the labels.

$n = |D|$ is the data size and n_y is the size of class y.

$D_k(x)$: the set of k-nearest neighbors of point x

$N_k(x)$: the k-occurrence frequency of point x, $N_k(x) = |x_i : x \in D_k(x_i)|$

$N_{k,y}(x)$: the k-occurrence frequency of point x within the kNN sets of points from class $Y = y$

OLD APPROACH: NHBNN

Neighbor occurrences are observed as random events and used as new defining features for query instances. Naïve Bayes rule is used to infer the class affiliation probabilities. Orphans are treated as a special case.

$$p(Y|D_k(X)) \propto p(Y) \prod_{t=1}^{k} p(X_t \in D_k(X)|Y)$$

Problem: Neighbor occurrences are not independent, so the basic assumption is severely violated for larger neighborhood sizes.

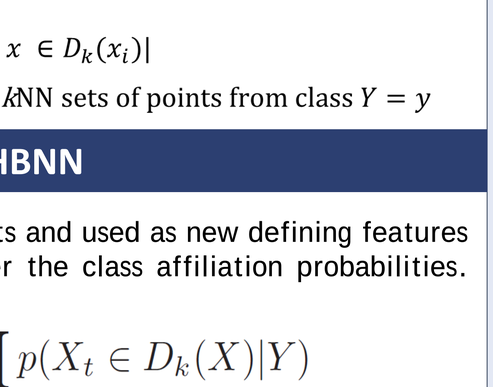
OUR IDEA: MODEL THE NEIGHBOR CO-OCCURRENCES

Hubs/regular points: $p(Y|D_k(X)) \propto p(Y) \prod_{t=1}^{k} p(X_t \in D_k(X)|Y, H_t(X,Y))$

We have used the Hidden Naïve Bayes (NHB) model for modeling hub co-occurrences. Hidden nodes are introduced that model the dependencies between different neighbor points.

We have modified the original NHB model and incorporated the hubness-aware factors derived from past occurrences.

We have named the new algorithm the Augmented Naïve Hubness-Bayesian *k*-Nearest Neighbor (**ANHBNN**).



Orphans/anti-hubs: $N_k(X_t) = 0 : p(X_t \in D_k(X)|Y, H_t(X,Y)) \approx$

$$p(X_t \in D_k(X)|Y) \approx \mathrm{AVG}_{Y_i=Y_t} p(X_i \in D_k(X)|Y) = \frac{N_{k,Y}(Y_t)}{k \cdot n_Y \cdot n_{Y_t}}$$

PROBABILITY ESTIMATES IN ANHBNN

$$p(X_t \in D_k(X)|Y, H_t(X,Y)) = \sum_{i=1, i\neq t}^{k} w_{it}^{Y} \cdot p(X_t \in D_k(X)|X_i \in D_k(X), Y)$$

$$p(X_t \in D_k(X)|X_i \in D_k(X), Y) \approx \begin{cases} \frac{N_{k,Y}(X_t, X_i)}{N_{k,Y}(X_i)}, & \text{if } N_{k,Y}(X_i) > 0, \\ 0, & \text{if } N_{k,Y}(X_i) = 0. \end{cases}$$

$$w_{it}^{Y} = \frac{\frac{I_P(X_i, X_t|Y)}{I_{k,Y}(X_i) \cdot H_k(X_i)}}{\sum_{j=1, j\neq t}^{k} \frac{I_P(X_j, X_t|Y)}{I_{k,Y}(X_j) \cdot H_k(X_i)}}$$

$$I_{k,Y}(X_i) = \log \frac{n_Y}{N_{k,Y}(X_i)}$$

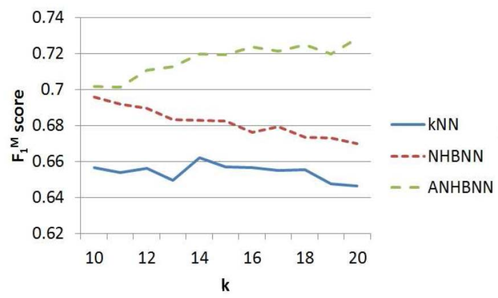
$$H_k(X_i) = \sum_{c \in C} \frac{N_{k,c}(X_i)}{n_c} \log \frac{n_c}{N_{k,c}(X_i)}$$

EVALUATION

The classifiers were compared on image and textual datasets, via cross-validation.

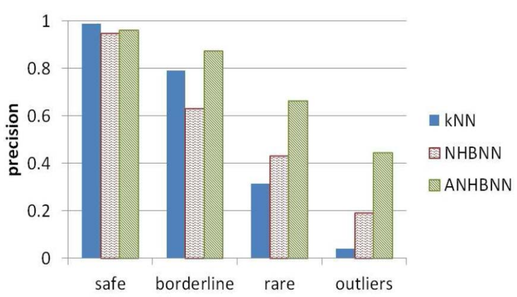
	ANHBNN	NHBNN	*k*NN	hw-*k*NN	h-FNN	HIKNN	Total Wins
ANHBNN	–	**13 (11)**	**14 (14)**	**14 (12)**	**15 (15)**	**11 (11)**	**67 (63)**
NHBNN	2 (1)	–	14 (10)	12 (9)	12 (11)	10 (7)	50 (38)
*k*NN	1 (0)	1 (1)	–	3 (2)	6 (6)	4 (4)	15 (13)
hw-*k*NN	1 (0)	3 (1)	11 (9)	–	11 (11)	8 (5)	34 (26)
h-FNN	0 (0)	3 (1)	8 (7)	3 (2)	–	1 (0)	15 (10)
HIKNN	**3 (2)**	5 (4)	9 (9)	6 (5)	13 (11)	–	36 (31)

Different neighborhood sizes:



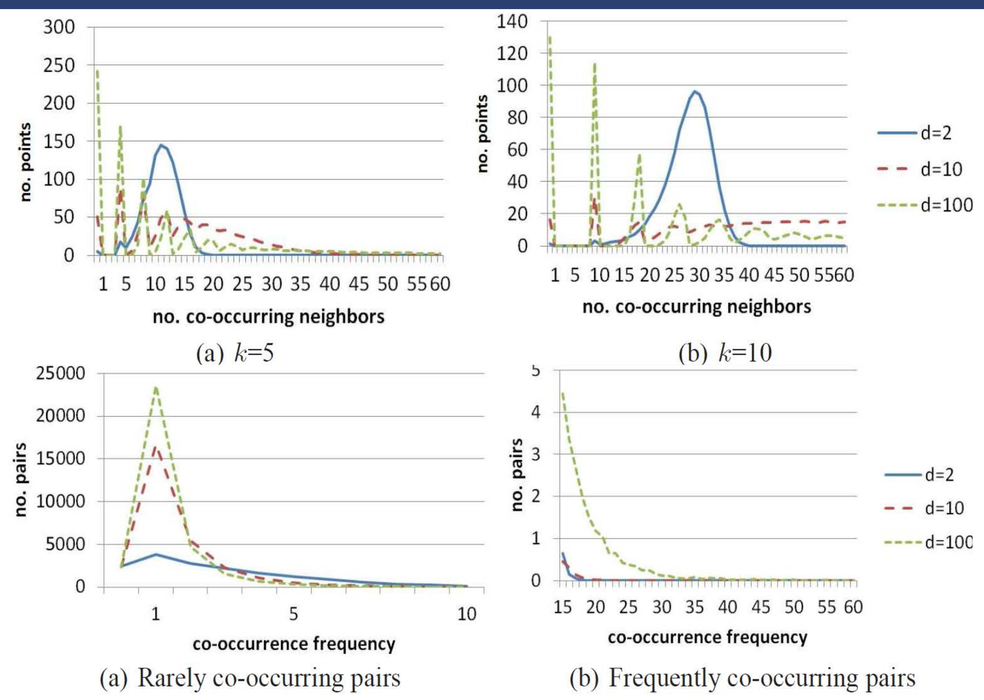
iNet3Imb

Different types of points:



iNet3Imb

CO-OCCURRENCE DISTRIBUTION IN HIGH-DIM. DATA: HUB LINKAGE



(a) Rarely co-occurring pairs (b) Frequently co-occurring pairs

CONCLUSIONS

- Hubness-aware classification methods increase the effectiveness of *k*NN classification in intrinsically high-dimensional data.
- We have proposed a novel *k*NN classification algorithm, ANHBNN. It offers significant improvements in classification performance over the compared baselines
- We have shown that it is possible to exploit the hub co-occurrence information to build better and more robust neighbor occurrence models for classification.
- We have examined the distribution of neighbor co-occurrences in intrinsically high-dimensional data and have uncovered some interesting regularities.

ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency, the ICT Programme of the EC under Xlike (ICT-STREP-288342) and RENDER (ICT-STREP-257790)

CONTACT

For more information on our work on **hubness**, visit: **http://ailab.ijs.si/nenad_tomasev/**