THE INFLUENCE OF WEIGHTING THE K-OCCURRENCES ON HUBNESS-AWARE CLASSIFICATION METHODS

Nenad Tomašev
Dunja Mladenić
Hubness and NN classification
Weighting the occurrences
The impact on data hubness
The impact on classification
Conclusions and future work
Nearest-neighbor methods in machine learning

- Similarity modeled as proximity in the feature space under some given distance measure

- The general principle: If we want to discover something new about a point, we will consult its $k$ closest nearest neighbors

- This approach is frequently used because of its simplicity
THE CURSE OF DIMENSIONALITY AND HOW IT AFFECTS $k$-NN METHODS

- Learning in many dimensions is often very difficult, since all data is **sparse** and estimates are less reliable.

- The contrast in proximity decreases, so it is hard to tell what is **close** and what is **distant**.

- Also, in high-dimensional data, **hubs** appear.
HUBS: THE INFLUENTIAL NEIGHBORS

- Some points tend to become closer on average to all other points from the same data cluster.

- This tendency gives rise to frequent nearest neighbors, known as hubs.

- Most other points are rarely observed as neighbors and we call them anti-hubs.
Why it matters

Good hubness + Bad hubness = Total hubness
Related work: Hubness-aware classification methods

- Hubness-based weighting to reduce the influence of bad hubs during classification (hw-\(k\)-NN, Radovanović et al., 2009, ICML)
- Hubness induced fuzzy measures in the h-FNN framework (Tomašev et al., 2011, MLDM)
- A naïve Bayesian approach: NHBNN (Tomašev et al., 2011, CIKM)
- An information-theoretic approach HINN (under review)
AN EXAMPLE: HUBNESS-BASED WEIGHTING

- The total number of occurrences is decomposed into:

\[ N_k(x_i) = GN_k(x_i) + BN_k(x_i) \]

\[ h_B(x_i) = \frac{NB_k(x_i) - \mu_{BN_k}}{\sigma_{BN_k}} \]

\[ w_i = e^{-h_B(x_i)} \]

- This was the second baseline (the first was \( kNN \)) in the experiments
THE IDEA

- Many $k$-NN methods use **distance-based weighting** of the neighbor votes.

- This works good because the same value of $k$ might not be appropriate for all the data points, due to **class imbalance**.

- So, if we were to include the occurrence weighting into the hubness-aware model, what would be the result?
Our goal

- Determine how the occurrence weighting would affect both the hubness of the data in general, as well as the performance of the subsequent classification by hubness-aware methods
THE WEIGHTED COUNTS

- Each occurrence is weighted by its relevance to the point of interest.
- The relevance is measured as a distance ratio, given the distance to the nearest neighbor.

\[ WN_k(x_i) = \sum_{x, x_i \in D_k(x)} \frac{d(x, NN(x))}{d(x, x_i)} \]
THE DATA

- UCI (we selected 10 datasets) and ImageNet (we constructed 5 datasets) repositories

- 10-times 10-fold cross-validation was performed when testing the classifiers

- Corrected resampled t-test was used to test for statistical significance

- We compared several recently proposed hubness-aware classification methods — $hw$-$kNN$, $h$-$FNN$, $HIKNN$
AN OBSERVED INCREASE IN HUBNESS

Figure 3: The difference between weighted and non-weighted k-occurrence skewness for datasets from Table 1,
AVERAGE RESULTS (k=30)

- An increase in accuracy was observed for h-FNN and HIKNN.
- A decrease in accuracy was present in hw-kNN, since the bad hubs were given higher weights.

<table>
<thead>
<tr>
<th>kNN</th>
<th>hw-kNN</th>
<th>h-FNN</th>
<th>HIKNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>72.14</td>
<td>74.89</td>
<td>73.86</td>
<td>77.08</td>
</tr>
<tr>
<td>72.41</td>
<td>75.15</td>
<td>77.81</td>
<td></td>
</tr>
</tbody>
</table>
SO, WHERE DOES THE IMPROVEMENT COME FROM?

- Most of the improvement was contained in two datasets: vowel and segment

<table>
<thead>
<tr>
<th>DS</th>
<th>kNN</th>
<th>hFNN</th>
<th>W-hFNN</th>
<th>HIKNN</th>
<th>W-HIKNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vowel</td>
<td>84.3</td>
<td>62.3</td>
<td>75.4</td>
<td>78.4</td>
<td>85.4</td>
</tr>
<tr>
<td>Segment</td>
<td>86.4</td>
<td>79.6</td>
<td>82.7</td>
<td>82.9</td>
<td>86.1</td>
</tr>
</tbody>
</table>

- On those two datasets, kNN was the better classifier, which is very rare, but it does occasionally happen. So, what is the nature of such data?
THE VOWEL DATASET: h-FNN

Figure 2: Accuracies of weighted and non-weighted class hubness implementations of h-FNN for $k = \{2,3,..30\}$ on vowel dataset. The basic kNN is given as a baseline for comparison.
Figure 1: Accuracies of weighted and non-weighted class hubness implementations of HIKNN for $k = \{2,3..30\}$ on vowel dataset. The basic kNN is given as a baseline for comparison.
The conclusion

- Introducing weighting into the occurrence score calculation has an impact on the performance of hubness-aware classification methods.
- The overall hubness of the data is increased, which might be good for subsequent clustering.
- It hampers the hubness-weighted approach and improves the class-hubness-based approaches.
- The overall improvement is small, so such occurrence weighting is of limited use.
- Other occurrence weighting schemes should also be explored.
Thank you for your attention

QUESTIONS?