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

Nenad Tomašev, Dunja Mladenić
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
e-mail: nenad.tomasev@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

Hubness is a phenomenon present in many high-dimensional data sets. It is related to the skewness in the distribution of k-occurrences, i.e. occurrences of data points in k-neighbor sets of other data points. Several hubness-aware methods that focus on exploiting this phenomenon have recently been proposed. In this paper, we examine the potential impact of weighting the *k*-occurrences, by taking into account the distance between the respective data points, on hubness-aware nearest-neighbor methods, more specifically hw-*k*NN, h-FNN and HIKNN. We show that such distance-based weighting can be both advantageous and detrimental and that it influences different methods in different ways.

1 INTRODUCTION

Nearest-neighbor classification methods are often used in machine learning tasks, due to their inherent simplicity and good asymptotic properties. They are based on a notion that similar data points often share the same label. Let $D = \{x_1, x_2, .. x_N\}$ be the training set. 1-NN classification rule is quite simple: given the point of interest *x*, find the point from *D* that is closest to *x* according to some appropriate distance function *d(x, .)*, and denote this point by NN(*x*); assign the label of NN(*x*) to the new point *x*. Due to the common sparsity of data, as well as noisy measurements and records, a generalized *k*NN rule is usually used instead, where a label of a new instance is determined by a majority vote of its *k* nearest neighbors [1].

Many extensions to the basic algorithm have been proposed over the years, offering modifications of various stages of the classification process. Attribute weighting was successfully used in conjunction with nearest-neighbor classification [2]. Fuzzy approaches were also shown to be quite promising [3][4][5]. Recently, large margin *k*NN was introduced, which achieves accuracy comparable to other state of the art classification methods [6].

Hubness is a phenomenon attributed to high-dimensional data which has potentially severe consequences for nearest-neighbor methods [7]. Denote by $N_k(x)$ the number of *k*-occurrences of *x,* i.e. the number of times *x* appears in *k*-neighbor sets of other data points. It has been noted that the distribution of $N_k(x)$ exhibits high skewness when the inherent dimensionality of the data is high. This leads to the emergence of *hubs,* influential data points. Hubs have been shown to appear frequently in many kinds of high-dimensional data, like time series, music and images. Several approaches to exploiting the hubness of the data have recently been proposed [7][8][9][10]. Improvement over the basic *k*NN was frequently present.

2 A BRIEF OVERVIEW OF THE USED HUBNESS-BASED NEAREST-NEIGHBOR CLASSIFICATION METHODS

We can distinguish between two sorts of hubs: the *good hubs* and the *bad hubs*, based on the usefulness of their influence in *k*NN classification. Consequently, we define *good hubness* ($GN_k(x)$) and *bad hubness* ($BN_k(x)$) so that $N_k(x) = GN_k(x) + BN_k(x)$, where $BN_k(x)$ denotes the number of label mismatches between *x* and data points where *x* appears in *k*-neighbor sets. Also, denote by $N_{k,c}(x)$ the number of *k*-occurrences of *x* in neighbor sets of elements belonging to class *c.* We will refer to this quantity as *class hubness*. Three types of approaches to exploiting previous occurrences have been proposed voting by label, voting by class hubness and combined voting.

2.1 VOTING BY LABEL

In hw-*k*NN[7], the basic weighted *k*-nearest neighbor voting framework is retained. Each neighbor votes by its own label and the label weight is determined so as to minimize the influence of bad hubs on classification outcome. The weights are set as $e^{-h_b(x)}$, where $h_b(x)$ is standardized bad hubness. Even this simple weighting scheme was shown to often lead to significant improvements over the basic *k*NN algorithm.

2.2 VOTING BY CLASS HUBNESS

Instead of observing only good and bad hubness, it is possible to take into account class-specific previous *k*-occurrences, i.e. class hubness [8][9]. The h-FNN algorithm is based on this notion and it integrates class hubness information into a fuzzy *k*-nearest neighbor voting framework. It uses a threshold to distinguish between low-hubness points (*anti-hubs*) and medium-to-high hubness points where inference based on class hubness is meaningful. Therefore, it requires a separate mechanism to deal with anti-hubs.

2.3 COMBINED VOTING

Hubness-information *k*-nearest neighbor (HIKNN) [10] is a robust algorithm which uses both the information contained in an instance label and the information contained in its previous occurrences. This approach is based on an information-theoretic perspective, so that the vote of *x* is shifted more towards using class hubness if $N_k(x)$ is high and more towards the label of *x* if $N_k(x)$ is low. The algorithm also weights all the individual fuzzy votes based on their total occurrence frequencies, so that more weight is given to anti-hubs, since they are considered more local to the point of interest and, therefore, more important when trying to determine its label.

3 WEIGHTING THE K-OCCURRENCES

Weighted voting in *k*NN helps in implicitly reshaping the neighborhood to give more emphasis to the closer neighbors. Choosing the proper *k* is far from trivial and sometimes no global neighborhood size gives satisfactory results.

Since this idea is commonly encountered, we wished to see what the effects would be if the same line of reasoning was applied when dealing with inverse neighbor sets. The final voting in h-FNN and HIKNN is also distance-weighted, so introducing some sort of weighting in class hubness calculations does seem somewhat reasonable.

On the other hand, hw-kNN does not employ distance-based weighting. It is based on the simple idea of weighting down the votes of bad hubs. Introducing some weights in the inverse neighbor sets might reduce the bad hubness estimates and increase the voting weights of bad hub points, which may in fact have an overall negative influence on the final classification accuracy. So, intuitively, we would expect to see differences in how these three algorithms change under weighted hubness scores.

We opted for testing a very simple distance-based weighting scheme for calculating class hubness scores. Denote by NN(*x*) the nearest-neighbor of *x*. Let $D_k(x)$ be the *k*-neighborhood of *x*. We define weighted hubness score of x_i as:

$$WN_k(x_i) = \sum_{x:x_i \in D_k(x)} \frac{d(x, NN(x))}{d(x, x_i)}$$

Weighted good and bad hubness ($WGN_k(x)$ and $WBN_k(x)$), as well as weighted class hubness scores ($WN_{k,c}(x_i)$), are defined analogously to their non-weighted counterparts.

4 EXPERIMENTAL SETUP

For small neighborhood sizes, weighted class hubness calculations would have little to no effect, since all neighbors would be close to the points of interest. For larger neighborhoods, the tendency of some neighbors to be much further away than others is amplified. This is why we chose to run all the experiments for a fixed value of k=30.

In order to reduce the influence of a particular anti-hub handling method in h-FNN, we opted for eliminating any such separate case of neighbor handling by setting the threshold value θ in h-FNN to zero. This means that in h-FNN every instance votes purely by class hubness scores from its previous *k*-occurrences. Since every element is by default included in its own neighborhood, $N_k(x) > 0$ for every *x*, which avoids the pathological case of zero hubness.

We selected 15 publicly available datasets, 10 from the UCI data repository and 5 from ImageNet repository [13]. 10-times 10-fold cross validation was performed on every data sets for all the algorithms and the corrected resampled *t*-test was used to check for statistical significance. The results are given in Table 1. . The shortened UCI dataset names in the table correspond to colonTumor, vowel, ecoli, parkinsons, sonar, ionosphere, vehicle, segment, isolet, mfeat-factors. Since UCI data exhibits low-to-medium *k*-occurrence skewness [8][9][10], some high-dimensional and high-skewness data was also included in the experiments. Datasets I-s3 to I-s7 represent five different subsets of images and they are composed of 3,4,5,6 and 7 categories, respectively [8]. The images are represented in a hybrid way, by combining the 400-dimensional quantized SIFT representation [11] and a 16-dimensional color histogram representation. SIFT features [12] capture the local information contained in highly textured image parts and are calculated at certain interesting *keypoints* [14]. Color histograms, on the other hand, capture the global color information. In our experiments, these two representation parts are given equal weight in distance calculations.

The Manhattan metric (sum of absolute differences) was used in all the experiments.

It is clear from the results shown in Table 1 that, most of the time, there is no significant difference if the weighted hubness scores are used. On the other hand, when there is a difference between the weighted and non-weighted implementations, it tends to be detrimental in hw-*k*NN and beneficial in h-FNN and HIKNN as expected.

So, in which cases does using the weighted hubness scores improve the classification result in h-FNN and HIKNN? In order to partially answer this question, we tested the algorithms on the vowel dataset for a range of different neighborhood sizes k = {2,3..30}. The resulting accuracies are shown in Figure 1 and Figure 2.

	*k*NN	hw-*k*NN	whw-*k*NN	h-FNN	wh-FNN	HI-*k*NN	wHI-*k*NN
cTum	72.3	65.4	63.1	62.7	63.4	64.1	64.7
vowel	84.3	57.4	60.3●	62.3	75.4●	78.4	85.4●
ecoli	82.0	85.6	84.4	86.5	86.3	86.3	86.0
psons	90.3	83.3	84.8	84.6	85.3	85.7	85.7
sonar	82.4	73.5	71.5	71.7	72.1	77.3	76.6
ionos	79.7	86.3	82.6	87.5	87.2	87.0	87.6
vehic	61.7	62.4	61.9	59.6	60.6	62.0	62.3
seg	86.4	78.6	78.8	79.6	82.7●	82.9	86.1●
isolet	74.2	85.4	87.5	84.5	85.1	87.2	87.8
mfact	94.5	94.2	93.6	94.0	94.3	94.9	94.9
I-s3	71.2	84.8	80◄	82.7	82.7	84.5	84.4
I-s4	55.4	68.4	63.7◄	63.9	64.1	67.4	67.4
I-s5	45.8	64.3	53.8◄	61.1	61.0	65.4	65.3
I-s6	58.9	70.7	70.2	68.0	68.0	70.9	70.9
I-s7	43.0	63.1	49.9◄	59.2	59.1	62.2	62.1
AVG	72.14	74.89	72.41	73.86	75.15	77.08	77.81

Table 1: *A comparison of classifier accuracies for neighborhood size of k=30. Algorithm implementations using the $WN_k(x_i)$ are given with prefix "w". A filled circle (●) marks those cases where the weighted implementations were significantly better. Inversely, a filled triangle (◄) shows when significant deterioration was observed. Significance level p=0.01 was used in all cases.*

Both vowel and segment (the two datasets where the improvement was observed) are datasets where the classification accuracy deteriorates with increasing neighborhood sizes. The rates of deterioration of *k*NN, h-FNN and HIKNN are not the same, though. The basic *k*NN falls to an accuracy plateu, while the accuracies of the two hubness-based algorithms continue to drop – more steeply in the case of h-FNN. The use of weighted hubness seems to reduce the deterioration rate, resulting in constantly better performance of the weighted implementations, over the entire *k*-range.

If it turns out that this is indeed the only case where one might improve by using the weighted hubness scores, then their usefulness is rather limited, since they do not, in fact, lead to an overall improvement on the dataset, given that the best results are obtained for k = 1 where there is no weighting.

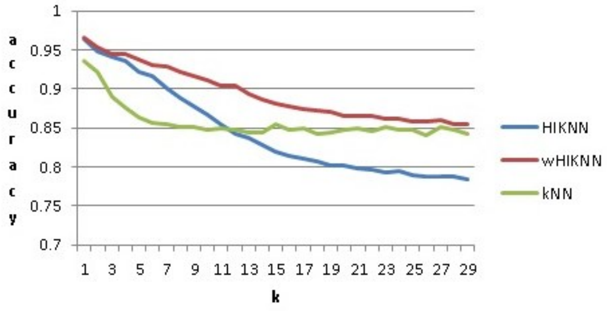

Figure 1: *Accuracies of weighted and non-weighted class hubness implementations of HIKNN for k = {2,3..30} on vowel dataset. The basic kNN is given as a baseline for comparison.*

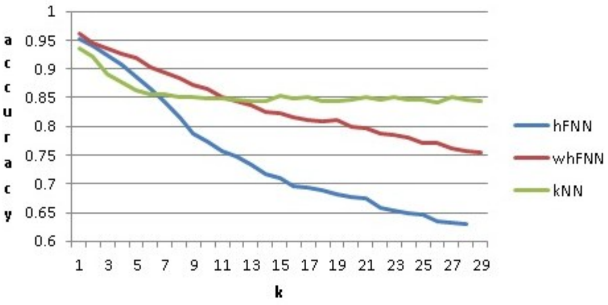

Figure 2: *Accuracies of weighted and non-weighted class hubness implementations of h-FNN for k = {2,3..30} on vowel dataset. The basic kNN is given as a baseline for comparison.*

As for the global influence of weighting on hubness scores, we compared the resulting skewness in *k*-occurrence distributions of weighted and non-weighted occurences. This is illustrated in Figure 3, where the difference between the two skewness values is shown for each of the used datasets. We see that in 14 out of 15 datasets there is a noticeable increase in the *k*-occurrence skewness when weighting is used.

Since any increase in the $N_k(x)$ distribution skewness entails higher hubness of the data, it is clear how this might occasionaly prove beneficial to hubness-based algorithms. On the other hand, even if these algorithms are designed so as to take data hubness into account, this does not imply that, for any specific dataset, they achieve the best performance when the hubness of the data is at the highest point. The observed increase in the skewness may even prove more useful in the unsupervised case, if used for hubness-proportional clustering (HPC) [15].

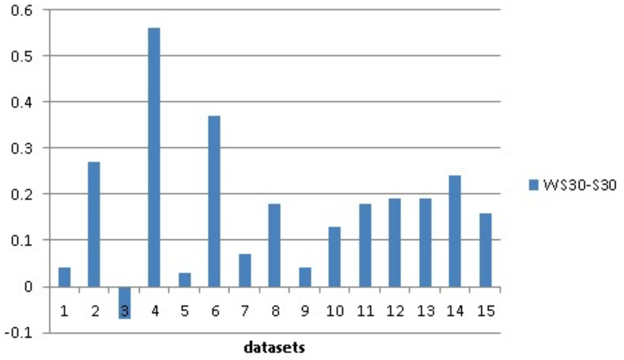

Figure 3: *The difference between weighted and non-weighted k-occurrence skewness for datasets from Table 1, given in the same order (1 – cTum, 2 – vowel, etc.).*

7 CONCLUSION

Data hubness, as a consequence of high inherent dimensionality, is a phenomenon of great importance for nearest-neighbor classification. We have explored how the potential weighting of class hubness scores affects several recently proposed hubness-based algorithms, namely hw-*k*NN, h-FNN and HIKNN. We observed occasional improvements in case of h-FNN and HIKNN, as well as performance deterioration in hw-*k*NN, which is in agreement with our starting hypothesis. We also detected a noticeable increase in *k*-occurrence skewness in the weighted case.

6 ACKNOWLEDGEMENTS

This work was supported by the Slovenian Research Agency, the ICT Programme of the EC under PASCAL2 (IST-NoE-216886) and PlanetData (ICT-NoE-257641).

References

[1] E.Fix and J.Hodges. Discriminatory analysis, nonparametric discrimination: consistency properties, Technical report, USAF School of Aviation Medicine, Randolph Field. Texas. 1951.

[2] E.H. Han and G. Karypis and V. Kumar. Text categorization using weight adjusted k-nearest neighbor classification. *Advances in Knowledge Discovery and Data Mining*. Springer. Berlin. pp. 53–65. 2001.

[3] J.E. Keller and M.R. Gray and J.A. Givens. A fuzzy k nearest-neighbor algorithm. In: IEEE Transactions on Systems, Man and Cybernetics. pp. 580–585. 1985.

[4] R. Jensen and C. Cornelis. A new approach to fuzzy-rough nearest neighbour classification. Rough Sets and Current Trends in Computing. Springer. Berlin. pp. 310–319. 2008.

[5] W. Shang and H. Huang and H. Zhu and Y. Lin and Y. Qu and H. Dong. An adaptive fuzzy knn text classifier. *Computational Science—ICCS* Springer. Berlin. pp. 216–223. 2006.

[6] K.Q. Weinberger and J. Blitzer and L.K Saul. Distance metric learning for large margin nearest-neighbor classification. *Proceedings of the NIPS conference.* MIT Press. 2006.

[7] M. Radovanović and A. Nanopulous. Nearest-neighbors in high-dimensional data: the emergence and influence of hubs. *Proceedings of 26th International Conference on Machine Learning (ICML)* pp. 865-872. 2009.

[8] N. Tomašev and M. Radovanović and D. Mladenić and M. Ivanović. Hubness-based fuzzy measures for high-dimensional k-nearest-neighbor classification. *In Proc. MDLM 2011, 7th International Conf. on Machine Learning and Data Mining*. New York. 2011.

[9] N. Tomašev and M. Radovanović and D. Mladenić and M. Ivanović. A Probabilistic approach to nearest-neighbor classification: Naive Hubness-Bayesian kNN. *In Proc. CIKM*. 2011.

[10] N. Tomašev and D. Mladenić. Nearest-neighbor voting in high dimensional data: learning from past occurrences. (under review)

[11] Z. Zhang and R. Zhang *Multimedia Data Mining*. Chapman & Hall. 2009.

[12] D. Lowe. Object recognition from local scale-invariant features, *Proceedings of the International Conference on Computer Vision.* pp. 1150-1157. 1999.

[13] IMAGENET. http://www.image-net.org/

[14] D. Lowe. SIFT Keypoint Detector. http://www.cs.ubc.ca/~lowe/keypoints/

[15] N. Tomašev and M.Radovanović and D. Mladenić and M. Ivanović. The Role of hubness in clustering high-dimensional data. *In Proc. of PAKDD*. Shenzhen. 2011.