Hub Co-occurrence Modeling for Robust High-dimensional kNN Classification

Nenad Tomašev and Dunja Mladenić

Institute Jožef Stefan
Artificial Intelligence Laboratory
Jamova 39, 1000 Ljubljana, Slovenia
nenad.tomasev@ijs.si, dunja.mladenic@ijs.si

Abstract. The emergence of hubs in k-nearest neighbor (kNN) topologies of intrinsically high dimensional data has recently been shown to be quite detrimental to many standard machine learning tasks, including classification. Robust hubness-aware learning methods are required in order to overcome the impact of the highly uneven distribution of influence. In this paper, we have adapted the Hidden Naive Bayes (HNB) model to the problem of modeling neighbor occurrences and co-occurrences in high-dimensional data. Hidden nodes are used to aggregate all pairwise occurrence dependencies. The result is a novel kNN classification method tailored specifically for intrinsically high-dimensional data, the Augmented Naive Hubness Bayesian k nearest Neighbor (ANHBNN). Neighbor co-occurrence information forms an important part of the model and our analysis reveals some surprising results regarding the influence of hubness on the shape of the co-occurrence distribution in high-dimensional data. The proposed approach was tested in the context of object recognition from images in class imbalanced data and the results show that it offers clear benefits when compared to the other hubness-aware kNN baselines.

Keywords: hubs, k-nearest neighbor, classification, curse of dimensionality, Bayesian, co-occurrences

1 Introduction

The basic k-nearest neighbor classification rule [1] is fairly simple, though often surprisingly effective, as it exhibits some favorable asymptotic properties [2]. Many extensions of the basic method have been proposed over the years. It is possible to use kNN in conjunction with kernels [3], perform large margin learning [4], multi-label classification [5], adaptively determine the neighborhood size [6], etc.

Even though kNN has mostly been replaced in general-purpose classification systems by support vector machines and some other modern classifiers [7], it is still very useful and quite effective in several important domains. Unlike many other methods, kNN has a relatively low generality bias and a rather high specificity bias. This makes it ideal for classification under class imbalance [8][9]. Many real-world class distributions are known to be very imbalanced and many examples can be found in medical diagnostic systems, spam filters, intrusion detection, etc. Nearest neighbor methods are

also currently considered as the state-of-the-art in time series classification when used in conjunction with the dynamic time-warping distance (DTW) [10]. Some recent experiments suggest that the kNN might also be quite appropriate for object recognition in images [11].

The curse of dimensionality [12] is an umbrella-term referring to many difficulties that are known to arise when dealing with high-dimensional feature representations. Many k-nearest neighbor methods are negatively affected by various aspects of the dimensionality curse. Most standard distance measures concentrate [13] and the overall contrast is reduced, which makes distinguishing between close/relevant and distant/irrelevant points difficult for any given query. The very concept of what constitutes nearest neighbors in high-dimensional data has rightfully been questioned in the past [14].

Hubness is a recently described consequence of high intrinsic dimensionality that is related specifically to k-nearest neighbor methods [15]. It was first noticed in music retrieval and recommendation systems [16], where some songs were appearing in the result sets of a surprisingly large proportion of queries [17]. Their occurrence frequency could not be explained by the semantics of the data alone and their apparent similarity to other songs was shown to be quite counter-intuitive. The initial thought was that this might be an artefact of the metric or the feature representation, though it was later shown [15][18] that hubness emerges naturally in most types of intrinsically high-dimensional data. *Hubs* become the centers of influence and the occurrence distribution asymptotically approaches a power law as the dimensionality increases. An illustrative example of the change in the distribution shape is shown in Figure 1. The almost scale-free topology of the k-nearest neighbor graph [18] and the skewed distribution of influence have profound implications for kNN learning under the assumption of hubness in high-dimensional data.

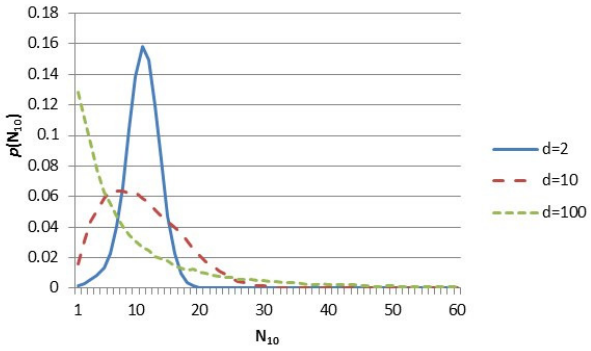


Fig. 1. The shape of the neighbor occurrence frequency distribution changes as the intrinsic dimensionality of the data increases. The example shows the distribution of 10-occurrences (N_{10}) of i.i.d. Gaussian data in case of 2, 10 and 100 dimensions.

The hubness among neighbor occurrences was previously unknown and is not even implicitly taken into account in most standard kNN classifiers. This can lead to some problems in applying the standard methods for high-dimensional data analysis. Pathological cases have even been shown to exist [19][20] where the influence of hubs reduces the overall kNN performance below that of zero-rule. Such cases are rare, but warn of the danger that lurks in ignoring the underlying occurrence distribution.

The presence of hubs could in principle be beneficial in that it reduces the overall impact of noise, but the network of influence also becomes more vulnerable to any inaccuracies that are contained in hubs themselves, when errors can propagate much more easily. Therefore, the overall stability of the learning process is compromised. As hubness is a geometric property that results by an interplay of representational features and metrics, it does not necessarily reflect the underlying semantics well. Many hub points are in fact known to induce severe misclassification [18]. Consequently, there is a rising awareness of a need for novel approaches to algorithm design for properly handling high-dimensional data in k-nearest neighbor methods.

Recent research has shown that learning from past occurrences and hub profiling can be successfully employed for improving the overall kNN classifier performance [21][22][23][24][25]. Hubness-aware metric learning also seems to be helpful [26][27][20]. The consequences of data hubness have recently been examined in the unsupervised context as well [28][29].

The Naive Bayesian interpretation of the observed k-neighbor occurrences (NHBNN) [23] was shown to be quite promising in high-dimensional data classification, especially in the context of learning from class imbalanced data [20]. Yet, NHBNN naively assumes independence between neighbor occurrences in the same k-neighbor set, an assumption that is clearly severely violated in most cases, as close points tend to co-occur as neighbors.

1.1 Goal and Contributions

Our goal was to extend and augment the existing naive NHBNN approach by including the co-occurrence dependencies between the observed neighbors in the underlying Bayesian model. This was done by introducing hidden nodes in the augmented topology, as in the recently proposed Hidden Naive Bayes method [30]. This work represents the first attempt to exploit the neighbor co-occurrence dependencies in high-dimensional neighbor occurrence models and we propose a novel classification algorithm named the Augmented Naive Hubness-Bayesian k-nearest Neighbor (ANHBNN).

Additionally, we justify our approach by examining how the increase in the intrinsic dimensionality of the data affects the distribution of neighbor co-occurrences. Our tests on synthetic Gaussian data reveal some surprising results. We have shown that the distribution of the number of distinct co-occurring neighbor points becomes multi-modal with modes located approximately around the multiples of $(k-1)$. We have also shown that the tail of the distribution of neighbor pair occurrence frequencies becomes thicker with increasing dimensionality, which indicates *hub linkage*, as some hub points tend to co-occur frequently. Also, the number of distinct pairs of co-occurring neighbors increases. These phenomena seem beneficial for co-occurrence modeling and they explain

why the proposed ANHBNN classifier works well in intrinsically high-dimensional data.

2 Related Work

As the emergence of hubs was shown to be potentially highly detrimental, hubness-aware classification of high-dimensional data has recently drawn some attention and several novel kNN classification methods have been proposed. The simplest approach (hw-kNN) was to include instance-specific weights that would reflect the nature of hubness of individual neighbor points [21]. This was later improved upon by including class-conditional occurrence profiles, as in [22][23][25]. In h-FNN [22], the occurrence profiles were used for forming fuzzy votes within the FNN [31] framework. HIKNN [25] was based on the information-theoretic re-interpretation of h-FNN, as less frequently occurring neighbors were judged to be more locally relevant and assigned higher weights, based on their occurrence self-information. On the other hand, the Naive Hubness Bayesian k-nearest Neighbor (NHBNN) [23] was based on a slightly different idea - interpreting the individual occurrences as random events and applying the Naive Bayes rule in order to perform classification. Prior tests have shown this to be a promising idea, so extending this basic approach will be the focus of this paper.

2.1 Naive Hubness-Bayesian *k*NN

In order to explain in detail the idea behind the Naive Hubness-Bayesian kNN (NHBNN) [23], it is necessary to introduce some formal notation.

Neighbor *k*-occurrence Models: Let $D = \{(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)\}$ be the data set, where x_i-s are the feature vectors and $y_i \in \{1 \dots C\}$ the class labels. Also, let $D_k(x)$ be the set of k-nearest neighbors of x. The neighbor k-occurrence frequency of x will be denoted by $N_k(x) = |x_i : x \in D_k(x_i)|$ and will also sometimes be referred to as *point hubness*. The total *hubness of a dataset* D is defined as the third standard moment (skewness) of the neighbor occurrence degree distribution and will be denoted by $SN_k = \frac{\frac{1}{n}\sum_{i=1}^{n}(N_k(x_i)-k)^3}{(\frac{1}{n}\sum_{i=1}^{n}(N_k(x_i)-k)^2)^{3/2}}$. High skewness indicates the long-tailed distribution where most k-neighbor sets are dominated by occurrences of a limited number of highly frequent neighbors, while most other points occur very rarely or not at all. The very frequently occurring points are called *hubs*, the infrequently occurring points *anti-hubs* and the points that never occur as neighbors *orphans*.

The total occurrence frequency is often decomposed into either *good and bad hubness* or alternatively *class-conditional hubness* in the following way: $N_k(x) = GN_k(x) + BN_k(x) = \sum_{c \in C} N_{k,c}(x)$. Good hubness is defined as the number of neighbor occurrences where neighbors share the same class label and bad hubness the number of occurrences where there is label mismatch, i.e. $GN_k(x) = |x_i : x \in D_k(x_i) \wedge y = y_i|$ and $BN_k(x) = |x_i : x \in D_k(x_i) \wedge y \neq y_i|$. Similarly, class-conditional hubness measures the occurrence frequency within the neighbor sets of a specific class:

$N_{k,c}(x) = |x_i : x \in D_k(x_i) \wedge y_i = c|$. These quantities are used to form an occurrence model from the training set that includes a neighbor occurrence profile for each neighbor point.

Naive Bayesian Interpretation of *k*NN: In the NHBNN method [23], the neighbor occurrences are interpreted as random events that can be used to deduce the class label in the point of interest. Equation 1 shows how the Naive Bayes rule [7] is used to deduce the class assignment probabilities for X based on its k-nearest neighbors from the training set. The label $y = \mathrm{argmax}_{c \in \{1 \ldots C\}} p(c|D_k(x))$ is assigned to x by this rule.

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

The order of neighbors is ignored in order to get better probability estimates in the model, so $p(X_t \in D_k(X)|Y)$ can be easily estimated from the class-specific hubness $N_{k,c}$ scores and the total class occurrences. Each point is trivially considered to be its own 0th nearest neighbor, for practical reasons.

$$p(X_t \in D_k(X)|Y) \approx \frac{N_{k,Y}(X_t) + \lambda}{n_Y \cdot (k+1) + \lambda|D|} \tag{2}$$

The actual algorithm is a bit more complex than this, mostly because there are points for which the $p(X_t \in D_k(X)|Y)$ can not be reliably estimated from the previous occurrences, orphans and anti-hubs. They need to be treated separately, as a special case. In analogy with the Naive Bayes classifier, it would be as if a completely new feature value was first encountered on the test data.

The obvious problem with this approach is that the Naive Bayes rule assumes independence between the random variables and this does not hold true among the k-neighbor occurrences, where close neighbors tend to co-occur together and there are clear dependencies between individual neighbor occurrences.

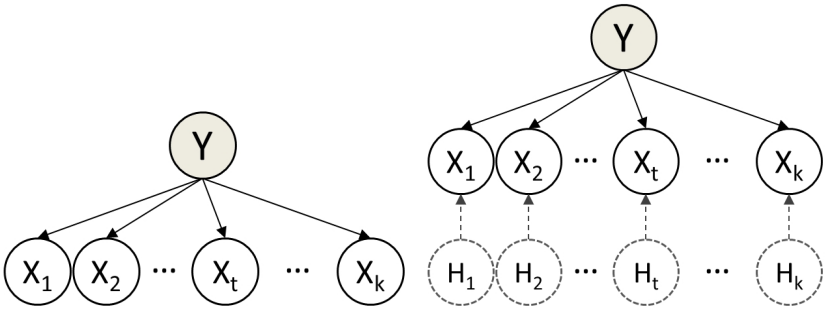
Naive Bayes sometimes works well even when the independence assumption does not hold [32] and the initial evaluation of the Naive Hubness-Bayesian k-nearest neighbor has shown it to be quite a promising approach to high-dimensional kNN classification. However, it was later observed that its performance quickly drops when the neighborhood size is increased and that it performs rather poorly for larger k values. It was hypothesized that this was a consequence of the independence assumption violation.

In order to test this hypothesis, we decided to proceed by including some sort of co-occurrence dependencies in the model, with the intent of increasing its robustness and overall performance. The extended algorithm was supposed be able to properly handle larger neighborhood sizes.

3 The Proposed Approach: Including the Co-occurrence Dependencies

Naive Bayes is the simplest among the Bayesian network models. The conditional independence assumption is often violated in practice, though its use can still be justified in

some cases [32]. Learning the optimal Bayesian network from the data can sometimes be intractable, as it was shown to be an NP-complete problem [33]. As the structure learning is the most time consuming step, assuming a certain type of underlying structure is common. We base our extension of the hubness-aware NHBNN classifier on the Hidden Naive Bayes model [30], shown in Figure 3. A hidden node is introduced for each variable that accounts for the influence from all other variables. In our case, the variables are the occurrences of points as neighbors in k-neighbor sets.



(a) The basic Naive Bayes model. All occurrences are conditioned only on the class label Y.

(b) The Hidden Naive Bayes model. An additional hidden node is introduced for each neighbor variable, modeling the dependencies on all other nodes.

Fig. 2. A comparison between the basic Naive Bayes and the Hidden Naive Bayes [30] models.

Hidden nodes help model the dependencies between neighbor co-occurrences.

Let $N_{k,c}(x_i, x_j)$ be the number of co-occurrences of x_i and x_j in neighborhoods of elements from c, i.e. $N_{k,c}(x_i, x_j) = |x : y = c \wedge x_i \in D_k(x) \wedge x_j \in D_k(x)|$. Calculating all the $N_{k,c}(x_i, x_j)$ paired class-conditional co-occurrence frequencies is possible in $O(nk^2)$, as this is the time required to consider and count all co-occurrences within the k-neighbor sets. In order to avoid the $O(Cn^2)$ memory complexity for storing all the co-occurrence counts, C hash tables can be used to store only the non-negative co-occurrence counts. Many $N_{k,c}(x_i, x_j)$ do equal zero, so this saves considerable memory space.

Classification in the extended Bayesian neighbor occurrence model is performed based on the class probability estimate shown in Equation 3 and it forms a similar expression as in NHBNN (Equation 1). The difference is that the probability of $X_t \in D_k(X)$ is now also conditioned on the hidden variable $H_t(X, Y)$.

$$p(Y|D_k(X)) \propto p(Y) \prod_{t=1}^{k} p(X_t \in D_k(X)|Y, H_t(X, Y)) \tag{3}$$

We will call the proposed algorithm that performs the k-nearest neighbor classification based on Equation 3 the Augmented Naive Hubness-Bayesian k-nearest Neighbor (ANHBNN).

3.1 Modeling the Influence of Hubs and Regular Points

In order to infer reliable probability estimates, a certain number of observed occurrences is required. We will first derive the estimates for frequent neighbor points and then focus on approximations for anti-hubs and orphans.

Assuming $N_k(X_t) > 0$, the conditional probabilities are expressed as a weighted sum of separate one-dependence estimators, as shown is Equation 4. This is a standard approach to modeling the influence of the hidden nodes within the HNB framework [30].

$$\begin{aligned} p(X_t \in D_k(X)|Y, H_t(X,Y)) &= \sum_{i=1, i\neq t}^{k} w_{it}^{Y} \cdot p(X_t \in D_k(X)|X_i \in D_k(X), Y) \\ p(X_t \in D_k(X)|X_i \in D_k(X), Y) &\approx \begin{cases} \frac{N_{k,Y}(X_t, X_i)}{N_{k,Y}(X_i)}, & \text{if } N_{k,Y}(X_i) > 0, \\ 0, & \text{if } N_{k,Y}(X_i) = 0. \end{cases} \end{aligned} \tag{4}$$

The weights in Equation 4 sum up to one and correspond to the strengths of individual influences. It is possible to try optimizing the weights via cross-validation, but it is overly time-consuming and is usually avoided. We propose to extend the original idea [30] of expressing the weights by normalized mutual information by including the class-conditional occurrence self-information $I_{k,Y}(X_i)$ (Equation 5) and the occurrence profile non-homogeneity (Equation 6) which is expressed as the reverse neighbor set entropy. These quantities are supposed to account for differences in hubness between different points.

The class-conditional occurrence self-information measures how unexpected it is to observe X_i in neighborhoods of class Y. Including the self-information in the denominator in Equation 8 allows us to increase the influence of very frequent neighbors. This is beneficial, as there is more past occurrence data for these points and the probability estimates are thus somewhat more reliable. On the other hand, neighbor points with less homogenous occurrence profiles often act as bad hubs and exhibit a detrimental influence, so favoring neighbors with homogenous profiles tends to improve the overall performance.

$$I_{k,Y}(X_i) = \log \frac{n_Y}{N_{k,Y}(X_i)} \tag{5}$$

$$H_k(X_i) = \sum_{c \in C} \frac{N_{k,c}(X_i)}{n_c} \log \frac{n_c}{N_{k,c}(X_i)} \tag{6}$$

The class-conditional mutual information $I_P(X_j, X_t|Y)$ between two neighbor occurrences X_j and X_t is estimated based on the previously observed occurrence profiles on the training data as outlined in Equation 7. The four factors in the outer sum correspond to the two neighbor points occurring together or separately or not at all.

$$I_P(X_j, X_t|Y) = \sum_{c=1}^{C} \left(\frac{N_{k,c}(X_j, X_t)}{n} \cdot \log \frac{\frac{N_{k,c}(X_j,X_t)}{n_c}}{\frac{N_{k,c}(X_j)}{n_c} \cdot \frac{N_{k,c}(X_t)}{n_c}} \right) +$$
$$+ \sum_{c=1}^{C} \left(\frac{N_{k,c}(X_j) - N_{k,c}(X_j, X_t)}{n} \cdot \log \frac{\frac{N_{k,c}(X_j)-N_{k,c}(X_j,X_t)}{n_c}}{\frac{N_{k,c}(X_j)}{n_c} \cdot (1 - \frac{N_{k,c}(X_t)}{n_c})} \right) +$$
$$+ \sum_{c=1}^{C} \left(\frac{N_{k,c}(X_t) - N_{k,c}(X_j, X_t)}{n} \cdot \log \frac{\frac{N_{k,c}(X_t)-N_{k,c}(X_j,X_t)}{n_c}}{(1 - \frac{N_{k,c}(X_j)}{n_c}) \cdot \frac{N_{k,c}(X_t)}{n_c}} \right) +$$
$$+ \sum_{c=1}^{C} \left(\frac{n_c - N_{k,c}(X_j) - N_{k,c}(X_t) + N_{k,c}(X_j, X_t)}{n} \cdot \log \frac{\frac{n_c-N_{k,c}(X_j)-N_{k,c}(X_t)+N_{k,c}(X_j,X_t)}{n_c}}{(1 - \frac{N_{k,c}(X_j)}{n_c}) \cdot (1 - \frac{N_{k,c}(X_t)}{n_c})} \right) \tag{7}$$

Finally, the co-dependency weights from Equation 4 are obtained from the class-conditional occurrence self-information, homogeneity and class-conditional neighbor mutual information as shown in Equation 8. Unlike in the original Hidden Naive Bayes [30] model, the weights here are also conditioned on the class, because of the class-conditional self-information. Some smoothing is needed in order to avoid zero divisions in cases when the denominator goes to zero.

$$w_{it}^{Y} = \frac{\frac{I_P(X_i,X_t|Y)}{I_{k,Y}(X_i) \cdot H_k(X_i)}}{\sum_{j=1,j\neq t}^{k} \frac{I_P(X_j,X_t|Y)}{I_{k,Y}(X_j) \cdot H_k(X_i)}} \tag{8}$$

The proposed extension of NHBNN embodied in the Augmented Naive Hubness-Bayesian k-nearest Neighbor (ANHBNN) does not have a significant impact on the overall computational complexity, as both algorithms are of the $O(n^2)$ complexity with respect to data size. Approximate k-neighbor set computations are possible and usually allow for considerable practical speed-ups in hubness-aware classifiers without sacrificing too much accuracy [25].

3.2 Dealing with Anti-Hubs and Orphans

For infrequently occurring points X_t, the $p(X_t \in D_k(X)|Y, H_t(X,Y))$ can not be estimated from their past occurrences properly. In principle, it would be possible to model their conditioned influence by the average conditioned influence exhibited by other points from their class, as in Equation 9.

$$p(X_t \in D_k(X)|Y, H_t(X,Y)) \approx \frac{\sum_{X_i:Y_i=Y_t \wedge N_k(X_i)>0} p(X_i \in D_k(X)|Y, H_i(X,Y))}{|X_i : Y_i = Y_t \wedge N_k(X_i) > 0|} \tag{9}$$

However, the exact $p(X_i \in D_k(X)|Y, H_i(X,Y))$ are not by default calculated during training, as they depend on the particular k-neighbor set and are inferred later from the pre-calculated one dependence estimators, mutual information and self-information. Therefore, approximating the influence of anti-hubs this way would require an additional time-consuming pass through the training data, as well as some initialization of $p(X_t \in D_k(X)|Y, H_i(X,Y))$ for anti-hubs anyway.

Luckily, points that never occur as neighbors on the training data very rarely occur as neighbors on the test data as well, so it is possible to employ very simple replacements in place of

the actual conditional estimates, as it is not possible to arrive at a reliable proper estimate anyway [22][23][25]. As hubs account for most occurrences, this does not have a significant influence on the algorithm performance. Therefore, we propose to use the hidden nodes only for regular points and hubs and approximate the influence of anti-hubs and orphans by the average class-to-class occurrence probabilities as in Equation 10. Here $N_{k,Y}(Y_t)$ denotes the total number of occurrences of elements from class Y_t in neighborhoods of elements from Y. A similar global anti-hub modeling approach was previously shown to be acceptable in NHBNN [23].

$$N_k(X_t) = 0 : p(X_t \in D_k(X)|Y, H_t(X,Y)) \approx$$
$$p(X_t \in D_k(X)|Y) \approx \text{AVG}_{Y_i=Y_t} p(X_i \in D_k(X)|Y) = \frac{N_{k,Y}(Y_t)}{k \cdot n_Y \cdot n_{Y_t}} \tag{10}$$

4 Neighbor Co-occurrences in High-dimensional Data

We hypothesized that the emergence of hubs in the kNN topologies of intrinsically high-dimensional data might have some influence on the distribution of neighbor co-occurrences. As our proposed hubness-aware classifier learns from the observed co-occurrences, we have run extensive tests in order to establish whether the hypothesis holds.

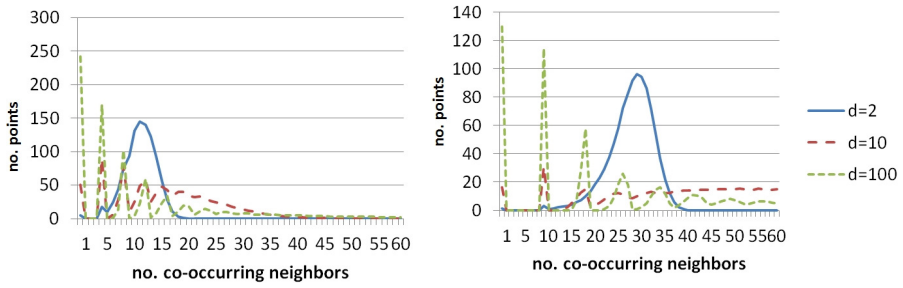
To our knowledge, no previous research has been done on the impact of high intrinsic dimensionality on the neighbor co-occurrence distribution and its connection to the hubness phenomenon. Therefore, we hope that the results presented here might shed some light on the more subtle consequences of the curse of dimensionality.

We have run the tests for three different dimensionalities: 2, 10 and 100. For each number of dimensions, a series of 200 randomly generated hyper-spherical zero-centered Gaussian distributions was generated and 1000 points were randomly drawn from each distribution as sample data. We have run tests for several different neighborhood sizes and we give the results for $k = 5$ and $k = 10$ here for comparison.

Figure 3 shows how the number of distinct neighbors that points co-occur with changes with increasing dimensionality. For $d = 2$, the distribution of the number of distinct co-occurring neighbors has a single mode. However, surprisingly, when the number of dimensions is increased, multiple modes appear and are centered approximately around the multiples of $(k - 1)$. We believe that this is a direct consequence of hubness, as there are many points in intrinsically high-dimensional data that occur in k-neighbor sets very rarely. When these points do occur as neighbors, it is possible that most of their $(k - 1)$ co-neighbors co-occur with the anti-hub point for the first time, hence the observed distribution modes.

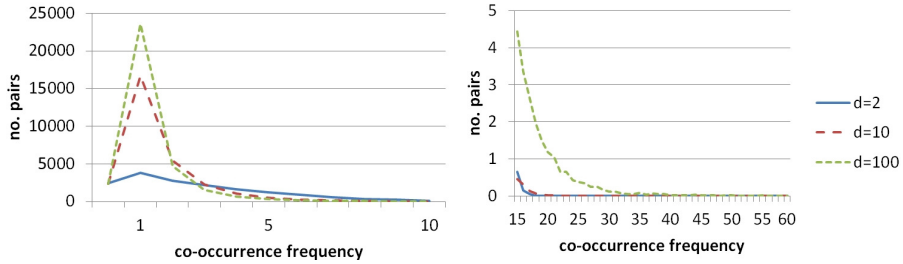
The emergence of hubs (and anti-hubs) also influences the distribution of the co-occurrence frequency of pairs of neighbor points, as shown in Figure 4. The number of very rarely co-occurring pairs increases significantly with increasing dimensionality, due to a large number of rarely occurring neighbor points. On the other hand, the distribution tail also becomes thicker, as the number of pairs of points that co-occur very frequently increases with increasing dimensionality. These very frequently co-occurring pairs emerge as a consequence of what we will denote as *hub linkage*, pairs of hub points that co-occur together in many k-neighbor sets. The linked hub pairs enable the proposed ANHBNN classifier to infer more reliable class-conditional co-occurrence estimates, which is an essential part of the model.

The overall number of distinct co-occurring pairs of neighbor points increases with increasing dimensionality, as shown in Figure 5. From the perspective of co-occurrence modeling, this is a good thing. It is therefore expected that there would be more pairs of neighbor points for which we would be able to derive some estimates of co-occurrence dependencies in intrinsically high-dimensional data.



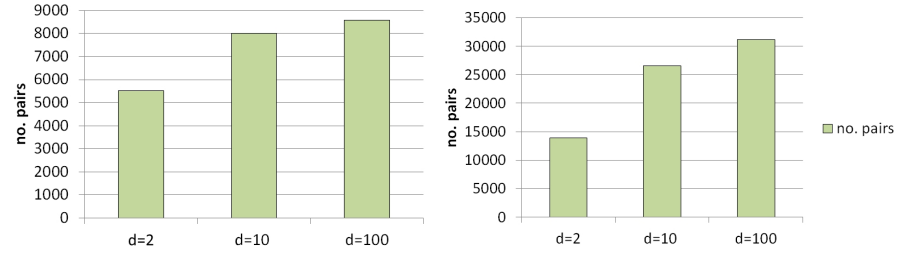
(a) k=5 (b) k=10

Fig. 3. The influence of increasing dimensionality on the distribution of number of different neighbors that points co-occur with. The distribution shape changes from a single modal to a multi-modal shape that has modes around multiples of $(k-1)$.



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

Fig. 4. The influence of increasing dimensionality on the distribution of co-occurrence frequency of pairs of neighbor points. The high-dimensional case shows two extremes: more very rarely co-occurring pairs and also more very frequently co-occurring pairs in the distribution tail. The results are given for $k = 10$.



(a) $k = 5$ (b) $k = 10$

Fig. 5. Increasing the intrinsic dimensionality of the data increases the number of distinct co-occurring neighbor pairs.

5 Experimental Evaluation

In order to evaluate whether the proposed approach offers any benefits, we have compared it with the other hubness-aware classifiers, namely NHBNN [23], hw-kNN [21], h-FNN [22] and

HIKNN [25], as well as with the baseline kNN. Comparisons were performed on a series of intrinsically high-dimensional datasets that have been shown to exhibit very high hubness.

5.1 Data

In experimental evaluation, we have focused on the task of object recognition from images. Image data is high-dimensional and known to exhibit significant hubness [19]. The basic properties of the datasets are outlined in Table 1. Some of the data is imbalanced and the class imbalance is measured by the relative imbalance factor $\mathrm{RImb} = \sqrt{(\sum_{c \in C} (p(c) - 1/C)^2)/((C-1)/C)}$, which is merely the normalized standard deviation of the class probabilities from the absolutely homogenous mean value of $1/c$.

Datasets iNet3-iNet7 and iNet3Imb-iNet7Imb represent different subsets of the public ImageNet repository (http://www.image-net.org/). These particular subsets have previously been used in several hubness-aware classification benchmarks [22][19][24][20], so they have been selected here for easier comparisons. Images were processed as quantized SIFT [34] bag-of-visual-words representations, extended by binned color histogram information, normalized to the $[0, 1]$ range. This sort of feature representation is known to be quite prone to hubness [19].

Datasets WiM1-WiM5 represent five non-trivial imbalanced binary classification problems defined on top of the WIKImage data [35], a set of publicly available images crawled from Wikipedia (http://www.wikipedia.org/). These images are available along with the associated text and their labels. We present the results on the textual data obtained from the labels, represented in a standard bag-of-words format, weighted by TF-IDF. The five selected datasets correspond to the presence/absence of following types of objects in the images: buildings and constructions, documents and maps, logos and flags, nature and scenic, sports.

Table 1. The summary of high-hubness datasets. Each dataset is described both by a set of basic properties (size, number of features, number of classes) and some hubness-related quantities for two different neighborhood sizes, namely: the skewness of the k-occurrence distribution (S_{N_k}), the percentage of *bad* k-occurrences (BN_k), the degree of the largest hub-point ($\max N_k$). Also, the relative imbalance of the label distribution is given [20], as well as the size of the majority class (expressed as a percentage of the total)

Data set	size	d	C	$S_{N_{15}}$	BN_{15}	$\max N_{15}$	RImb	$p(c_M)$
iNet3	2731	416	3	9.27	29.7%	901	0.40	50.2%
iNet4	6054	416	4	8.99	48.9%	968	0.14	35.1%
iNet5	6555	416	5	12.10	57.2%	1888	0.20	32.4%
iNet6	6010	416	6	14.26	44.4%	1901	0.26	30.9%
iNet7	10544	416	7	12.29	59.2%	1741	0.09	19.2%
iNet3Imb	1681	416	3	2.22	18.8%	136	0.72	81.5%
iNet4Imb	3927	416	4	5.44	40.5%	374	0.39	54.1%
iNet5Imb	3619	416	5	7.35	44.4%	513	0.48	58.7%
iNet6Imb	3442	416	6	3.93	44.2%	268	0.46	54.0%
iNet7Imb	2671	416	7	4.35	45.6%	301	0.46	52.1%
WiM1	1007	3182	2	12.31	36.9%	997	0.26	62.8%
WiM2	1007	3182	2	12.31	7.0%	997	0.84	92.1%
WiM3	1007	3182	2	12.31	37.3%	997	0.91	95.7%
WiM4	1007	3182	2	12.31	22.4%	997	0.60	79.9%
WiM5	1007	3182	2	12.31	4%	997	0.93	96.9%
AVG	3484.6	1338	4	9.45	36.03%	931.73	0.48	59.70%

The quantities shown in Table 1 illustrate the consequences of high dimensionality and the hubness phenomenon. Neighbor k-occurrence distribution skewness is considerable, as anything above $SN_k = 1$ is usually considered high-hubness data [18]. The most frequently occurring hub points dominate and appear in unexpectedly many k-neighbor sets. For instance, the major hub on iNet3 data appears in about 30% of all neighbor sets for $k = 15$, while the major hub in WiM1 appears in nearly all neighbor sets, 997 out of 1007 for $k = 15$. The situation is somewhat more bearable for smaller neighborhood sizes in a sense that the major hubs cover fewer neighbor sets, but the overall occurrence skewness is usually higher.

Removal of such frequently occurring hub-points is possible, but their positions in the k-neighbor sets are taken by other points and this often leads to emergence of new hubs and they exhibit their own detrimental influence on data analysis. Reducing the hubness of the data is, in general, a difficult task, though certain feature types, metrics and normalization methods are known to be somewhat less prone to the dimensionality curse [19]. As there is no guarantee that the preprocessing would significantly reduce the overall hubness of the data, robust hubness-aware learning methods are to be preferred.

5.2 Classification Experiments

All experiments and classifier comparisons were run as 10-times 10-fold cross-validation. Corrected re-sampled t-test was used to determine statistical significance. The L_1 Manhattan distance was used to measure the dissimilarity between quantized image pairs and cosine similarity to determine the distance between textual feature vectors.

All algorithms were run with standard parameter configurations, as given in the respective papers. As some datasets exhibit class imbalance, the macro-averaged F_1 score, denoted by F_1^M, was used to measure classifier performance [7]. The summary of results for neighborhood size $k = 15$ is given in Table 2. In principle, ANHBNN requires slightly larger neighborhood sizes, as it provides it with more co-occurrence information. Trivially, for $k = 1$, there would be no co-occurrences at all. The algorithm also performs rather poorly for $k = 2$ or $k = 3$, which is understandable. However, as the results show, it achieves very good results for larger k values.

This is not the case with NHBNN, as it was already noticed that its performance drops significantly with increasing neighborhood size, as the independence assumption between different neighbor occurrences becomes more severely violated. As this is what ANHBNN aims at improving, the neighborhood size of $k = 15$ was used in most experiment runs. A more detailed comparison of algorithm performance under varying neighborhood size is shown in Figure 6, demonstrating that the performance of the proposed approach is not very sensitive to the choice of k, once it exceeds some lower threshold value. Its performance remains stable when k is increased, suggesting that it succeeds in modeling the hub co-occurrence dependencies.

The results in Table 2 suggest that the proposed ANHBNN does indeed outperform NHBNN in the evaluated context. Furthermore, it achieves the best overall F_1^M score on the examined data. Table 3 provides a summary of pairwise classifier comparisons by showing the number of wins and statistically significant wins in each individual comparison. The proposed approach achieves the highest number of wins against any given baseline, as well as the highest total number of wins (67) and statistically significant wins (63).

Even though these results seem quite encouraging, some caution is still required when comparing different approaches. Namely, both NHBNN and ANHBNN assume high underlying hubness of the data and are not well suited for applications on datasets that exhibit low hubness or no hubness at all. In that sense, they are not general-purpose classification algorithms. Instead, they are tailored specifically for classifying intrinsically high-dimensional data. This is not the case with h-FNN, hw-kNN or HIKNN. Even though these remaining three methods are hubness-aware, they perform rather well even when the data exhibits only low to moderate k-occurrence

Table 2. An overview of algorithm performance for $k = 15$. The macro-averaged F-score F_1^M percentage is given for Augmented Naive hubness-Bayesian kNN (ANHBNN), Naive hubness-Bayesian kNN (NHBNN), kNN, hubness-weighted kNN (hw-kNN), hubness-based fuzzy nearest neighbor (h-FNN) and hubness information k-nearest neighbor (HIKNN). The symbols •/○ denote statistically significant worse/better performance ($p < 0.05$) compared to ANHBNN. The best result in each line is in bold.

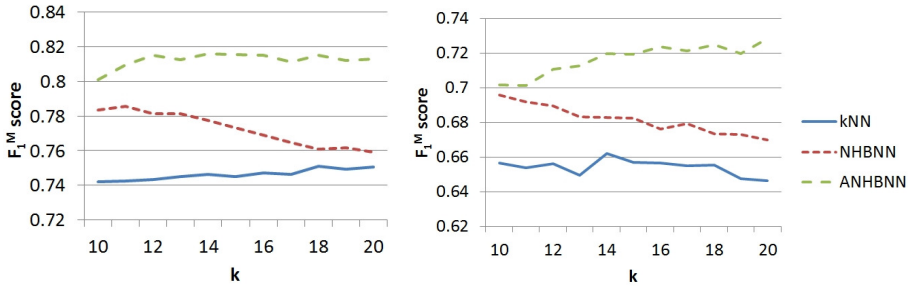
Data set	**ANHBNN**	NHBNN	k**NN**	hw-kNN	h-FNN	HIKNN
iNet3	**81.1 ± 1.1**	77.3 ± 1.6 •	74.7 ± 1.7 •	78.3 ± 2.4 •	78.4 ± 1.7 •	80.3 ± 1.3 •
iNet4	65.9 ± 1.3	63.3 ± 1.4 •	62.4 ± 1.5 •	65.5 ± 1.7	63.4 ± 1.5 •	**66.9 ± 1.4** ○
iNet5	**62.8 ± 1.2**	59.8 ± 1.3 •	47.5 ± 1.3 •	56.1 ± 2.5 •	53.7 ± 1.6 •	59.3 ± 1.3 •
iNet6	56.1 ± 1.3	**57.0 ± 1.4** ○	56.2 ± 1.2	56.4 ± 1.3	51.3 ± 1.5 •	56.2 ± 1.2
iNet7	**59.9 ± 1.3**	56.3 ± 0.9 •	45.3 ± 1.0 •	55.5 ± 2.8 •	56.9 ± 1.0 •	59.1 ± 0.8 •
iNet3Imb	**71.9 ± 1.4**	67.6 ± 2.1 •	65.9 ± 2.0 •	65.3 ± 1.4 •	55.0 ± 1.6 •	64.7 ± 1.3 •
iNet4Imb	**67.1 ± 1.6**	60.1 ± 1.5 •	56.7 ± 1.4 •	57.9 ± 1.6 •	45.2 ± 1.5 •	54.6 ± 1.5 •
iNet5Imb	**56.8 ± 1.6**	52.7 ± 1.8 •	35.3 ± 1.9 •	43.2 ± 1.9 •	31.1 ± 1.6 •	38.1 ± 1.6 •
iNet6Imb	52.8 ± 1.3	52.4 ± 1.5	49.2 ± 1.6 •	52.7 ± 1.6	50.5 ± 1.7 •	**54.1 ± 1.4** ○
iNet7Imb	**47.8 ± 1.3**	46.1 ± 1.2 •	33.3 ± 1.9 •	44.0 ± 2.1 •	35.7 ± 2.1 •	42.4 ± 2.2 •
WiM1	**69.1 ± 2.8**	64.4 ± 2.7 •	66.4 ± 2.2 •	53.9 ± 3.5 •	46.0 ± 3.1 •	54.3 ± 2.8 •
WiM2	75.2 ± 1.2	**75.7 ± 1.1**	58.1 ± 1.3 •	72.7 ± 1.2 •	69.1 ± 1.1 •	68.5 ± 1.2 •
WiM3	**72.1 ± 1.4**	72.0 ± 1.5	59.5 ± 1.3 •	67.6 ± 1.7 •	69.9 ± 1.3 •	**72.1 ± 1.4**
WiM4	**71.8 ± 3.0**	70.0 ± 2.8 •	69.8 ± 2.7 •	62.7 ± 2.9 •	54.1 ± 3.1 •	56.8 ± 2.6 •
WiM5	**54.2 ± 2.9**	49.9 ± 2.7 •	49.2 ± 2.7 •	49.2 ± 2.7 •	49.2 ± 2.7 •	49.2 ± 2.7 •
AVG	**64.30**	61.64	55.30	58.73	53.96	58.26

Table 3. Pairwise comparison of classifiers on the examined data: number of wins (with the statistically significant ones in parenthesis)

	ANHBNN	NHBNN	kNN	hw-kNN	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)
kNN	1 (0)	1 (1)	–	3 (2)	6 (6)	4 (4)	15 (13)
hw-kNN	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)

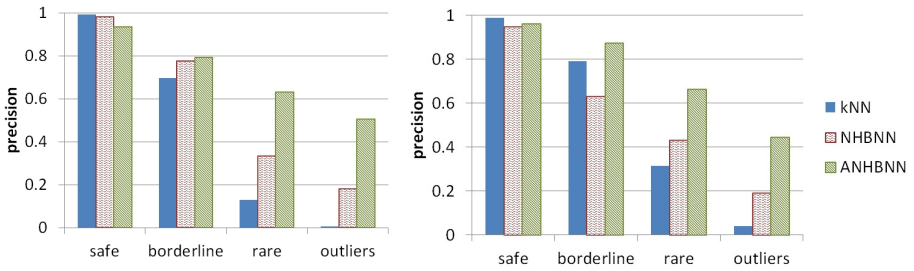
distribution skewness [25]. In our initial experiments, we have determined that HIKNN is to be preferred in such cases, as for example on UCI datasets (http://archive.ics.uci.edu/ml/datasets.html).

In order to examine the nature of the observed differences in performance on the test data, we have analyzed the precision that the algorithms achieve on certain types of points. Not all points are equally hard to classify by k-nearest neighbor methods and a point characterization scheme based on the proportion of label mismatches in k-neighbor sets was recently proposed [36]. Four different point types were observed: *safe* points, *borderline* points, *rare* points and *outliers*, the latter being much more difficult to handle. A comparison between kNN, NHBNN and ANHBNN on two different datasets is shown in Figure 7. The proposed approach clearly outperforms NHBNN in terms of rare point and outlier classification precision and also achieves a slightly higher precision when classifying borderline points. In other words, ANHBNN achieves its improvements by being able to better handle very difficult points that lie far away from class interiors. This is a highly desired property.



(a) iNet3 (b) iNet3Imb

Fig. 6. The influence of increasing the neighborhood size k. Neighbor occurrence dependencies induce a drop in NHBNN performance, while the ANHBNN performance slowly increases with additional neighbor occurrence and co-occurrence information.



(a) iNet3 (b) iNet3Imb

Fig. 7. Precision achieved by the classification algorithms on specific types of points: safe points, borderline points, rare points and outliers.

6 Conclusions and Future Work

Hubness is an important aspect of the dimensionality curse that affects most k-nearest neighbor methods in severely negative ways, as hub points tend to dominate the k-neighbor sets and induce many label mismatches. Hubness-aware classification methods are required in order to properly deal with the emerging hubs.

We have proposed an extension of one such hubness-aware kNN classifier and have named it the Augmented Naive Hubness-Bayesian k-nearest neighbor (ANHBNN). The previous approach (NHBNN) failed to take the neighbor co-occurrences into account, which led to poor performance for larger neighborhood sizes. Our proposed approach (ANHBNN) overcomes this issue by adapting the Hidden Naive Bayes model to the problem of modeling neighbor k-occurrences. We have also proposed a novel set of hubness-aware weights for combining the one-dimensional estimators in the model.

We have performed an analysis of the high-dimensional neighbor co-occurrence distributions for Gaussian mixture data. The analysis has revealed several surprising facts. The distribution of the number of distinct co-occurring neighbor points becomes multi-modal with modes located approximately around the multiples of $(k-1)$. Additionally, there seems to be a phenomenon of *hub linkage*, as the tail of the co-occurrence frequency distribution becomes thicker with in-

creasing dimensionality, indicating that some pairs of hub points co-occur frequently. The overall number of distinct co-occurring pairs also increases, which allows us to estimate more pairwise dependencies in high-dimensional data.

Our evaluation in the context of object recognition from images shows that the proposed approach clearly outperforms the compared baselines and offers additional benefits in achieving higher precision when classifying points that lie far from class interiors and are otherwise difficult to handle. Unlike NHBNN, the performance of the proposed ANHBNN classifier does not decrease when the neighborhood size k is increased, which was the main issue with the previous approach.

As many of the co-occurrence dependencies are somewhat difficult to estimate directly from the occurrence data, in our future work we intend to explore the possibilities for using the Poisson processes for neighbor occurrence modeling, in order to try and achieve a more robust k-occurrence model.

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

References

1. Fix, E., Hodges, J.: Discriminatory analysis, nonparametric discrimination: consistency properties. Technical report, USAF School of Aviation Medicine, Randolph Field (1951)
2. T.M.Cover, P.E.Hart: Nearest neighbor pattern classification. IEEE Transactions on Information Theory **IT-13**(1) (1967) 21–27
3. Peng, J., Heisterkamp, D.R., Dai, H.K.: Adaptive quasiconformal kernel nearest neighbor classification. IEEE Trans. Pattern Anal. Mach. Intell. **26**(5) (2004) 656–661
4. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. J. Mach. Learn. Res. **10** (June 2009) 207–244
5. ling Zhang, M., hua Zhou, Z.: Ml-knn: A lazy learning approach to multi-label learning. PATTERN RECOGNITION **40** (2007) 2007
6. Ougiaroglou, S., Nanopoulos, A., Papadopoulos, A.N., Manolopoulos, Y., Welzer-druzovec, T.: Adaptive k-nearest neighbor classification based on a dynamic number of nearest neighbors. In: Proceedings of ADBIS Conference. ADBIS 2007 (2007)
7. Han, J.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2005)
8. Holte, R.C., Acker, L.E., Porter, B.W.: Concept learning and the problem of small disjuncts. In: Proc. 11th Int. Conf. AI - Volume 1, Morgan Kaufmann Publishers Inc. (1989) 813–818
9. van den Bosch, A., Weijters, T., Herik, H.J.V.D., Daelemans, W.: When small disjuncts abound, try lazy learning: A case study (1997)
10. Xing, Z., Pei, J., Yu, P.S.: Early prediction on time series: a nearest neighbor approach. In: Proceedings of the 21st international jont conference on Artifical intelligence. IJCAI'09, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (2009) 1297–1302
11. Boiman, O., Shechtman, E., Irani, M.: In Defense of Nearest-Neighbor Based Image Classification. In: CVPR. (2008)
12. Bellman, R.E.: Adaptive control processes - A guided tour. Princeton University Press, Princeton, New Jersey, U.S.A. (1961)
13. François, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE Transactions on Knowledge and Data Engineering **19**(7) (2007) 873–886
14. Hinneburg, A., Aggarwal, C., Keim, D.A.: What is the nearest neighbor in high dimensional spaces?, Morgan Kaufmann (2000) 506–515

15. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research **11** (2010) 2487–2531
16. Aucouturier, J., Pachet, F.: Improving timbre similarity: How high is the sky? Journal of Negative Results in Speech and Audio Sciences **1** (2004)
17. Gasser M., Flexer A., S.D.: Hubs and orphans - an explorative approach. In: Proceedings of the 7th Sound and Music Computing Conference. SMC'10 (2010)
18. Miloš, R.: Representations and Metrics in High-Dimensional Data Mining. Izdavačka knjižarnica Zorana Stojanovića, Novi Sad, Serbia (2011)
19. Tomašev, N., Brehar, R., Mladenić, D., Nedevschi, S.: The influence of hubness on nearest-neighbor methods in object recognition. In: Proceedings of the 7th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP). (2011) 367–374
20. Tomašev, N., Mladenić, D.: Hubness-aware shared neighbor distances for high-dimensional k-nearest neighbor classification. Knowledge and Information Systems (2013) 1–34
21. Radovanović, M., Nanopoulos, A., Ivanović, M.: Nearest neighbors in high-dimensional data: The emergence and influence of hubs. In: Proc. 26th Int. Conf. on Machine Learning (ICML). (2009) 865–872
22. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: Hubness-based fuzzy measures for high-dimensional k-nearest neighbor classification. In: Proc. MLDM. (2011)
23. Tomašev, N., Radovanović, M., Mladenić, D., Ivanovicć, M.: A probabilistic approach to nearest neighbor classification: Naive hubness bayesian k-nearest neighbor. In: Proceeding of the CIKM conference. (2011)
24. Tomašev, N., Mladenić, D.: Nearest neighbor voting in high-dimensional data: Learning from past occurrences. In: ICDM PhD Forum. (2011)
25. Tomašev, N., Mladenić, D.: Nearest neighbor voting in high dimensional data: Learning from past occurrences. Computer Science and Information Systems **9** (2012) 691–712
26. Schnitzer, D., Flexer, A., Schedl, M., Widmer, G.: Using mutual proximity to improve content-based audio similarity. In: ISMIR'11. (2011) 79–84
27. Tomašev, N., Mladenić, D.: Hubness-aware shared neighbor distances for high-dimensional k-nearest neighbor classification. In: Proceedings of the 7th International Conference on Hybrid Artificial Intelligence Systems. HAIS '12 (2012)
28. Tomašev, N., Radovanović, M., Mladenić, D., Ivanovic, M.: The role of hubness in clustering high-dimensional data. In: PAKDD (1)'11. (2011) 183–195
29. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: The role of hubness in clustering high-dimensional data. IEEE Transactions on Knowledge and Data Engineering **99**(PrePrints) (2013) 1
30. Jiang, L., Zhang, H., Cai, Z.: A novel bayes model: Hidden naive bayes. Knowledge and Data Engineering, IEEE Transactions on **21**(10) (Oct.) 1361–1371
31. Keller, J.E., Gray, M.R., Givens, J.A.: A fuzzy k-nearest-neighbor algorithm. In: IEEE Transactions on Systems, Man and Cybernetics. (1985) 580–585
32. Rish, I.: An empirical study of the naive Bayes classifier. In: Proc. IJCAI Workshop on Empirical Methods in Artificial Intelligence. (2001)
33. Chickering, D.M.: Learning bayesian networks is np-complete. In: Learning from Data: Artificial Intelligence and Statistics V, Springer-Verlag (1996) 121–130
34. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (November 2004) 91
35. Pracner, D., Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: WIKImage: Correlated Image and Text Datasets. In: SiKDD: Information Society. (2011)
36. Napierala, K., Stefanowski, J.: Identification of different types of minority class examples in imbalanced data. In Corchado, E., Snasel, V., Abraham, A., Wozniak, M., Graa, M., Cho, S.B., eds.: Hybrid Artificial Intelligent Systems. Volume 7209 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg (2012) 139–150