Nearest Neighbors in High-Dimensional Data: The Emergence and Influence of Hubs

Miloš Radovanović RADACHA@DMI.UNS.AC.RS
Department of Mathematics and Informatics, University of Novi Sad, Trg D. Obradovića 4, 21000 Novi Sad, Serbia

Alexandros Nanopoulos NANOPOULOS@ISMLL.DE
Institute of Computer Science, University of Hildesheim, Marienburger Platz 22, D-31141 Hildesheim, Germany

Mirjana Ivanović MIRA@DMI.UNS.AC.RS
Department of Mathematics and Informatics, University of Novi Sad, Trg D. Obradovića 4, 21000 Novi Sad, Serbia

Abstract

High dimensionality can pose severe difficulties, widely recognized as different aspects of the curse of dimensionality. In this paper we study a new aspect of the curse pertaining to the distribution of k-occurrences, i.e., the number of times a point appears among the k nearest neighbors of other points in a data set. We show that, as dimensionality increases, this distribution becomes considerably skewed and hub points emerge (points with very high k-occurrences). We examine the origin of this phenomenon, showing that it is an inherent property of high-dimensional vector space, and explore its influence on applications based on measuring distances in vector spaces, notably classification, clustering, and information retrieval.

1. Introduction

It is widely recognized that high-dimensional spaces pose severe difficulties, regarded as different aspects of the *the curse of dimensionality* (Bishop, 2006). One aspect of this curse is *distance concentration*, which directly affects machine learning applications. It refers to the tendency of distances between all pairs of points in high-dimensional data to become almost equal. Concentration of distances and the meaningfulness of finding nearest neighbors in high-dimensional spaces has been studied thoroughly (Beyer et al., 1999; Aggarwal et al., 2001; François et al., 2007).

There is another aspect of the curse of dimensionality that is related to nearest neighbors (NNs). Let D be a set of points and $N_k(x)$ the number of *k-occurrences* of each point $x \in D$, i.e., the number of times x occurs among the k NNs of all other points in D. Under certain conditions, as dimensionality increases, the distribution of N_k becomes considerably skewed to the right, resulting in the emergence of *hubs*, i.e., points which appear in many more k-NN lists than other points. Unlike distance concentration, the skewness of N_k has not been studied in depth. As will be described in Section 2.2, the two phenomena are related but distinct. In this paper we study the causes and the implications of this aspect of the dimensionality curse.

1.1. Related Work

The skewness of N_k recently started to be observed in fields like audio retrieval (Aucouturier & Pachet, 2007; Doddington et al., 1998) and fingerprint identification (Hicklin et al., 2005), where it is described as a problematic situation. Singh et al. (2003) notice possible skewness of N_1 on real data and account for it in their reverse NN search algorithm. Nevertheless, these works neither analyze the causes of skewness nor generalize it to other applications.

The distribution of k-occurrences has been explicitly studied in the applied probability community (Newman et al., 1983; Yao & Simons, 1996). No skewness was observed because of the different properties of the settings studied, which will be explained in Section 2.2.

1.2. Motivation and Contributions

Since the skewness of k-occurrences has been observed in the contexts of specific applications, the question remains whether it is limited to them by being an artifact of the data or the modeling algorithms. In this paper we show that it is actually an inherent property of high-dimensional vector spaces under widely used assumptions. To the best of

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009. Copyright 2009 by the author(s)/owner(s).

our knowledge, there has been no study relating this phenomenon with the properties of vector space and the dimensionality curse. It is worth to examine its origin and consequences because of its influence on applications based on distances in vector spaces, notably classification, clustering, and information retrieval.

We make the following contributions. First, we demonstrate and explain the emergence of skewness of k-occurrences (Section 2). We then study its implications on widely used techniques (Sections 3–5). As this is a preliminary examination of the problem, we provide a list of directions for future work (Section 6).

2. The Skewness of k-occurrences

In this section we first demonstrate the emergence of skewness in the distribution of N_k and then explain its causes.

2.1. A Motivating Example

We start with an illustrative experiment which demonstrates the changes in the distribution of N_k with varying dimensionality. Consider a random data set consisting of 10000 d-dimensional points drawn uniformly from the unit hypercube $[0, 1]^d$, and the following distance functions: Euclidean (l_2), fractional $l_{0.5}$ (proposed for high-dimensional data by Aggarwal et al. (2001)), and cosine. Figure 1 shows the empirically observed distributions of N_k, with $k = 5$, for (a) $d = 3$, (b) $d = 20$, and (c) $d = 100$.

For $d = 3$ the distributions of N_5 for the three distance functions (Fig. 1(a)) are consistent with the binomial distribution. This is expected when considering k-occurrences as node in-degrees in the k-nearest neighbor digraph. For uniformly distributed points in low dimensions, this digraph follows the Erdős-Rényi (ER) random graph model, in which the degree distribution is binomial (Erdős & Rényi, 1959).

As dimensionality increases, the observed distributions of N_5 depart from the random graph model and become more skewed to the right (Fig. 1(b, c)). We verified this by being able to fit the tails of the distributions with the log-normal distribution, which is highly skewed (fits were supported by the χ^2-test at 0.05 confidence level). We made similar observations with various k values, distance measures (l_p norm for both $p \geq 1$ and $0 < p < 1$, Bray-Curtis, normalized Euclidean, and Canberra), and distributions, like the normal. In all these cases, skewness exists and produces hubs, i.e., points with high N_k.

2.2. The Causes of Skewness

The skewness of k-occurrences appears to be related with the phenomenon of distance concentration, which is usually expressed as the ratio between some measure of spread and some measure of magnitude of distances of all points in a data set to some arbitrary reference point (Aggarwal et al., 2001; François et al., 2007). If this ratio converges to 0 as dimensionality goes to infinity, it is said that the distances concentrate.

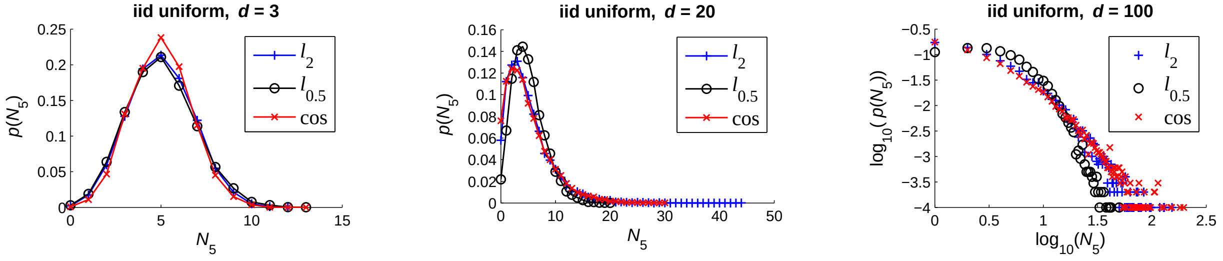
To ease comprehension, consider again the iid uniform random data examined in the previous section and select as the reference point the mean of the distribution. Figure 2 plots, for each point x, its $N_5(x)$ against its Euclidean distance from the mean, for $d = 3,\ 20,\ 100$. As dimensionality increases, stronger correlation emerges, meaning that points closer to the mean tend to become hubs. We need to understand why some points tend to be closer to the mean and, thus, become hubs. Based on existing theoretical results (Beyer et al., 1999; Aggarwal et al., 2001), high dimensional points are approximately lying on a hypersphere centered at the data set mean. Moreover, the results by Demartines (1994) and François et al. (2007) specify that the distribution of distances to the data set mean has a non-negligible variance for any finite d.[1] Hence, the existence of a non-negligible number of points closer to the data set mean is *expected* in high dimensions. These points, by being closer to the mean, tend to be closer to all other points – a tendency which is amplified (in relative terms) by high dimensionality, making points closer to the data set mean have increased inclusion probability into k-NN lists, even for small values of k.

Note that the non-negligible variance has an additional "side": we also expect points farther from the mean and, thus, with much lower N_k than the rest. Such points correspond to the bottom-right parts of Fig. 2(b, c), and can be regarded as *outliers* since they are also far away from all other points (Tan et al., 2005). Outliers will be analyzed further in Section 4.

Research in applied probability describes that, within the Poisson process setting, as $d \to \infty$ the distribution of k-occurrences converges to the Poisson distribution with mean k (Newman et al., 1983; Yao & Simons, 1996), which implies no skewness. However, a Poisson process produces an *unbounded* infinite set of points for which no meaningful data set mean exists, and distances do not concentrate (their spread and magnitude are infinite). Through simulation of this setting we verified that, once boundaries are introduced (as in the majority of practical cases), skewness of N_k emerges.[2]

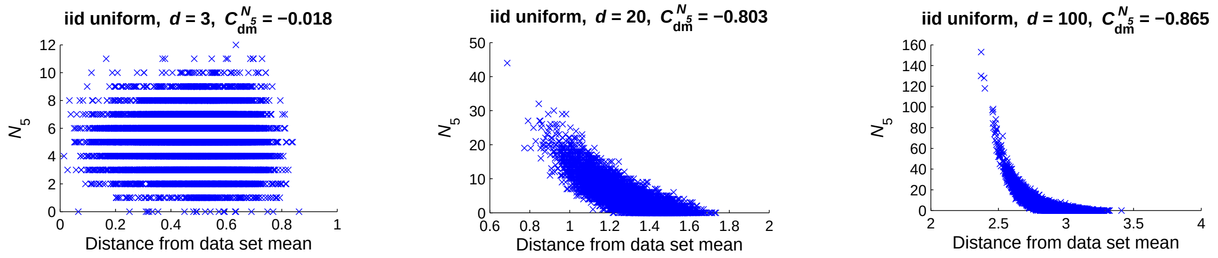
[1] These results apply to l_p distances, but our numerical simulations suggest that other mentioned distance functions behave similarly.

[2] With the exception of combinations of (bounded) data distributions and distances without meaningful means, e.g. centered normal distribution and cosine distance.



(a) (b) (c)

Figure 1. Distribution of 5-occurrences for Euclidean, $l_{0.5}$, and cosine distances on iid uniform random data sets with dimensionality (a) $d = 3$, (b) $d = 20$, and (c) $d = 100$ (log-log plot).



(a) (b) (c)

Figure 2. Scatter plots and Spearman correlation of $N_5(x)$ against the Euclidean distance of point x to the data set mean for iid uniform random data with (a) $d = 3$, (b) $d = 20$, and (c) $d = 100$.

2.3. Skewness in Real Data

In this section we examine the skewness of N_k in real data which, unlike previously examined random data, present two factors: (a) they usually have dependent attributes, therefore we need to consider their *intrinsic dimensionality* (measured by the maximum likelihood estimator (Levina & Bickel, 2005)), (b) they are usually clustered, hence we need to consider more than one group of points.

We examined 50 real data sets belonging to three categories: UCI multidimensional data, gene expression data, and textual data. Due to space considerations, Table 1 lists mainly the data sets used in later sections. Columns describe data set sources, basic statistics (whether attributes were standardized or not, the number of points (n), embedding dimensionality (d), estimated intrinsic dimensionality (d_{mle}), number of classes), the number of groups (clusters) for expressing correlations with N_k, and the distance metric (Euclidean or cosine). We took care to ensure that the choice of distance metric and preprocessing (i.e., standardization) corresponds to a realistic scenario for the particular data set. The value of k will be fixed at 10.

We characterize the asymmetry of N_k with the standardized third moment $S_{N_k} = \mathrm{E}(N_k - \mu_{N_k})^3/\sigma^3_{N_k}$ (μ_{N_k}, σ_{N_k} are the mean and standard deviation of N_k, resp.). The corresponding (10th) column of Table 1 shows that N_{10} of all examined data are skewed to the right.[3] Moreover, we define for each point x its *standardized hubness* score $h(x)$ (used in later sections):

$$h(x) = (N_k(x) - \mu_{N_k})/\sigma_{N_k} \,. \qquad (1)$$

To examine the first factor (intrinsic dimensionality), for each data set we randomly permuted the elements within every attribute. This way, attributes preserve their individual distributions, but the dependencies between them are lost and intrinsic dimensionality increases (François et al., 2007). In Table 1 (11th column) we give the skewness, denoted $S^S_{N_k}$, of the modified data. In most cases $S^S_{N_k}$ is considerably higher than S_{N_k}, implying that skewness depends on the intrinsic rather than embedding dimensionality.

To examine the second factor (many groups), for every data set we measured: (i) the Spearman correlation, denoted as $C^{N_{10}}_{\text{dm}}$ (12th column), of N_k and the distance from the data set mean, and (ii) the correlation, denoted as $C^{N_{10}}_{\text{cm}}$ (13th column), of N_k and the distance to the closest group mean. Groups are determined using K-means clustering, where the number of clusters was computed for each data set by exhaustive search of values between 2 and $\lfloor\sqrt{n}\rfloor$, in order to maximize $C^{N_{10}}_{\text{cm}}$.[4] In most cases, $C^{N_{10}}_{\text{cm}}$ is much stronger than $C^{N_{10}}_{\text{dm}}$. Consequently, in real data hubs are closer than other points to their respective cluster centers (which we verified by examining the actual scatter plots).

[3] If $S_{N_k} = 0$ there is no skewness, positive (negative) values signify skewness to the right (left).

[4] We report averages of $C^{N_{10}}_{\text{cm}}$ over 10 runs of K-means clustering with different random seeding, in order to reduce the effects of chance.

Table 1. Real data sets (portion). Data sources are the UCI Machine Learning Repository, and Kent Ridge Bio-medical Repository (KR).

Name	Src.	Stan.	n	d	d_{mle}	Cls.	Clu.	Dist.	$S_{N_{10}}$	$S^S_{N_{10}}$	$C^{N_{10}}_{\mathrm{dm}}$	$C^{N_{10}}_{\mathrm{cm}}$	CAV	$\widetilde{BN}_{10}$	$C^{N_{10}}_{BN_{10}}$
ecoli	UCI	yes	336	7	4.13	8	8	l_2	0.116	0.208	−0.396	−0.792	0.193	0.223	0.245
ionosphere	UCI	yes	351	34	13.57	2	18	l_2	1.717	2.051	−0.639	−0.832	0.259	0.185	0.464
mfeat-factors	UCI	yes	2000	216	8.47	10	44	l_2	0.826	5.493	−0.113	−0.688	0.145	0.063	0.001
mfeat-fourier	UCI	yes	2000	76	11.48	10	44	l_2	1.277	4.001	−0.350	−0.596	0.415	0.272	0.436
musk1	UCI	yes	476	166	6.74	2	17	l_2	1.327	3.845	−0.376	−0.752	0.474	0.237	0.621
optdigits	UCI	yes	5620	64	9.62	10	74	l_2	1.095	3.789	−0.223	−0.601	0.168	0.044	0.097
page-blocks	UCI	yes	5473	10	3.73	5	72	l_2	−0.014	0.470	−0.063	−0.289	0.068	0.049	−0.046
pendigits	UCI	yes	10992	16	5.93	10	104	l_2	0.435	0.982	−0.062	−0.513	0.156	0.014	−0.030
segment	UCI	yes	2310	19	3.93	7	48	l_2	0.313	1.111	−0.077	−0.453	0.332	0.089	0.074
sonar	UCI	yes	208	60	9.67	2	8	l_2	1.354	3.053	−0.550	−0.771	0.461	0.286	0.632
spambase	UCI	yes	4601	57	11.45	2	49	l_2	1.916	2.292	−0.376	−0.448	0.271	0.139	0.401
spectrometer	UCI	yes	531	100	8.04	10	17	l_2	0.591	3.123	−0.269	−0.670	0.242	0.200	0.225
vehicle	UCI	yes	846	18	5.61	4	25	l_2	0.603	1.625	−0.162	−0.643	0.586	0.358	0.435
vowel	UCI	yes	990	10	2.39	11	27	l_2	0.766	0.935	−0.252	−0.605	0.598	0.313	0.691
lungCancer	KR	no	181	12533	59.66	2	6	l_2	1.248	3.073	−0.537	−0.673	0.136	0.052	0.262
ovarian-61902	KR	no	253	15154	9.58	2	10	l_2	0.760	3.771	−0.559	−0.773	0.399	0.164	0.467
mini-newsgroups	UCI	no	1999	7827	3226.43	20	44	cos	1.980	1.765	−0.422	−0.704	0.526	0.524	0.701
reuters-transcribed	UCI	no	201	3029	234.68	10	3	cos	1.165	1.693	−0.781	−0.763	0.595	0.642	0.871

2.4. Skewness and Intrinsic Dimensionality

The results in Table 1 suggest that the skewness of N_k is strongly correlated with d (Spearman correlation over *all* 50 data sets is 0.62), and especially with the intrinsic dimensionality d_{mle} (Spearman correlation over *all* 50 data sets is 0.80). We elaborate further on the interplay of skewness and intrinsic dimensionality by considering dimensionality reduction (DR) techniques. The main question is whether DR can alleviate the issue of the skewness of k-occurrences altogether.

We examined the widely used principal component analysis (PCA) dimensionality reduction method. Figure 3 depicts for several real data sets, and iid uniform random data, the relationship between the percentage of features maintained by PCA, and S_{N_k} ($k = 10$). For real data, S_{N_k} stays relatively constant until a small percentage of features is left, after which it suddenly drops. This is the point where the intrinsic dimensionality is reached, and further reduction incurs loss of information. Such behavior is in contrast with the case of iid uniform random data, where S_{N_k} steadily reduces with the decreasing number of (randomly) selected features (PCA is not meaningful in this case), because intrinsic and embedded dimensionalities are equal. These observations indicate that dimensionality reduction does not have a significant effect on the skewness of k-occurrences when the number of features is above the intrinsic dimensionality, a result that is useful in most practical cases since otherwise loss of significant information may occur.

3. Influence on Classification

From this section we start to investigate possible implications of the skewness of N_k on widely used machine learning methods, beginning with supervised learning.

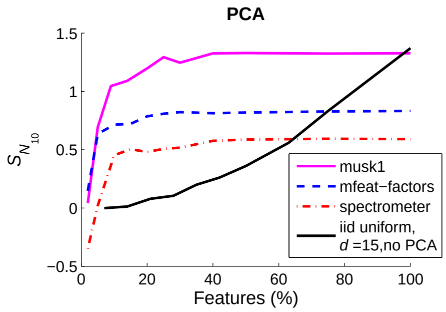


Figure 3. Skewness of N_{10} in relation to the percentage of PCA features kept.

3.1. "Good" and "Bad" k-occurrences

When labels are present, k-occurrences can be distinguished based on whether labels of neighbors match. We define the number of *"bad" k-occurrences* of x, $BN_k(x)$, as the number of points from D for which x is among the first k NNs *and* the labels of x and the points in question *do not match*. Conversely, $GN_k(x)$, the number of *"good" k-occurrences* of x, is the number of such points where labels do match. Naturally, for every $x \in D$, $N_k(x) = BN_k(x) + GN_k(x)$.

To account for labels, Table 1 includes $\widetilde{BN}_{10}$ (15th column), the sum of all "bad" 10-occurrences of a data set normalized by $\sum_x N_{10}(x) = 10n$. This measure is intended to express the total amount of "bad" k-occurrences within a data set. Also, to compute the amount of information regular k-occurrences contain about "bad" k-occurrences in a data set, $C^{N_{10}}_{BN_{10}}$ (16th column) denotes the Spearman correlation between BN_{10} and N_{10} vectors. The motivation behind this measure is to express the degree to which BN_k and N_k follow a similar distribution.

"Bad" hubs, i.e., points with high BN_k, are of particular interest to supervised learning, since they carry more information about the location of the decision boundaries than other points, and affect classification algorithms (as will be described next). To understand the origins of "bad" hubs in real data, we rely on the notion of the *cluster assumption* from semi-supervised learning (Chapelle et al., 2006), which roughly states that most pairs of points in a high density region (cluster) should be of the same class. To measure the degree to which the cluster assumption is violated in a particular data set, we simply define the *cluster assumption violation* (CAV) coefficient as follows. Let a be the number of pairs of points which are in different classes but in the same cluster, and b the number of pairs of points which are in the same class and cluster. Then, we define $\mathrm{CAV} = a/(a+b)$, which gives a number in range $[0, 1]$, higher if there is more violation. To reduce the sensitivity of CAV to the number of clusters (too low and it will be overly pessimistic, too high and it will be overly optimistic), we choose the number of clusters to be 3 times the number of classes of a particular data set. Clustering is performed with K-means.

For all examined (50) real data sets, we computed the Spearman correlations between $\widetilde{BN}_{10}$ and CAV (14th column of Table 1), and found it strong (0.85). Another significant correlation (0.39) is observed between $C^{N_{10}}_{BN_{10}}$ and intrinsic dimensionality. In contrast, $\widetilde{BN}_{10}$ is not correlated with intrinsic dimensionality nor with the skewness of N_{10} (both correlations are around 0.03). The latter fact indicates that high dimensionality and skewness of N_k are not sufficient to induce "bad" hubs. Instead, based on the former fact, we can argue that there are two, mostly independent, forces at work: violation of the cluster assumption on one hand, and high intrinsic dimensionality on the other. "Bad" hubs originate from putting the two together; i.e., the consequences of violating the cluster assumption can be more severe in high dimensions than in low dimensions, not in terms of the total amount of "bad" k-occurrences, but in terms of their distribution, since strong regular hubs are now more prone to "pick up" bad k-occurrences than non-hub points. This is supported by the positive correlation between $C^{N_{10}}_{BN_{10}}$ with intrinsic dimensionality, meaning that in high dimensions BN_k tends to follow a more similar distribution to N_k than in low dimensions.

3.2. Influence on Classification Algorithms

We now examine how the skewness of N_k and the existence of ("bad") hubs affects well-known classification techniques, focusing on the k-NN classifier, support vector machines (SVM), and AdaBoost.

k-NN classifier. The k-NN classifier is negatively affected by the presence of "bad" hubs, because they provide erroneous class information to many other points. To validate this assumption, we tried a simple weighting scheme. For each point x, we calculate its standardized "bad" hubness score, $h_B(x)$, by adapting its $h(x)$ score (Equation 1) to consider $BN_k(x)$ instead of $N_k(x)$. Thus, during majority voting, when a point x participates a k-NN list, its vote is weighted by $e^{-h_B(x)}$. Figure 4(a–d) compares the resulting accuracy of k-NN classifier with and without this weighting scheme for some data sets of Table 1. Leave-one-out evaluation is performed using Euclidean distance, whereas the k value for N_k is naturally set to the k value used by the k-NN classifier. The reduced accuracy of the unweighted scheme signifies the negative influence of "bad" hubs.

Support vector machines. We consider SVM with the RBF (Gaussian) kernel, which is a smooth monotone function of the Euclidean distance between points. Therefore, N_k values in the kernel space are exactly the same as in the original space.[5] To examine the influence of "bad" hubs on SVM, Fig. 4(e, f) illustrates 10-fold cross-validation accuracy results when points are progressively removed from the training sets: (i) by decreasing BN_k ($k = 5$), and (ii) at random. Accuracy drops with removal by BN_k, indicating that bad hubs are important for SVMs.

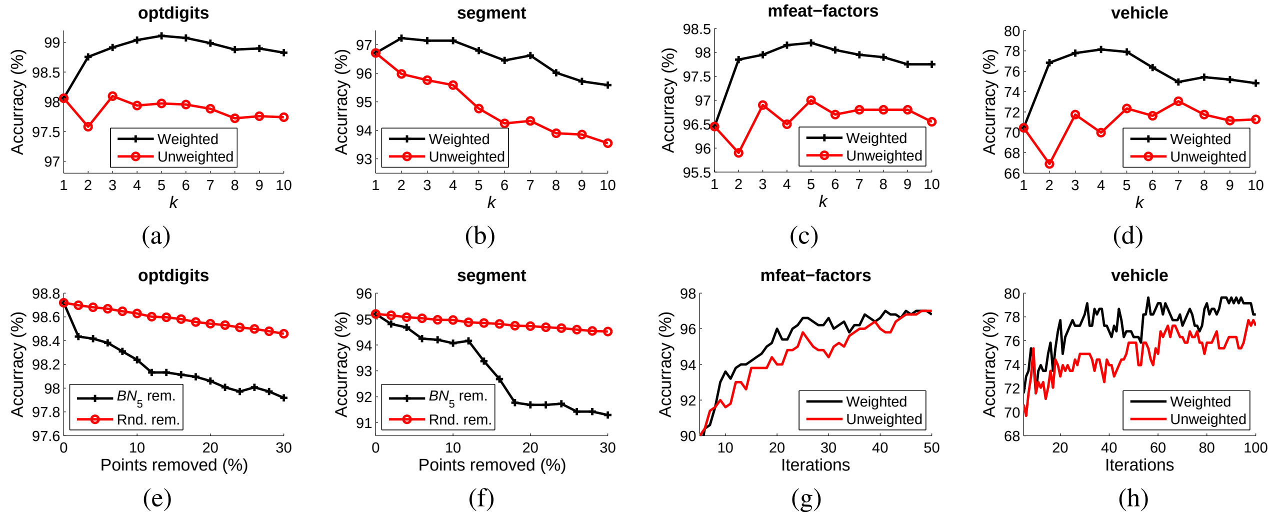
The reason is that in high-dimensional data, points with high BN_k can comprise good support vectors. Table 2 exemplifies this point by listing, for several data sets, normalized average ranks of support vectors in the 10-fold cross-validation models with regards to decreasing BN_k. The ranks are in the range $[0, 1]$, with the value 0.5 expected from a random set of points. Lower values of the ranks indicate that the support vectors, on average, tend to have high BN_k.

Table 2. Normalized average support vector ranks with regards to decreasing BN_5.

Data set	SV rank	Data set	SV rank
mfeat-factors	0.218	page-blocks	0.267
mfeat-fourier	0.381	segment	0.272
optdigits	0.189	vehicle	0.464

AdaBoost. Boosting algorithms take into account the "importance" of points in the training set for classification by weak learners, usually by assigning and updating weights of individual points – the higher the weight, the more attention is to be paid to the point by following learners. We consider the classical AdaBoost algorithm in conjunction with CART trees of maximal depth 3, and set the initial weight of each point x in the training set to $1/(1 + |h(x)|)$, normalized by the sum over all points ($h(x)$ was defined in Equation 1). The motivation behind the weighting scheme is to assign less importance to both hubs and outliers than

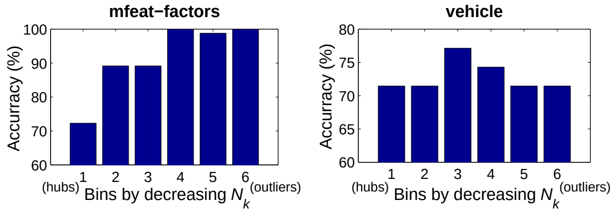
[5] Centering the kernel matrix changes the N_k of points in the kernel space, but we observed that the overall distribution (i.e. its skewness) does not become radically different. Therefore, the following arguments still hold for centered kernels, providing N_k is calculated in the kernel space.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4. (a–d) Accuracy of k-NN classifier with and without the weighting scheme. (e, f) Accuracy of SVM with RBF kernel and points being removed from the training sets by decreasing BN_5, and at random (averaged over 20 runs). (g, h) Accuracy of AdaBoost with and without the weighting scheme: (g) $k = 20$, (h) $k = 40$.



(a) (b)

Figure 5. Binned accuracy of AdaBoost, by decreasing N_k.

other points (this is why we take the abs of $h(x)$). Figure 4(g, h) illustrates on two data sets from Table 1 how the weighting scheme helps AdaBoost achieve better generalization in fewer iterations, showing the classification accuracy on a 2:1:1 training-validation-test set split (validation sets are used to determine the values of k). While it is known that AdaBoost is sensitive to outliers, this suggests that hubs should be regarded in an analogous manner, i.e. both hubs and outliers are intrinsically more difficult to classify correctly, and the attention of the weak learners should initially be focused on regular points.

To further support this claim Fig. 5 depicts binned accuracies of unweighted AdaBoost trained in one fifth of the iterations shown in Fig. 4(g, h), for points sorted by decreasing k-occurrences. It exemplifies how in earlier phases of ensemble training the generalization power with hubs and/or outliers is worse than with regular points.

4. Influence on Clustering

The main objectives of clustering algorithms are to minimize intra-cluster distance and maximize inter-cluster distance. The skewness of k-occurrences in high-dimensional data influences both objectives.

Intra-cluster distance may be increased due to points with low k-occurrences. As mentioned in Section 2.2, such points are far from all the rest, acting as outliers. A common outlier score of a point is its distance from its k-th nearest neighbor (Tan et al., 2005). Low N_k values and high outlier scores are correlated as exemplified in Fig. 6(a, b) (in their lower-right parts) for two data sets from Table 1. Outliers and their influence on clustering are well-studied subjects (Tan et al., 2005): outliers do not cluster well because they have high intra-cluster distance, thus they are often discovered and eliminated beforehand. The existence of outliers is attributed to various reasons (e.g., erroneous measurements). Nevertheless, the skewness of N_k suggests that in high-dimensional data outliers are also expected due to inherent properties of vector space.

Inter-cluster distance may be reduced due to points with high k-occurrences, i.e., hubs. Like outliers, hubs do not cluster well, but for a different reason: they have low inter-cluster distance, because they are close to many points, thus also to points from other clusters. In contrast to outliers, the influence of hubs on clustering has not attracted significant attention.

To examine the influence of both outliers and hubs, we used the popular silhouette coefficients (SC) (Tan et al., 2005). For the i-th point, a_i is its average distance to all points in its cluster (a_i corresponds to intra-cluster distance), whereas b_i is the minimum average distance to points of other clusters (b_i corresponds to inter-cluster distance). The SC of the i-th point is $(b_i - a_i)/\max(a_i, b_i)$, in range $[-1, 1]$ (higher values are preferred). We examined several clustering algorithms and distance measures, but due to lack of space report results for the algorithm of Meilă and Shi (2001) and Euclidean distance. Follow-

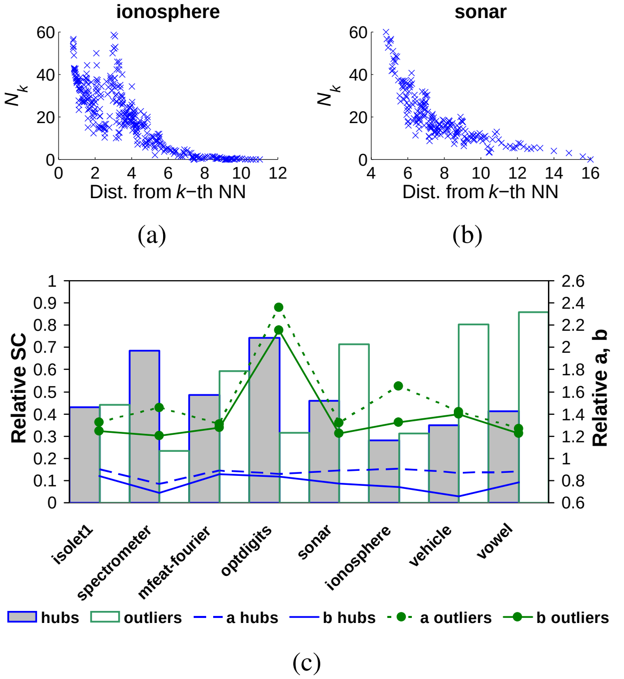


Figure 6. (a, b) Correlation between low N_k and outlier score ($k = 20$). (c) Relative silhouette coefficients for hubs (gray filled bars) and outliers (empty bars). Relative values for a and b coefficients are also plotted (referring to the right vertical axes).

ing a standard method, we select as hubs those points x with $h(x) > 2$, i.e., $N_k(x)$ more than 2 standard deviations higher than the mean (note that $h(x)$ ignores labels). Let n_h be their number. Moreover, we select as outliers the n_h points with the lowest k-occurrences. Finally, we randomly select n_h points from the remaining points (we report averages for 100 different selections). To compare hubs and outliers against random points, we measure as relative SC of hubs (outliers) the quotient of the mean SC of hubs (outliers) divided by mean SC of random points. For several data sets from Table 1, Figure 6(c) plots with bars the relative SC. As expected, outliers have relative SC lower than one, meaning that they cluster worse than random points. Notably, the same holds for hubs, too.[6]

To gain further insight, the same figure plots with lines (referring to the right vertical axes) for hubs and outliers their relative mean values of a_i and b_i (dividing with those of randomly selected points). Outliers have high relative a_i values, indicating higher intra-cluster distance. Hubs, in contrast, have low relative b_i values, indicating reduced inter-cluster distance. In conclusion, when clustering high-dimensional data, hubs should receive analogous attention as outliers.

5. Influence on Information Retrieval

The existence of hubs affects applications that are based on nearest neighbor queries. A typical example in information retrieval is finding documents that are most similar to a query document. Hub documents will be frequently included in the result list without necessarily being relevant to the query. This may harm the precision of results and the users' experience, by having "persistent" irrelevant results.

To reduce the inclusion probability of hubs in the result list, we can increase their distance from query documents. For this purpose we use a simple scheme. Given a database, D, of documents, for each $x \in D$, let $d(x, q)$ be the distance (in the range $[0, 1]$) between x and a query document q. Being consistent with the findings in the previous section, we ignore labels and consider as hubs those documents with $h(x) > 2$. For hubs, we increase $d(x, q)$ as follows:

$$d(x, q) \leftarrow d(x, q) + (1 - d(x, q)) \frac{h(x)}{\max_{y \in D} h(y)} \, .$$

Thus, the higher the $h(x)$, the more the distance is increased, remaining however in the range $[0, 1]$.

We examine the influence of hubs by comparing this scheme with the regular case (no increment). We use cosine distance, leave-one-out cross validation, and measure precision (fraction of results with same class label as the query) versus the number of retrieved documents. Figure 7 reports (with bars) precision with and without using the increment scheme, for two selected (for space considerations) text data sets given in Table 1. The same figure depicts with lines (referring to the right axis) the probability of a hub to be included in the result list. In the regular case (no increment), hubs have much higher inclusion probability, resulting in lower precision.

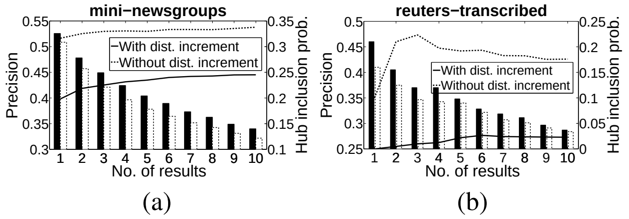


Figure 7. Precision with (solid bars) and without (dashed bars) the distance increment scheme: (a) $k = 10$, (b) $k = 1$. Inclusion probability of hubs is also plotted (against the right vertical axes) with (solid line) and without (dashed line) the increment scheme.

6. Conclusion

We explored the emergence of skewness of k-occurrences in high-dimensional data, and possible implications on several important applications, namely classification, clustering, and information retrieval. Experimental results suggest these applications are affected by the skewness phe-

[6] The statistical significance of differences between the mean SC of hubs and randomly selected points has been verified with the double t-test at 0.05 confidence level.

nomenon, and that explicitly taking skewness into account can improve the accuracy of different methods. Although this was a preliminary examination, we hope we demonstrated that the phenomenon may be significant to many fields of machine learning, data mining and information retrieval, and that it warrants further investigation.

Possible directions for future work include a more formal and theoretical study of the interplay between the skewness of k-occurrences and various distance-based machine learning models, possibly leading to approaches that account for the phenomenon at a deeper level. Supervised learning methods may deserve special attention, as it was also observed in another study (Caruana et al., 2008) that the k-NN classifier and boosted decision trees can experience problems in high dimensions. Further directions of research may involve determining whether the phenomenon is applicable to probabilistic models, (unboosted) decision trees, and other techniques not explicitly based on distances between points; and also to algorithms based on general metric spaces. Since we determined a correlation between hubness and the proximity to cluster centers for K-means clustering of high-dimensional data, it would be interesting to explore how this can be used in seeding iterative clustering algorithms, like K-means or self-organizing maps. The interplay between dimensionality reduction and skewness of N_k may also be worth further study. Other fields that could directly benefit from an investigation into the skewness of N_k include outlier detection and reverse k-NN queries. Finally, as we determined high correlation between intrinsic dimensionality and the skewness of N_k, it would be interesting to see whether some measure of skewness of the distribution of N_k can be used for estimation of the intrinsic dimensionality of a data set.

Acknowledgments

The authors would like to thank the reviewers for their valuable comments. The presented work was supported by project *Abstract Methods and Applications in Computer Science* (no. 144017A), of the Serbian Ministry of Science.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional spaces. *Proc. Int. Conf. on Database Theory* (pp. 420–434).

Aucouturier, J.-J., & Pachet, F. (2007). A scale-free distribution of false positives for a large class of audio similarity measures. *Pattern Recognition*, *41*, 272–284.

Beyer, K. S., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is “nearest neighbor” meaningful? *Proc. Int. Conf. on Database Theory* (pp. 217–235).

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. *Proc. Int. Conf. on Machine Learning* (pp. 96–103).

Chapelle, O., Schölkopf, B., & Zien, A. (Eds.). (2006). *Semi-supervised learning*. The MIT Press.

Demartines, P. (1994). *Analyse de données par réseaux de neurones auto-organisés*. Doctoral dissertation, Institut Nat’l Polytechnique de Grenoble, France.

Doddington, G., Liggett, W., Martin, A., Przybocki, M., & Reynolds, D. (1998). SHEEP, GOATS, LAMBS and WOLVES: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. *Proc. Int. Conf. on Spoken Language Processing*. Paper 0608.

Erdős, P., & Rényi, A. (1959). On random graphs. *Publicationes Mathematicae Debrecen*, *6*, 290–297.

François, D., Wertz, V., & Verleysen, M. (2007). The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, *19*, 873–886.

Hicklin, A., Watson, C., & Ulery, B. (2005). *The myth of goats: How many people have fingerprints that are hard to match?* (Technical Report). National Institute of Standards and Technology.

Levina, E., & Bickel, P. J. (2005). Maximum likelihood estimation of intrinsic dimension. *Advances in Neural Information Processing Systems 17* (pp. 777–784).

Meilă, M., & Shi, J. (2001). Learning segmentation by random walks. *Advances in Neural Information Processing Systems 13* (pp. 873–879).

Newman, C. M., Rinott, Y., & Tversky, A. (1983). Nearest neighbors and voronoi regions in certain point processes. *Advances in Applied Probability*, *15*, 726–751.

Singh, A., Ferhatosmanoğlu, H., & Tosun, A. S. (2003). High dimensional reverse nearest neighbor queries. *Proc. Int. Conf. on Information and Knowledge Management* (pp. 91–98).

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Addison Wesley.

Yao, Y.-C., & Simons, G. (1996). A large-dimensional independent and identically distributed property for nearest neighbor counts in Poisson processes. *Annals of Applied Probability*, *6*, 561–571.