Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Localized Centering: Reducing Hubness in Large-Sample Data

Kazuo Hara*
kazuo.hara@gmail.com
National Institute of Genetics
Mishima, Shizuoka, Japan

Ikumi Suzuki*
suzuki.ikumi@gmail.com
National Institute of Genetics
Mishima, Shizuoka, Japan

Masashi Shimbo
shimbo@is.naist.jp
Nara Institute of Science and Technology
Ikoma, Nara, Japan

Kei Kobayashi
kei@ism.ac.jp
The Institute of Statistical Mathematics
Tachikawa, Tokyo, Japan

Kenji Fukumizu
fukumizu@ism.ac.jp
The Institute of Statistical Mathematics
Tachikawa, Tokyo, Japan

Miloš Radovanović
radacha@dmi.uns.ac.rs
University of Novi Sad
Novi Sad, Serbia

Abstract

Hubness has been recently identified as a problematic phenomenon occurring in high-dimensional space. In this paper, we address a different type of hubness that occurs when the number of samples is large. We investigate the difference between the hubness in high-dimensional data and the one in large-sample data. One finding is that centering, which is known to reduce the former, does not work for the latter. We then propose a new hub-reduction method, called *localized centering*. It is an extension of centering, yet works effectively for both types of hubness. Using real-world datasets consisting of a large number of documents, we demonstrate that the proposed method improves the accuracy of k-nearest neighbor classification.

Introduction

The k-nearest neighbor (kNN) classifier (Cover and Hart 1967) is a simple instance-based classification algorithm. It predicts the class label for a query sample using a majority vote from its k most similar samples in a dataset, and thus no traning is required beforehand. In spite of its simplicity, the performance of kNN is comparable to other methods in tasks such as text classification (Yang and Liu 1999; Colas and Brazdil 2006).

However, the kNN classifier is vulnerable to *hubness*, a phenomenon known to occur in high-dimensional data; i.e., some samples in a high-dimensional dataset emerge as *hubs* that frequently appear in the k nearest (or k most-similar) neighbors of other samples (Radovanović, Nanopoulos, and Ivanović 2010a).

The emergence of hubness often affects the accuracy of kNN classification, as it incurs a bias in the prediction of the classifier towards the labels of the hubs. This happens because the predicted label of a query sample is determined by the labels of its kNN samples, in which hubs are very likely included.

*Equally contributed

To mitigate the influence of hubness in the kNN classification, Suzuki et al. (2013) reported that *centering*—that is, shifting the origin of the vector space to the centroid of the dataset—is effective at reducing hubs, when sample similarity is measured by the inner product.

Contributions

As in Suzuki et al. (2013), this paper also addresses how to alter the inner product similarity into the one which produces less hubs.

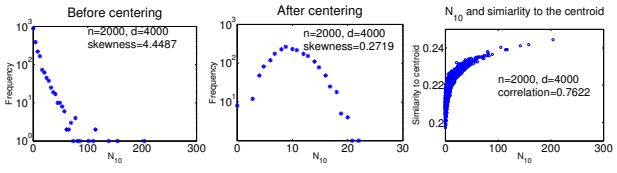
Thus far, it is known that hubness emerges in high-dimensional datasets, and that hubs are the samples similar to the data centroid (Radovanović, Nanopoulos, and Ivanović 2010a). In such cases, centering successfully reduces hubness.

We point out that there exists a different type of hubness which occurs when the number n of samples is large, with the dimension d of the vector space not necessarily very high; e.g., $n = 10,000$, $d = 500$. Unlike the hubs in high-dimensional datasets, these hubs are generally not so similar to the centroid of the entire dataset. Consequently, centering fails to reduce them.

We demonstrate that such a "new" hub sample is the one highly similar to its *local centroid*, the mean of samples within its local neighborhood. On the basis of this finding, we propose a new hub-reduction method, which we call *localized centering*. Our experimental results using synthetic data indicate that localized centering reduces the hubness not suppressed by classical centering. Using large real-world datasets, moreover, we show that the proposed method improves the performance of document classification with kNN insofar as it reduces hubness.

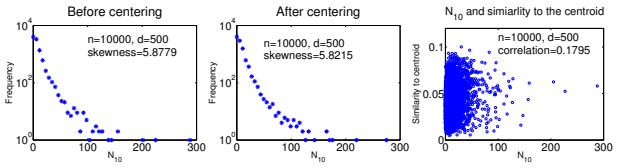
Hubness in High-Dimensional Data

Hubness is known as a phenomenon concerning nearest neighbors in high dimensional space (Radovanović, Nanopoulos, and Ivanović 2010a). Let $D \subset \mathbb{R}^d$ be a dataset in d-dimensional space, and let $N_k(x)$ denote the number of times a sample $x \in D$ occurs in the kNNs of other samples



(a) (b) (c)

Figure 1: Hubness in high-dimensional data ($d = 4000$): The N_{10} distribution (a) before centring, and (b) after centering; (c) scatter plot of samples with respect to the N_{10} value and the similarity to the data centroid (before centering).



(a) (b) (c)

Figure 2: Hubness in large-sample data ($|D| = 10,000$): The N_{10} distribution (a) before centering, and (b) after centering; (c) scatter plot of samples with respect to the N_{10} value and the similarity to the data centroid (before centering).

in D under some similarity measure. As the dimension d increases, the shape of the distribution for N_k changes such that it has a longer right tail, with a small number of samples taking unexpectedly large N_k values. Such samples are called *hubs*, and this phenomenon (or the extent to which a sample is a hub) is called *hubness*.

Here we demonstrate the emergence of hubness using synthetic data. To simulate a document collection represented as "bag-of-words" vectors, we generate a set D of sparse d-dimensional vectors holding non-negative elements. Let $n = |D|$ be the number of samples to generate. Initially, all n vectors in D are null. For each dimension $i = 1, \cdots, d$, a real number is drawn from the *LogNormal*$(5, 1)$ distribution. Let n_i be its rounded integer. We make exactly n_i items hold a non-zero value in the ith component, by choosing n_i vectors from D uniformly at random, and replacing their ith component with a random number drawn uniformly from $[0, 1]$. Finally all sample vectors are normalized to unit length.

We use the inner product as the measure of similarity. Since all sample vectors have unit length, the sample similarity is equivalent to cosine. Note however that for vectors not belonging to D, the similarity (i.e., inner product) is not necessarily equivalent to cosine. This also applies to the centroid (i.e., the sample mean) of D, which may not have a unit length even if all samples in D do.

Figure 1(a) shows the distribution of N_{10} (i.e., N_k with $k = 10$) for $n = 2000$ samples in high-dimensional space ($d = 4000$). In the figure, we can observe the presence of *hubs*, i.e., samples with an extremely large N_{10} value.

Following Radovanović, Nanopoulos, and Ivanović (2010a), we evaluate the degree of hubness by the skewness of the N_k distribution, $S_{N_k} = \mathbb{E}[(N_k - \mu_{N_k})^3]/\sigma_{N_k}^3$, where $\mathbb{E}[\cdot]$ is the expectation operator, and μ_{N_k} and σ_{N_k} are the mean and the standard deviation of the N_k distribution, respectively. Skewness is a standard measure for the degree of symmetry of the distributions. Its value is zero for a symmetric distribution like Gaussian, and it takes a positive or negative value for distributions with a long right or left tail. In particular, a large (positive) S_{N_k} indicates strong hubness in a dataset. Indeed, S_{N_k} is 4.45 in this synthetic dataset (Figure 1(a)).

Figure 1(c) gives the scatter plot of samples with respect to the N_{10} value and the similarity to the data centroid. As the figure shows, a strong correlation exists between them. Thus for this high-dimensional dataset, the samples similar to the centroid are making hubs. See Radovanović, Nanopoulos, and Ivanović (2010a) for more examples.

Centering as a Hubness Reduction Method

Centering is a transformation that involves shifting the origin of the feature space to the data centroid $c = (1/|D|)\sum_{x \in D} x$. It is a classic technique for removing observation bias in the data, but only recently has it been identified as a method for reducing hubness (Suzuki et al. 2013).

An important property of centering is that it makes the inner product between a sample and the centroid c uniform (actually zero). To see this, consider the similarity of a sample $x \in D$ to a given query $q \in \mathbb{R}^d$, given by their inner product:

$$\mathbf{Sim}(x; q) \equiv \langle x, q \rangle.$$

After centering, the similarity is changed to

$$\mathbf{Sim}^{\mathsf{CENT}}(x; q) \equiv \langle x - c, q - c \rangle. \quad (1)$$

Substituting $q = c$, we have for any $x \in \mathbb{R}^d$,

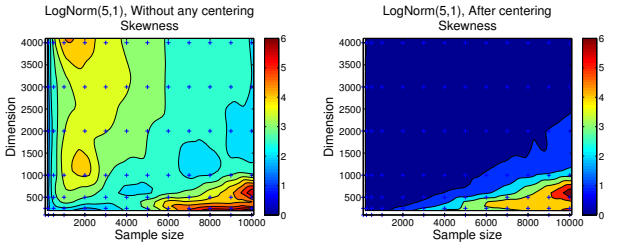
$$\mathbf{Sim}^{\mathsf{CENT}}(x; c) = 0,$$

which means that no samples are specifically similar to the centroid. Thus, hubness is expected to decrease after centering, at least in the dataset used to plot Figure 1(a) and (c), for which samples similar to the centroid constitute hubs.

As expected, S_{N_k} for this dataset decreases from 4.45 to 0.27 after centering; observe that the N_k distribution shown in Figure 1(b) is nearly symmetric. Suzuki et al. (2013) give a more detailed explanation as to why hubs are suppressed by centering.

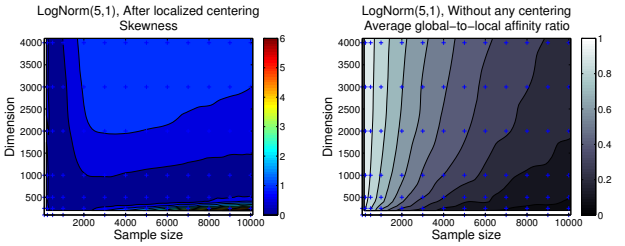
Hubness in Large-Sample Data

Nevertheless, there are cases where hubs are not suppressed by centering. Let us generate a synthetic dataset D in the same way as before, except that this time, the number of samples is larger ($|D| = 10,000$) and the feature space is not especially high-dimensional ($d = 500$). For this data, the skewness $S_{N_{10}}$ of the N_{10} distribution is 5.88 before centering and 5.82 after centering; thus, hubness is not reduced



(a) Skewness (b) Skewness after centering

Figure 3: Contour plots for the skewness of the N_{10} distribution (a) before and (b) after centering, with variations in the sample size and dimensions (the + mark corresponds to a dataset). A warmer color represents the existence of hubness.



(a) Skewness after localized centering (b) Global-to-local affinity ratio

Figure 4: Contour plots for (a) the skewness of the N_{10} distribution after localized centering, and (b) the global-to-local affinity ratio, for the same datasets used in Figure 3.

much. Indeed, as shown in Figure 2(a) and (b), the shape of the N_{10} distribution is nearly identical before and after centering. Moreover, the scatter plot in Figure 2(c) shows that the requisite condition for the success of the centering method—samples with a larger N_k value (i.e. hubs) have a higher similarity to the data centroid—is unmet in this dataset.

We further generate a range of datasets by changing the number of samples $|D|$ between 100 and $10,000$ and dimension d between 100 and 4000. For each dataset, the skewness of the N_{10} distribution is calculated, both before and after the data is centered. For each combination of $|D|$ and d, we generate ten datasets and take the average skewness.

Figure 3 shows the contour plot for the skewness of the resulting N_{10} distribution; panel (a) plots the skewness before centering, and panel (b) the one after centering. In panel (a), hubness occurs in the upper-left and lower-right regions. By comparing the two panels, we see that the hubness in the upper-left region disappears after centering, whereas the one in the lower-right region persists to almost the same degree as it did before centering.

To investigate why centering fails to reduce hubness in the lower-right region, we introduce two quantities defined for each sample: *local affinity* and *global affinity*. We use these quantities to describe how a sample is populated in a dataset.

$\mathbf{LocalAffinity}(x)$ is defined for a sample $x \in D$ as the average similarity between x and the samples belonging to the κNN of x, calculated by

$$\mathbf{LocalAffinity}(x) \equiv \frac{1}{\kappa} \sum_{x' \in \kappa\mathrm{NN}(x)} \langle x, x' \rangle = \langle x, c_\kappa(x) \rangle, \tag{2}$$

where κ, called *local segment size*, is a parameter determining the size of the local neighborhood,[1] and

$$c_\kappa(x) = \frac{1}{\kappa} \sum_{x' \in \kappa\mathrm{NN}(x)} x'$$

is the *local centroid* for x.

$\mathbf{GlobalAffinity}(x)$ is defined as the average similarity between x and all samples in D.

$$\mathbf{GlobalAffinity}(x) \equiv \frac{1}{|D|} \sum_{x' \in D} \langle x, x' \rangle = \langle x, c \rangle, \tag{3}$$

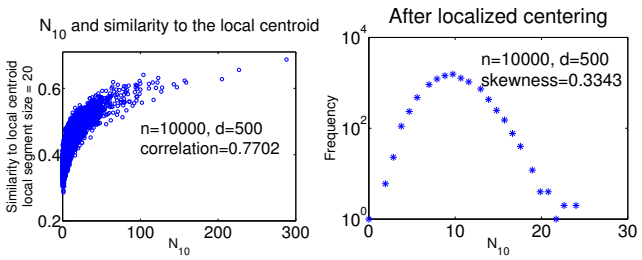
where $c = (1/|D|) \sum_{x' \in D} x'$ is the data centroid. As shown in the last equality, global affinity for x is simply the inner product similarity between x and c.

If D is a set of vectors with non-negative components, then $\mathbf{LocalAffinity}(x) \geq \mathbf{GlobalAffinity}(x) \geq 0$ for all $x \in D$. In text classification, a sample text is normally represented as a (tf-idf weighted) word-count vector, and the above inequality holds. Further note that if $\kappa = |D|$, $\mathbf{LocalAffinity}(x) = \mathbf{GlobalAffinity}(x)$.

We now consider the *global-to-local affinity ratio*, i.e., $\mathbf{GlobalAffinity}(x)/\mathbf{LocalAffinity}(x)$. For a dataset of non-negative vectors, this ratio falls within $[0, 1]$. It can be used as a measure of how a sample x is not "localized" in D; if the ratio is close to zero for some $\kappa \ll |D|$, the sample x has a local set of samples to which x has especially high similarity. If, to the contrary, the ratio is near one regardless of κ, x does not possess this special set of similar samples.

Let us calculate the global-to-local affinity ratio for the datasets used to draw Figure 3. Because this ratio is defined for individual samples, we take the average over all samples in a dataset. Figure 4(b) displays a contour plot in terms of the average ratio, calculated over datasets of varying sample sizes and dimensions. Here, the local segment size is fixed at $\kappa = 20$. Comparing this plot to the one in Figure 3(b), we see that the lower-right region, where the skewness remains

[1]The parameter κ in $\mathbf{LocalAffinity}$ can be different from the parameter k of the kNN classification performed subsequently. Indeed, in later experiments, we will tune κ so as to maximize the correlation with the N_{10} skewness, independently from the kNN classification.



(a) (b)

Figure 5: (a) Correlation between N_{10} and local affinity (before localized centering) and (b) the N_{10} distribution after localized centering applied to the same large-sample dataset used to draw Figure 2.

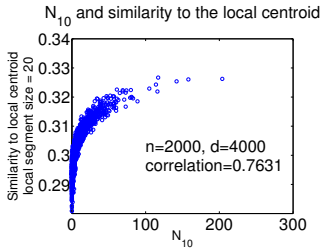


Figure 6: Correlation between N_{10} and local affinity for the same high-dimensional dataset used to draw Figure 1.

high after centering, corresponds to the region in Figure 4(b) where the global-to-local affinity ratio is small. This indicates that samples are "localized" in a dataset in this region, which in turn suggests that the local affinity is more relevant for reducing hubness than the global affinity in such a dataset.

Indeed, for the large-sample dataset ($|D| = 10,000$) of Figure 2, a strong correlation is observed between the N_{10} value and the local affinity (i.e., the similarity to the local centroid), as shown in Figure 5(a). Recall that by contrast, correlation is weak between the N_{10} value and the global affinity (i.e., the similarity to the data centroid), as seen in Figure 2(c).

Furthermore, as shown in Figure 6, the N_{10} value correlates well with the local affinity even for the high-dimensional data ($d = 4000$) of Figure 1. These results indicate that the local affinity can be a general indicator of hubness, not just in large-sample data, but also in high-dimensional data.

Proposed Method: Localized Centering

Inspired by the above observation that the local affinity is a good indicator of hubness, we propose a new method of hubness reduction based on the local affinity, which is effective for both high-dimensional space, and large-sample data.

To this end, we draw analogy from Equation (1), the similarity of a sample $x \in D$ to a query $q \in \mathbb{R}^d$ after centering. When nearest neighbors of a fixed query q is concerned, q is a constant as well as c. Thus from Equation (1),

$$\begin{aligned}\mathbf{Sim}^{\mathsf{CENT}}(x;q) &\equiv \langle x-c, q-c\rangle \\ &= \langle x,q\rangle - \langle x,c\rangle + \text{constant}. \end{aligned} \quad (4)$$

In other words, the global affinity $\langle x,c\rangle$ is subtracted from the original similarity score $\langle x,q\rangle$. In the same vein, the new method, which we call *localized centering*, subtracts the local affinity of x from the original similarity to q, as follows:[2]

$$\mathbf{Sim}^{\mathsf{LCENT}}(x;q) \equiv \langle x,q\rangle - \langle x, c_\kappa(x)\rangle\,. \quad (5)$$

Equation (5) contains the parameter κ to determine the size of the local neighborhood. We systematically select κ depending on the dataset, so that the correlation between $N_k(x)$ and the local affinity $\langle x, c_\kappa(x)\rangle$ is maximized. Note that if $\kappa = |D|$, the proposed method reduces to the standard centering.

After the transformation with Equation (5), for any sample $x \in D$, the similarity to its local centroid $c_\kappa(x)$ is constant, since substituting $q = c_\kappa(x)$ in Equation (5) yields

$$\mathbf{Sim}^{\mathsf{LCENT}}(x; c_\kappa(x)) = 0.$$

In other words, no samples have specifically high similarity to their local centroids after the transformation. Taking into account the observation that the samples with high similarity to their local centroids become hubs, we expect this transformation to reduce hubness. This expectation is borne out for the dataset illustrated in Figure 5(b), where hubs disappear after localized centering.

Localized centering also reduces hubness in other datasets of varying sample sizes and dimensions. In Figure 4(a), we display the contour plot for the skewness of the N_{10} distribution after localized centering. From the figure, we can see that localized centering reduces both type of hubness: the one occurs in large-sample data (corresponding to the lower-right region of the contour plot), and the one occurs in high-dimensional data (the upper-left region).

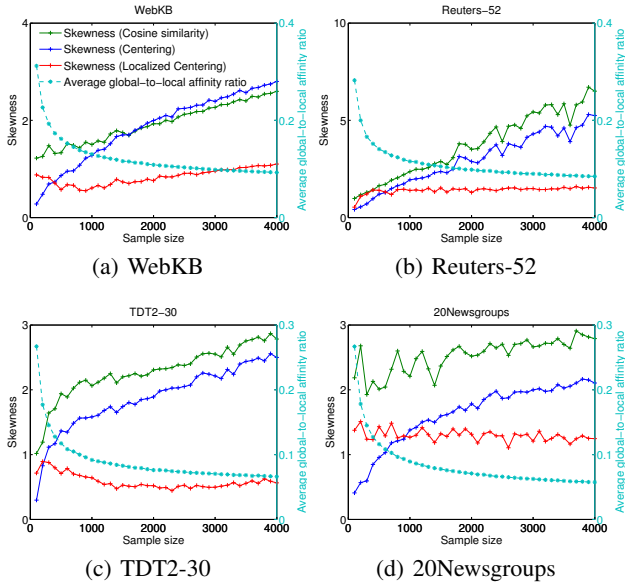
Even for high-dimensional data, the localized centering (Equation (5)) is as effective as the standard centering (Equation (4)). The explanation is that in such a dataset, the global-to-local affinity ratio is close to one as indicated in the upper-left (pale-colored) region of Figure 4(b). This implies $\langle x, c_\kappa(x)\rangle \approx \langle x, c\rangle$ for most samples x in datasets of this region, and hence Equation (5) becomes nearly identical to Equation (4).

We can further extend Equation (5). The second term on the right side of the formula can be interpreted as a penalty term used to render the similarity score smaller depending on how likely x is to become a hub. Now, in order to control the degree of penalty, we introduce a parameter γ, to some extent heuristically, such that

$$\mathbf{Sim}^{\mathsf{LCENT}}_{\gamma}(x;q) \equiv \langle x,q\rangle - \langle x, c_\kappa(x)\rangle^{\gamma}\,. \quad (6)$$

Parameter γ can be tuned so as to maximally reduce the skewness of the N_k distribution.

[2] Unlike the standard centering method, transformation by localized centering can no longer be interpreted as shifting the origin to somewhere in the vector space. Moreover, $\mathbf{Sim}^{\mathsf{LCENT}}$ is not a symmetric function; i.e., $\mathbf{Sim}^{\mathsf{LCENT}}(x;y) = \mathbf{Sim}^{\mathsf{LCENT}}(y;x)$ does not generally hold.



(a) WebKB (b) Reuters-52

(c) TDT2-30 (d) 20Newsgroups

Figure 7: Relation between the sample size and the skewness of the N_{10} distribution: baseline cosine similarity without any centering (green), with centering (blue), and with localized centering (red). The average global-to-local affinity ratio (light blue) is also plotted.

Experiments Using Real-world Data

We applied localized centering to real-world datasets for multiclass document classification, in order to evaluate its effect in reducing hubness and to determine whether it improves the classification accuracy. The datasets are: WebKB, Reuters-52, and 20Newsgroups, all preprocessed and distributed by Cardoso-Cachopo (2007), and TDT2-30 distributed by Cai, He, and Han (2005). We represented each document as a tf-idf weighted "bag-of-word" vector normalized to unit length. Throughout the experiment, inner product is used as the measure of similarity. Therefore the baseline similarity measure is equivalent to cosine, since we use normalized vectors.

Hubness Reduction

In the first experiment, we examined whether localized centering reduces hubness in real large-sample datasets, as it did with the synthetic datasets in the previous sections. We randomly selected a subset of samples from an entire dataset, increasing the size of the subset from 100 to 4000. For each subset size, we repeated the random selection of a subset ten times, and computed the average skewness of the N_{10} distribution with each repetition. We compared the average skewness values before and after centering, and the one after localized centering.

The results are shown in Figure 7. Throughout the four datasets that were examined, we observe the same trend. As the sample size is increased, the N_{10} skewness for the original similarity also increases; i.e., hubness is enlarged. With centering, the skewness also increases with the sample size, even though the amount of increase is slightly smaller than the original similarity measure. By contrast, with localized centering the skewness remains at approximately the same value, regardless of the sample size. Standard centering is effective when the number of samples is small (i.e., fewer than approximately 500), but not for datasets with more samples. Localized centering on the other hand keeps hubness at a low level even for large datasets.

Moreover, we calculated the average global-to-local affinity ratio for each dataset using Equations (2) and (3), and overprinted the plot in Figure 7. Here, we clearly observe the same tendency as previously observed with the synthetic datasets: When the global-to-local affinity ratio is small, localized centering suppresses hubness, whereas standard centering fails to do so.

kNN Classification Accuracy

We examined whether reduction of hubness by localized centering leads to an improved kNN classification accuracy. The task is to classify a document into one of the predefined categories. To simulate a situation in which the number of training samples is large, we ignored the predefined training-test splits provided with the datasets. Instead, the performance was evaluated by the accuracy of the leave-one-out cross validation over all samples. We predicted the label for a test sample (document) with kNN classification, using the remaining documents as training samples, and calculated the rate of correct predictions.

Besides the baseline inner product (cosine) similarity (COS), we tried five similarity measures transformed from the baseline similarity: the standard centering (CENT), localized centering (LCENT)[3], commute-time kernel (CT), mutual proximity (MP), and local scaling (LS). The commute-time kernel was originally presented in Saerens et al. (2004) to define the similarity between graph nodes based on graph Laplacian, and was later shown effective in reducing hubness (Suzuki et al. 2012).

Mutual proximity[4] (Schnitzer et al. 2012) and *local scaling* (Zelnik-Manor and Perona 2005) are hub reduction methods for distance metrics, which attempt to symmetrize the nearest-neighbor relations by rescaling the distance between samples. Since the value calculated as one minus the cosine (i.e., the baseline similarity) can be considered a distance, MP and LS are applicable to the task here.

The results are shown in Table 1. Methods based on local transformations (i.e., LCENT and LS) achieved the best overall performance in terms of accuracy. The standard centering (CENT) results in improved accuracy and skewness within a small margin of the baseline similarity. In contrast, the localized centering (LCENT) reduces skewness (i.e., reduces hubness) and increases accuracy considerably.

[3] As mentioned previously, the parameters κ and γ in Equation (6) were selected with respect to the N_{10} skewness, without using the label for the data.

[4] We used a MATLAB script norm_mp_empiric.m distributed at http://ofai.at/~dominik.schnitzer/mp.

(a) WebKB (4168 samples, 7770 features, 4 classes)

	Accuracy						Skewness					
k	COS	CENT	LCENT	CT	MP	LS	COS	CENT	LCENT	CT	MP	LS
10	0.753	0.756	0.761	0.743	0.761	**0.764**	2.64	2.85	1.13	3.52	0.74	1.15
20	0.769	0.766	0.774	0.757	0.771	**0.777**	2.01	2.12	0.87	2.51	0.69	0.65
30	0.770	0.776	0.780	0.764	0.780	**0.786**	1.74	1.70	0.73	2.06	0.65	0.47
40	0.777	0.778	0.781	0.772	0.778	**0.785**	1.63	1.42	0.66	1.78	0.52	0.34
50	0.776	0.780	0.785	0.781	0.783	**0.790**	1.57	1.25	0.63	1.60	0.39	0.22

(b) Reuters-52 (9100 samples, 19, 241 features, 52 classes)

	Accuracy						Skewness					
k	COS	CENT	LCENT	CT	MP	LS	COS	CENT	LCENT	CT	MP	LS
10	0.872	0.885	**0.901**	0.858	0.898	0.895	14.82	11.04	1.76	6.93	0.82	2.36
20	0.876	0.894	**0.913**	0.885	0.908	0.904	7.92	6.42	1.58	4.92	0.77	1.76
30	0.875	0.896	**0.913**	0.889	0.909	0.904	5.74	4.64	1.61	4.04	0.78	1.37
40	0.869	0.896	**0.913**	0.889	0.903	0.901	4.68	3.77	1.63	3.55	0.81	1.08
50	0.866	0.894	**0.910**	0.892	0.900	0.897	4.02	3.27	1.63	3.22	0.80	0.90

(c) TDT2-30 (9394 samples, 36, 093 features, 30 classes)

	Accuracy						Skewness					
k	COS	CENT	LCENT	CT	MP	LS	COS	CENT	LCENT	CT	MP	LS
10	**0.964**	0.963	**0.964**	0.953	0.962	0.961	3.63	3.18	0.91	4.20	0.47	0.80
20	**0.965**	0.963	0.964	0.955	0.962	0.962	3.13	2.81	0.81	3.09	0.38	0.77
30	**0.966**	0.964	**0.966**	0.958	0.963	0.963	2.72	2.42	0.45	2.62	0.31	0.71
40	0.965	0.963	**0.966**	0.959	0.963	0.963	2.50	2.16	0.45	2.34	0.25	0.64
50	0.965	0.963	**0.967**	0.960	0.963	0.965	2.36	1.98	0.46	2.15	0.22	0.61

(d) 20Newsgroups (18, 820 samples, 70, 216 features, 20 classes)

	Accuracy						Skewness					
k	COS	CENT	LCENT	CT	MP	LS	COS	CENT	LCENT	CT	MP	LS
10	0.859	0.860	**0.877**	0.838	0.865	0.874	3.25	2.99	0.81	5.53	0.49	0.77
20	0.845	0.845	0.861	0.841	0.852	**0.862**	3.14	2.69	0.89	4.09	0.49	12.90
30	0.834	0.836	0.849	0.837	0.846	**0.855**	3.06	2.47	1.05	3.42	0.43	6.71
40	0.831	0.832	0.841	0.833	0.840	**0.849**	3.04	2.37	1.27	2.97	0.39	4.54
50	0.825	0.827	0.835	0.833	0.835	**0.844**	3.01	2.27	1.36	2.67	0.36	3.06

Table 1: Accuracy (higher the better) of document classification via kNN, and skewness (smaller the better) for the N_k distribution calculated using five similarity or distance measures for different k. COS: cosine similarity, CENT: centering, LCENT: localized centering, CT: commute-time kernel, MP: mutual proximity, LS: local scaling. The bold letter indiates best accuracy.

Discussion and Related Work

Radovanović, Nanopoulos, and Ivanović (2010b) remarked that hubs tend to be more similar to their respective cluster centers. However, to find clusters we have to choose the right clustering algorithms and parameters (such as the number of clusters for the K-means clustering). The localized centering instead employs the local centroids of individual samples to detect hubs. Local centroids are straightforward to compute, and the parameter κ can be systematically tuned with respect to the N_k skewness.

The existence of hubness in a dataset depends on two factors: (1) high intrinsic dimensionality, and (2) spatial centrality, in the sense that a "central" point exists in the data space with respect to the given distance (or similarity) measure (Radovanović, Nanopoulos, and Ivanović 2010a). If at least one of the two factors is absent, hubness will not emerge. Localized centering can be understood as an approach that eliminates the second factor.

A key observation behind localized centering is the connection between the hubness occurring in a large-sample dataset and the low global-to-local affinity ratio; i.e., we pointed out that samples are "localized" and not uniformly populated in the dataset. On the other hand, Low et al. (2013) argued that the hubness phenomenon is directly related to the existence of density gradients, rather than high dimensionality, and hence the phenomenon can be made to occur in low-dimensional data. Although our argument is based on similarity measures, whereas Low et al. (2013) assume distance, the two arguments seem consistent, as the non-uniformity of a sample population in a dataset can be a source of a density gradient, and therefore a cause of hubness.

Localized centering and local scaling (Zelnik-Manor and Perona 2005) have similar formulations, in that both penalize the original similarity/distance on the basis of the information on the neighborhood of individual samples. Moreover, localized centering subtracts the local affinity from the original similarity score, whereas local scaling divides the original distance by the *local scale*, which is the distance to the kth nearest neighbor. It is intriguing that these formulations are derived with different objectives in mind: Local-

ized centering tries to decrease the correlation between N_k and local affinity, but local scaling aims at making neighborhood relations more symmetric.

Concerning the problematic behavior of large-sample data, von Luxburg, Radl, and Hein (2010) showed that the commute-time distance becomes meaningless in a graph with a large number of nodes; in the limit, it gives the same distance rankings regardless of the query node, which implies that the top-ranked nodes are in fact hubs.

Conclusion

Although hubness is known to occur in high-dimensional spaces, we observed that hubness also occurs when the number of samples is large relative to the dimension of the space. Analyzing the difference between these two cases of hubness, we proposed localized centering, a new hub-reduction method that works effectively for both types of hubness. Using real-world data, we demonstrated that localized centering effectively reduces hubness and improves the performance of document classification. A theoretical analysis of hubness reduction with localized centering shall be pursued in future work.

References

Cai, D.; He, X.; and Han, J. 2005. Document clustering using locality preserving indexing. *IEEE Transactions on Knowledge and Data Engineering* 17(12):1624–1637. Datasets available at http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html.

Cardoso-Cachopo, A. 2007. *Improving Methods for Single-label Text Categorization*. Phd thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa. Datasets available at http://web.ist.utl.pt/acardoso/datasets/.

Colas, F., and Brazdil, P. 2006. Comparison of SVM and some older classification algorithms in text classification tasks. *Artificial Intelligence in Theory and Practice* 169–178.

Cover, T. M., and Hart, P. E. 1967. Nearest-neighbor pattern classification. *IEEE Transactions on Information Theory* 13(1):21–27.

Low, T.; Borgelt, C.; Stober, S.; and Nrnberger, A. 2013. The hubness phenomenon: Fact or artifact? In *Towards Advanced Data Analysis by Combining Soft Computing and Statistics*, volume 285 of *Studies in Fuzziness and Soft Computing*. Springer. 267–278.

Radovanović, M.; Nanopoulos, A.; and Ivanović, M. 2010a. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research* 11:2487–2531.

Radovanović, M.; Nanopoulos, A.; and Ivanović, M. 2010b. On the existence of obstinate results in vector space models. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, 186–193.

Saerens, M.; Fouss, F.; Yen, L.; and Dupont, P. 2004. The principal components analysis of graph, and its relationships to spectral clustering. In *Proceedings of the 15th European Conference on Machine Learning (ECML '04)*, 371–383.

Schnitzer, D.; Flexer, A.; Schedl, M.; and Widmer, G. 2012. Local and global scaling reduce hubs in space. *Journal of Machine Learning Research* 13(1):2871–2902.

Suzuki, I.; Hara, K.; Shimbo, M.; Matsumoto, Y.; and Saerens, M. 2012. Investigating the effectiveness of Laplacian-based kernels in hub reduction. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Suzuki, I.; Hara, K.; Shimbo, M.; Saerens, M.; and Fukumizu, K. 2013. Centering similarity measures to reduce hubs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 613–623.

von Luxburg, U.; Radl, A.; and Hein, M. 2010. Getting lost in space: Large sample analysis of the resistance distance. In *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc. 2622–2630.

Yang, Y., and Liu, X. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, 42–49.

Zelnik-Manor, L., and Perona, P. 2005. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*. MIT Press. 1601–1608.