

# Flattening the Density Gradient for Eliminating Spatial Centrality to Reduce Hubness

**Kazuo Hara\***

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

**Ikumi Suzuki\***

suzuki.ikumi@gmail.com  
Yamagata University  
Yonezawa, Yamagata, 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

*Spatial centrality*, whereby samples closer to the center of a dataset tend to be closer to all other samples, is regarded as one source of *hubness*. Hubness is well known to degrade  $k$ -nearest-neighbor ( $k$ -NN) classification. Spatial centrality can be removed by *centering*, i.e., shifting the origin to the global center of the dataset, in cases where inner product similarity is used. However, when Euclidean distance is used, centering has no effect on spatial centrality because the distance between the samples is the same before and after centering. As described in this paper, we propose a solution for the hubness problem when Euclidean distance is considered. We provide a theoretical explanation to demonstrate how the solution eliminates spatial centrality and reduces hubness. We then present some discussion of the reason the proposed solution works, from a viewpoint of *density gradient*, which is regarded as the origin of spatial centrality and hubness. We demonstrate that the solution corresponds to flattening the density gradient. Using real-world datasets, we demonstrate that the proposed method improves  $k$ -NN classification performance and outperforms an existing hub-reduction method.

## Introduction

### Background

The  $k$ -nearest neighbor ( $k$ -NN) classifier is vulnerable to the *hubness* problem, which is a phenomenon that occurs in high-dimensional spaces (Radovanović, Nanopoulos, and Ivanović 2010; Schnitzer et al. 2012; Suzuki et al. 2013; Tomašev and Mladenić 2013). Hubness refers to the property by which some samples in a dataset become *hubs*, frequently occurring in the  $k$ -NN lists of other samples. The emergence of hubs often affects  $k$ -NN classification accuracy. The predicted label of a query sample is determined by

the labels of its  $k$ -NN samples, in which hubs are likely to be included.

According to Radovanović et al. (2010), hubness occurs because of the existence of spatial centrality and high dimensionality. *Spatial centrality* is the tendency of samples that are closer to the center of a dataset to be closer to all other samples. As dimensionality increases, this tendency of such samples is amplified, causing the samples closer to the center to become hubs.

To reduce hubness, Suzuki et al. (2013) showed that shifting the origin to the global centroid, known as *centering*, is effective when an inner product-based similarity is used. More precisely, in a high-dimensional dataset with a global centroid vector  $c$ , hubness occurs when the inner product  $\langle x_i, x_j \rangle$  is used to measure the similarity between samples  $x_i$  and  $x_j$ . However, hubness does not occur if  $\langle x_i - c, x_j - c \rangle$  is used instead. This result occurs because of the elimination of spatial centrality through the process of centering. However, the centering process cannot eliminate hubness as measured by Euclidean distance because the distance between samples remains the same before and after the centering.

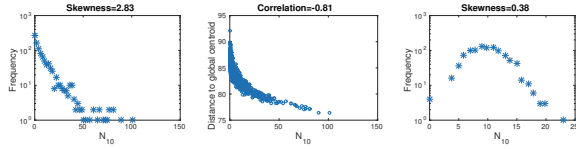
### Contributions

We propose a solution for the hubness problem considering Euclidean distance, rather than inner product similarity. We introduce a value called the *sample-wise centrality*: for each sample  $x$ , we define this value as the (squared) distance from the global centroid vector  $c$ ,  $\|x - c\|^2$ . The proposed method subtracts the sample-wise centrality from the (squared) original Euclidean distance. Subsequently, we provide a theoretical explanation of how the solution eliminates the spatial centrality and reduces hubness.

As our second contribution, from a viewpoint of *density gradient*, we explain why the proposed solution works, i.e., the reason for the reduction of hubness by the elimination of spatial centrality. After verifying that the origin of hubness lies in the density gradient and high-dimensionality (Low et al. 2013), we demonstrate that subtracting sample-wise centrality from the (squared) original Euclidean distance flattens

\*Equally contributed

Copyright © 2016, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



(a) Hubness exists. (b) Spatial centrality. (c) No hubness.

Figure 1: Illustrative example using a dataset generated from an i.i.d.  $Gaussian(\vec{0}, \mathbf{I})$  with sample size  $n = 1000$  and dimension  $d = 1000$ . (a) Hubness occurs when samples have a large  $N_{10}$  value, and when the  $N_{10}$  distribution is skewed to the right. (b) Correlation between the  $N_{10}$  value and the distance to the global centroid is strong. (c) Hubness is reduced successfully (with lower  $N_{10}$  value and skewness) by the proposed transformation from Equation (3).

the density gradient of the isotropic Gaussian distribution.

However, when dealing with real-world datasets, the form of distributions from which datasets are generated is not generally the isotropic Gaussian distribution. As our third contribution, we propose a new hub-reduction method for such practical situations. The method relies on the assumption that each sample locally follows an isotropic Gaussian distribution. Using real-world datasets involving gene expression profile classification, and handwritten digit or spoken letter recognition, we demonstrate that the proposed method improves  $k$ -NN classification performance and that it outperforms an existing hub-reduction method.

### Property of Hubness

The *hubness* phenomenon is known to occur when nearest neighbors in high-dimensional spaces are considered (Radovanović, Nanopoulos, and Ivanović 2010). Letting  $D \subset \mathbb{R}^d$  be a dataset in  $d$ -dimensional space and letting  $N_k(x)$  denote the number of times a sample  $x \in D$  occurs in the  $k$  NNs of other samples in  $D$ , then the shape of the  $N_k$  distribution skews to the right, and a few samples have large  $N_k$  values when the dimension is large. Such samples that are close to many other samples are called *hubs*. This phenomenon is known as *hubness*.

Here, we demonstrate the emergence of hubness using artificial data. We generate a dataset from an i.i.d.  $Gaussian(\vec{0}, \mathbf{I})$  with sample size  $n = 1000$  and dimension  $d = 1000$ , where  $\vec{0}$  is a  $d$ -dimensional vector of zeros and  $\mathbf{I}$  is a  $d \times d$  identity matrix. The distribution of  $N_{10}$  is presented in Figure 1(a), where one can observe the presence of hubs, i.e., samples with particularly large  $N_{10}$  values.

Following Radovanović et al. (2010), we evaluate the degree of hubness by the skewness of the  $N_k$  distribution. Skewness is a standard measure of the degree of symmetry in a distribution. Its value, which is zero for a symmetric distribution such as a Gaussian distribution, takes positive or negative values for distributions with a long right or left tail. Particularly, a large skewness indicates strong hubness in a dataset. Indeed, skewness is large, i.e., 2.83, in Figure 1(a).

### Spatial Centrality

For the artificial dataset described above, we present a scatter plot of samples with respect to the  $N_{10}$  value and the distance to the center of the dataset (Fig. 1(b)). Clearly, a strong correlation exists. It is called *spatial centrality* (Radovanović, Nanopoulos, and Ivanović 2010).

Spatial centrality refers to the fact that samples closer to the center of a dataset tend to be closer to other samples, and therefore, tend to have large  $N_k$  values. We now show that the emergence of spatial centrality is inherent in the (squared) Euclidean distance, where the distance is computed as  $\|x - q\|^2$  between a database sample  $x \in D$  and a query sample  $q \in D$ .<sup>1 2</sup>

**Proposition 1.** *Let us consider two database samples  $a, b \in D$  located proximate to or distant from the global centroid  $c = \frac{1}{|D|} \sum_{q \in D} q \equiv \mathbb{E}_q[q]$ , such that*

$$\|a - c\|^2 \leq \|b - c\|^2. \quad (1)$$

Then

$$\mathbb{E}_q[\|a - q\|^2] \leq \mathbb{E}_q[\|b - q\|^2]. \quad (2)$$

*Proof.* Because Equation (1) is equivalent to

$$-2\langle a, c \rangle + \|a\|^2 + 2\langle b, c \rangle - \|b\|^2 \leq 0,$$

$$\begin{aligned} & \mathbb{E}_q[\|a - q\|^2] - \mathbb{E}_q[\|b - q\|^2] \\ &= -2\langle a, \mathbb{E}_q[q] \rangle + \|a\|^2 + 2\langle b, \mathbb{E}_q[q] \rangle - \|b\|^2 \\ &= -2\langle a, c \rangle + \|a\|^2 + 2\langle b, c \rangle - \|b\|^2 \leq 0. \end{aligned}$$

Therefore, we obtain  $\mathbb{E}_q[\|a - q\|^2] \leq \mathbb{E}_q[\|b - q\|^2]$ .  $\square$

Proposition 1 suggests that, on average, sample  $a$ , which is near the global centroid, is closer to the query samples than sample  $b$ , which is distant from the centroid. Therefore, spatial centrality exists in the squared Euclidean distance.<sup>3</sup>

Using Euclidean distance or squared Euclidean distance does not affect the performance of subsequent  $k$ -NN classification because the nearest neighbors of a query sample selected from database samples are the same irrespective of the metric used. Therefore, we continue to use the squared Euclidean distance in the sections below.

### Solution for Reducing Hubness by Eliminating Spatial Centrality

The existence of spatial centrality is regarded as one of the principal causes for hubness (Radovanović, Nanopoulos, and Ivanović 2010). Therefore, we expect that hubness

<sup>1</sup>We use the terminology “database sample” and “query sample” because we assume  $k$ -NN classification by which database samples are sorted in ascending order based on the distance from a given query sample.

<sup>2</sup>For brevity, we consider the case where the set of samples in the database and the queried set of samples are identical.

<sup>3</sup>A similar argument using Euclidean distance was presented in a report of an earlier study (Radovanović, Nanopoulos, and Ivanović 2010), where samples were assumed to follow the Gaussian distribution. In our argument, however, Inequality (2) holds for any distribution.

will be suppressed if spatial centrality is removed. In this section, we propose a hub-reduction method that transforms the (squared) Euclidean distance such that the transformed distance does not generate spatial centrality.

As noted previously, we do not consider the Euclidean distance, but instead work with the squared Euclidean distance. Therefore, for a given query sample  $q \in D$  and database sample  $x \in D$ , we use the squared Euclidean distance  $\|x - q\|^2$ .

To remove spatial centrality with respect to the global centroid, we define *sample-wise centrality* for database sample  $x$  and query sample  $q$ , respectively as  $\|x - c\|^2$  and  $\|q - c\|^2$ , which are the (squared) distances from the (global) centroid  $c$ . We then transform the squared Euclidean distance by subtracting the sample-wise centrality of  $x$  and  $q$ , such that

$$DisSim^{Global}(x, q) \equiv \|x - q\|^2 - \|x - c\|^2 - \|q - c\|^2. \quad (3)$$

This can take a negative value. Therefore, it is regarded as a *dissimilarity* (i.e., we designate by  $DisSim$ ), not *distance*. However, non-negativity does not affect the process of  $k$ -NN classification, where database samples are sorted in ascending order based on their dissimilarity with a given query sample  $q$ .

Now, substituting  $q = c$  in Equation (3) yields

$$DisSim^{Global}(x, c) = 0. \quad (4)$$

This fact indicates that the dissimilarity of any database sample  $x \in D$  with the centroid  $c$  is the same (i.e., 0). In other words, Equation (4) implies that spatial centrality does not exist after the transformation, because no samples have a specific small dissimilarity with the centroid.

Next, we show that the transformation based on Equation (3) reduces hubness.

**Theorem 1.** *The mean of the dissimilarity defined in Equation (3) between a database sample  $x \in D$  and query samples is constant, i.e.,*

$$\mathbb{E}_q[DisSim^{Global}(x, q)] = const.$$

*Proof.* Using  $c = \mathbb{E}_q[q]$ ,

$$\begin{aligned} \mathbb{E}_q[DisSim^{Global}(x, q)] &= \mathbb{E}_q[-2(\langle x, q \rangle - \langle x, c \rangle - \langle q, c \rangle + \|c\|^2)] \\ &= -2(\langle x, \mathbb{E}_q[q] \rangle - \langle x, c \rangle - \langle \mathbb{E}_q[q], c \rangle + \|c\|^2) \\ &= -2(\langle x, c \rangle - \langle x, c \rangle - \langle c, c \rangle + \|c\|^2) = 0, \end{aligned}$$

which takes a constant value (i.e., 0) that is independent of database sample  $x$ .  $\square$

Theorem 1 shows that any two database samples  $a, b \in D$  are equally close to the query samples on average, even if they are selected to satisfy Inequality (1). Recall that, without the proposed transformation, sample  $a$  is closer to the query samples on average than sample  $b$ , as indicated by Inequality (2). In contrast, the proposed method does not cause some database samples to be specifically closer to the query samples. Therefore the proposed method is expected to suppress hubness.

Indeed, the proposed method reduces hubness in the dataset used to create Figure 1. After the proposed transformation according to Equation (3) has been applied, the skewness decreases from 2.83 (Fig. 1(a)) to 0.38 (Fig. 1(c)).

## Why the Solution Works?

Although Fig. 1 (b) shows strong correlation between hubness (i.e.,  $N_k$  value) and centrality (i.e., distance to global centroid), it does not mean in general that equalizing the distance to the global centroid reduces hubness. This section, presents the reason that the proposed solution works from a viewpoint of *density gradient*.

### Density Gradient: A Cause of Hubness

The origin of hubness can also be viewed to lie in density gradient and high dimensionality (Low et al. 2013). To illustrate, we consider a dataset in which each sample in the dataset is a real-valued vector  $x$  generated from continuous probability density function  $f(x)$ . If the value of  $f(x)$  varies over  $x$ , then we say that the dataset has a density gradient, which means that samples are concentrated around the region where  $f(x)$  is large, but samples are sparsely observed in the region where  $f(x)$  is small. In other words, a density gradient exists in any dataset generated from a distribution other than a uniform distribution in which  $f(x)$  takes a constant value irrespective of  $x$ .<sup>4</sup>

We now discuss relations between density gradient and hubness, using the isotropic Gaussian distribution as an example of a density gradient, and the uniform distribution that has no density gradient.

We start from an observation in one dimension. In Figure 2, the 1-NN relations between samples are represented as arrows. Each sample has an out-going arrow. The sample indicated by the arrow is the closest sample (i.e., 1-NN sample) to the sample in which the arrow goes out.

It is noteworthy that the directions of the arrows are random in datasets generated from the uniform distribution (Fig. 2(b)), but they are not random in datasets generated from the Gaussian distribution. Precisely, the arrows tend to direct to the center of the dataset (Fig. 2(a)) because the closer to the center a point  $x$  lies, the greater the density  $f(x)$  of the Gaussian distribution becomes, meaning that samples are more likely generated in the region closer to the center. As a result, samples closer to the center are likely to be selected as 1-NN by samples that are more distant from the center. In contrast, in the case of the uniform distribution, all samples are equally likely to be selected as 1-NN, irrespective of their position.

However, in one dimension, hubs (which are the samples with a large  $N_1$  value here) do not occur because  $N_1$  takes a value from  $\{0, 1, 2\}$ . Therefore, the maximum of  $N_1$  cannot become greater than 2. Consequently, the existence of density gradient is insufficient for hubness to occur.

We then proceed to the case of a higher dimension. Figure 2(c) shows 1-NN relations between samples generated from the isotropic two-dimensional Gaussian distribution.

<sup>4</sup>If a uniform distribution has a bounded support, hubness actually appears. This is because boundaries exist—a region where  $f(x)$  is constant, and elsewhere  $f(x) = 0$ , so there will be a sharp density increase/drop. Hence, to be precise, a dataset generated by a boundless uniform distribution (e.g., a uniform distribution on a spherical surface) or a *Poisson process* (uniformly spread all over the space) does not have a density gradient.

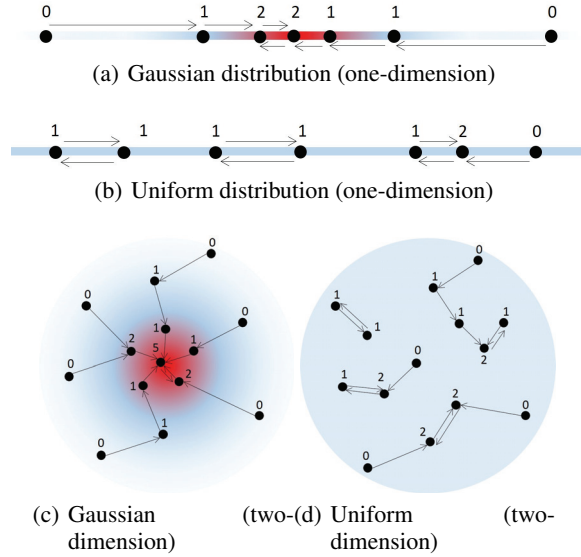


Figure 2: For datasets generated illustratively from one or two-dimensional Gaussian and uniform distributions, the 1-NN relations between samples are indicated by arrows. The red regions represent high density whereas shaded blue stands for low density. The number attached to a sample denotes the  $N_1$  value for the sample. The sample with a large  $N_1$  value is a hub.

Like that in one dimension, a tendency is apparent by which arrows are directed towards the center because of the existence of density gradient. However, contrary to the one-dimensional case, the upper limit of  $N_1$  is not 2 but it takes the value of the *kissing number*, which grows exponentially with the number of dimensions (Conway, Sloane, and Bannai 1987). Consequently, in higher dimensions, samples closer to the center tend to be pointed by many arrows starting from samples farther from the center. Therefore, they become hubs.

### Flattening the Density Gradient

Necessary ingredients that cause hubness include density gradient and high dimensionality. Therefore, we can expect that a method that is intended to flatten the density gradient is effective to reduce hubness.<sup>5</sup>

Herein, we show that the hub-reduction method according to Equation (3) in fact flattens the density gradient of the isotropic Gaussian distribution. The isotropic Gaussian assumes squared Euclidean distance  $d(x, y) = \|x - y\|^2$  between two points  $x$  and  $y$ . Using the distance, its probability density function  $f(x)$  is defined as

$$f(x) \propto \exp(-\beta d(x, \mu)) = \exp(-\beta \|x - \mu\|^2),$$

<sup>5</sup>Dimensionality reduction is also effective to reduce hubness. However, it has been pointed out in earlier studies that if dimensionality is reduced below intrinsic dimensionality, then a loss of information can occur (Radovanović, Nanopoulos, and Ivanović 2010).

where  $\mu$  and  $\beta$  respectively denote location and scale parameters.

Then, we assume that we are given a dataset  $D$  generated from the isotropic Gaussian, and that  $c = \frac{1}{|D|} \sum_{x \in D} x$  denotes the center, or the global centroid of the dataset. The center  $c$  is known to approach to  $\mu$  (i.e.,  $c \approx \mu$ ) as the size of dataset becomes large.

If the squared distance is transformed using Equation (3) such that  $d_{new}(x, y) = \|x - y\|^2 - \|x - c\|^2 - \|y - c\|^2$ , then the density function is also modified, which gives

$$\begin{aligned} f_{new}(x) &\propto \exp(-\beta d_{new}(x, \mu)) \\ &= \exp(-\beta (\|x - \mu\|^2 - \|x - c\|^2 - \|\mu - c\|^2)) \\ &\approx \exp(-\beta (\|x - \mu\|^2 - \|x - \mu\|^2 - \|\mu - \mu\|^2)) \\ &= \exp(0) = 1. \end{aligned}$$

Therefore, the density becomes constant irrespective of  $x$ . In other words, **the density gradient flattens or disappears**. Consequently, the transformation using Equation (3) reduces the hubness occurring in a dataset generated from the isotropic Gaussian distribution.

### Derivation of the Solution

Thus far, we assumed that the solution is given in the form of Equation (3). Although we have discussed the benefits of using it, i.e., reduction of hubness through elimination of spatial centrality or flattening of the density gradient, we have not yet described the rationale related to the derivation of the solution. We address this issue next, with presentation of some necessary assumptions for the derivation.

Assuming that we are given a distance function  $d(x, y)$  and a probability density function

$$f(x) = \frac{1}{\gamma_d} \exp(-\beta d(x, \mu)),$$

where  $\gamma_d = \int \exp(-\beta d(x, \mu)) dx$ . Then the goal is to obtain a new dissimilarity function  $d_{new}(x, y)$  by remaking  $d(x, y)$  so that the resulting density function  $f_{new}(x) \propto \exp(-\beta d_{new}(x, \mu))$  has no density gradient.

The simplest but trivial solution to obtaining  $f_{new}(x) = const$  is to give  $d_{new}(x, y) = const$ . However, this is not desirable because the solution ignores relations between points that are provided originally by  $d(x, y)$ . To avoid this, we consider a loss function

$$\int (d(x, y) - d_{new}(x, y)) f(x) f(y) dx dy, \quad (5)$$

and we will find  $d_{new}(x, y)$  that minimizes the loss.

However, the loss is minimized when  $d_{new}(x, y) = d(x, y)$ , which is also undesirable. To avoid this, we introduce a non-negative function  $h(x) \geq 0$  to restrict the new dissimilarity  $d_{new}(x, y)$  in the form

$$d_{new}(x, y) = d(x, y) - h(x) - h(y). \quad (6)$$

In other words, the new dissimilarity  $d_{new}(x, y)$  is restricted to those obtained by subtracting  $h(x)$  and  $h(y)$ , the factors for discounting dissimilarity depending on individual points. Then, to prohibit  $h(x) = 0$  that yields  $d_{new}(x, y) = d(x, y)$ ,

using a fixed value  $\gamma$  that does not depend on  $h(x)$ , we make a rather strong assumption<sup>6</sup>

$$\int \exp(-\beta h(x)) dx = \gamma,$$

which means that  $\psi(x) = \frac{1}{\gamma} \exp(-\beta h(x))$  provides a probability density function having the same scale parameter  $\beta$  used in the given probability density function  $f(x)$ .

Consequently, the loss (i.e., Equation (5)) becomes

$$\int (h(x) + h(y)) f(x) f(y) dx dy = 2 \int h(x) f(x) dx,$$

and the problem reduces to finding  $h(x) = -\frac{1}{\beta} \log(\psi(x)) - \frac{1}{\beta} \log \gamma$  that minimizes the loss, which is equivalent to finding  $\psi(x)$  that minimizes

$$- \int \log(\psi(x)) f(x) dx.$$

This process is known as cross-entropy minimization, which gives

$$\psi(x) = f(x).$$

Then, by taking the logarithm of both sides of the above equation, the form of  $h(x)$  is determined as

$$h(x) = d(x, \mu) - \frac{1}{\beta} \log\left(\frac{\gamma}{\gamma_d}\right).$$

Finally, replacing  $h(x)$  in Equation (6) with  $d(x, \mu)$  produces  $d_{new}(x, y) = d(x, y) - d(x, \mu) - d(y, \mu) + \frac{2}{\beta} \log\left(\frac{\gamma}{\gamma_d}\right)$ , and omitting the constant term which has no effect on determining  $f_{new}(x)$ , we obtain

$$d_{new}(x, y) = d(x, y) - d(x, \mu) - d(y, \mu). \quad (7)$$

Note that Equation (7) is a general form of Equation (3). Applying  $d(x, y) = \|x - y\|^2$  to Equation (7) with a substitution of  $c$  (the center, or the global centroid of the dataset) for the parameter  $\mu$  yields Equation (3).

## A More Practical Solution

Up to this point, we assumed that datasets are generated from the isotropic Gaussian, but this is not always true. In addition, it may sometimes be impractical to assume that all samples in a dataset follow the same unique distribution.

To be free from the limitations that might arise from such assumptions, we present a more practical solution, which approximates that each sample in a dataset locally follows a different isotropic Gaussian distribution.

More precisely, assuming that distance is given as  $d(x, y) = \|x - y\|^2$ , we approximate that each sample  $x$  in a dataset is generated individually from a probability density function

$$f(x) \propto \exp(-\beta d(x, \mu(x))),$$

where the location parameter  $\mu(x)$  depends on  $x$ . Then, following the derivation described in the previous section,  $h(x)$

<sup>6</sup>Relaxing this assumption is left as a subject for future work.

in Equation (6) is now determined as  $h(x) = d(x, \mu(x))$ . We therefore obtain a new dissimilarity

$$\begin{aligned} d_{new}(x, y) &= d(x, y) - d(x, \mu(x)) - d(y, \mu(y)) \\ &= \|x - y\|^2 - \|x - \mu(x)\|^2 - \|y - \mu(y)\|^2, \end{aligned}$$

which is expected to reduce hubness. To estimate  $\mu(x)$ , we use the *local centroid*,  $c_\kappa(x) = \frac{1}{\kappa} \sum_{x' \in \kappa\text{NN}(x)} x'$ , the mean vector of the  $\kappa$  nearest neighbor samples of  $x$ . By substituting  $c_\kappa(x)$  for  $\mu(x)$ , we obtain the solution below.

$$\begin{aligned} DisSim^{\text{Local}}(x, y) \\ \equiv \|x - y\|^2 - \|x - c_\kappa(x)\|^2 - \|y - c_\kappa(y)\|^2. \quad (8) \end{aligned}$$

For the additional parameter  $\kappa$ , we select a value from  $[1, n-1]$  such that the hubness is maximally reduced, where  $n$  is the dataset size.

## Experiment

### Reduction of Hubness

To evaluate the two proposed dissimilarity measures,  $DisSim^{\text{Global}}$  (Equation (3)) and  $DisSim^{\text{Local}}$  (Equation (8)), we first conducted a simulation study to ascertain whether they reduce hubness, using artificial data.

We generated datasets of three types: one that follows the isotropic Gaussian (i.e., *Gaussian*( $\vec{0}, \mathbf{I}$ ), where  $\vec{0}$  is the all-zeros vector and  $\mathbf{I}$  is the identity matrix), one that follows the non-isotropic Gaussian (i.e., *Gaussian*( $\vec{0}, \mathbf{M}$ ), where  $\mathbf{M}$  is a randomly generated positive-semidefinite matrix), and one that is generated from a mixture of two isotropic Gaussians (i.e., *Gaussian*( $\vec{0}, \mathbf{I}$ ) and *Gaussian*( $\vec{1}, \mathbf{I}$ ), where  $\vec{1}$  is the all-ones vector). For the dataset of each type, we fixed dimension  $d = 1000$ . The number of samples was increased from 500 to 5000. We computed the skewness of the  $N_{10}$  distribution for each dataset and used it to evaluate the hubness (i.e., large skewness denotes the existence of hubness). For each setting, we generated a dataset 10 times and reported the averaged skewness.

The results are presented in Figure 3. Whereas the proposed dissimilarity  $DisSim^{\text{Global}}$  (Equation (3)) greatly reduced hubness for the isotropic Gaussian, it failed to reduce hubness for both of the non-isotropic Gaussian and the mixture of two isotropic Gaussians. However, the proposed dissimilarity  $DisSim^{\text{Local}}$  (Equation (8)) coped effectively with datasets of all three examined types.

### $k$ -NN Classification

We examined whether the reduction of hubness attained using the proposed methods  $DisSim^{\text{Global}}$  (Equation (3)) and  $DisSim^{\text{Local}}$  (Equation (8)) improved the accuracy of  $k$ -NN classification using datasets for gene expression profile classification and handwritten digit or spoken letter recognition. We used the two datasets from the Kent Ridge Biomedical Dataset Repository, Leukemia and Lung Cancer,<sup>7</sup> as well as the two datasets in the UCI machine learning repository, MFeat and ISOLET.<sup>8</sup> The task was to classify a sample into one of the predefined categories. We compared the

<sup>7</sup><http://datam.i2r.a-star.edu.sg/datasets/krbd>

<sup>8</sup><https://archive.ics.uci.edu/ml/datasets.html>

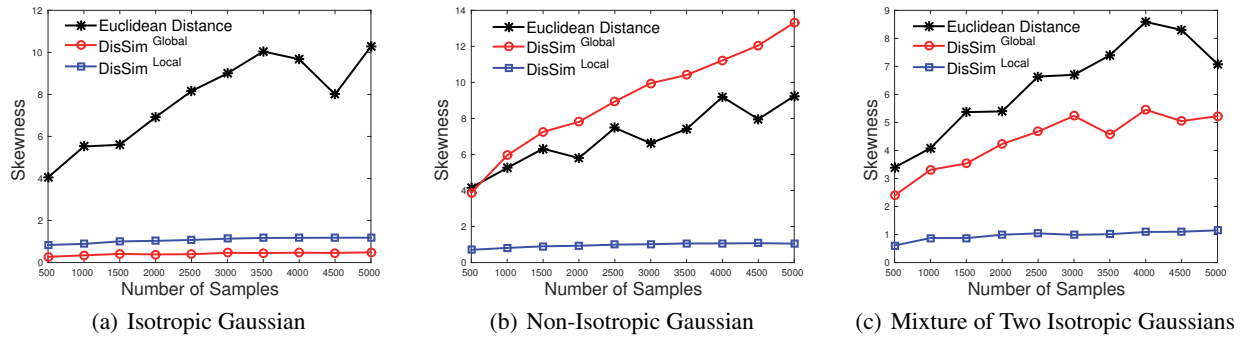


Figure 3: Skewness of the  $N_{10}$  distribution computed using the baseline Euclidean distance and that transformed using the proposed methods, i.e.,  $DisSim^{Global}$  (Equation (3)) and  $DisSim^{Local}$  (Equation (8)). The smaller the skewness was, the better the result which was obtained.

performance of classification using the baseline Euclidean distance and its transformations, obtained using Equation (3) and Equation (8) and mutual proximity<sup>9</sup> (Schnitzer et al. 2012) that renders symmetric the nearest-neighbor relations, which are used with  $k$ -NN classifier. We assessed performance according to the accuracy of the prediction using leave-one-out cross-validation.

Table 1 presents the results. Compared with Euclidean distance, the proposed method  $DisSim^{Local}$  (Equation (8)) reduced skewness and increased the accuracy of  $k$ -NN classification, and outperformed mutual proximity overall.

### Related Work

Previously, the reduction of hubness under Euclidean distance has been studied using the approach that aims to symmetrize the nearest-neighbor relations (Zelnik-Manor and Perona 2005; Schnitzer et al. 2012). In contrast, this paper is the first attempt to explore spatial centrality and density gradient to solve this problem.

For the hubness problem under the inner product similarity, several studies have removed spatial centrality. Among them, this paper was particularly inspired by the studies conducted by Suzuki et al. (2013) who used the similarity to the global centroid, and Hara et al. (2015) who investigated the use of the local centroid.

However, important differences exist between this paper and those two studies: (i) We were interested in the hubness under Euclidean distance, but Suzuki et al. (2013) and Hara et al. (2015) addressed hubness under inner product similarity. (ii) We were aware that the two notions *density gradient* and *spatial centrality* are closely interrelated, and therefore, proposed to flatten the density gradient to reduce hubness, but the two previously mentioned studies merely eliminated the spatial centrality to reduce hubness. (iii) We pointed out that the method using the global or local centroid corresponds to flattening the density gradient of the global or local isotropic Gaussian distribution, but the two previously mentioned studies did not present such a discussion.

<sup>9</sup>We used a MATLAB script `norm_mp_gaussi.m` distributed at <http://ofai.at/~dominik.schnitzer/mp>.

It can be said that our method is among the techniques labeled “Approach 1” in slides presented by Radovanović (2015).<sup>10</sup> Here, hubness is reduced with the expected effect of redistributing responsibility for errors produced by models more uniformly among the points.

According to Bellet, Habrard, and Sebban (2013), most previous studies have been undertaken to improve Euclidean distance, including Weinberger and Saul (2009), who use supervised metric learning, and also do not consider hubness. Our approach is unsupervised, however, so our proposed method would be presented in Table 2 (page 9) of the arXiv Tech Report by Bellet, Habrard, and Sebban (2013) as “Supervision = unsupervised” and “Regularizer = hubness.”

### Conclusion

We proposed a solution for the hubness problem when Euclidean distance is considered. After providing a theoretical explanation for how the solution eliminates *spatial centrality*, a source of hubness, we showed that the solution corresponds to flattening of the *density gradient*, a notion closely related to spatial centrality and hubness. We demonstrated empirically that flattening of the density gradient for eliminating spatial centrality produces an effect on reducing hubness and  $k$ NN classification.

### References

Bellet, A.; Habrard, A.; and Sebban, M. 2013. A survey on metric learning for feature vectors and structured data. *CoRR* abs/1306.6709.

Conway, J. H.; Sloane, N. J. A.; and Bannai, E. 1987. *Sphere-packings, Lattices, and Groups*. New York, NY, USA: Springer-Verlag New York, Inc.

Hara, K.; Suzuki, I.; Shimbo, M.; Kobayashi, K.; Fukumizu, K.; and Radovanović, M. 2015. Localized centering: Reducing hubness in large-sample data. In *Proceedings of the Twenty-Nine AAAI Conference on Artificial Intelligence*, 2645–2651.

<sup>10</sup><http://perun.pmf.uns.ac.rs/radovanovic/publications/Radovanovic-HubsInNNGraphs4.pdf>

Table 1: Accuracy of  $k$ -NN classification and skewness of the  $N_k$  distribution, computed using the baseline Euclidean distance and that transformed using the proposed methods (i.e.,  $DisSim^{Global}$  of Equation (3) and  $DisSim^{Local}$  of Equation (8)) and the mutual proximity, for different  $k \in \{1, 5, 10, 20\}$ . The numbers in bold are the best results.

	Accuracy / Skew ( $k = 1$ )		Accuracy / Skew ( $k = 5$ )		Accuracy / Skew ( $k = 10$ )		Accuracy / Skew ( $k = 20$ )	
Euclidean distance	0.722	11.92	0.817	10.07	0.789	7.52	0.740	4.60
$DisSim^{Global}$	0.774	4.70	0.841	2.43	0.850	1.51	0.875	0.79
$DisSim^{Local}$	0.804	<b>1.94</b>	<b>0.881</b>	<b>1.35</b>	<b>0.881</b>	<b>0.50</b>	<b>0.884</b>	<b>0.01</b>
Mutual proximity	<b>0.838</b>	2.99	0.856	1.77	0.869	1.29	0.838	0.89

(a) Leukemia (327 samples, 10,533 features, 7 classes)

	Accuracy / Skew ( $k = 1$ )		Accuracy / Skew ( $k = 5$ )		Accuracy / Skew ( $k = 10$ )		Accuracy / Skew ( $k = 20$ )	
Euclidean distance	0.887	5.65	0.901	3.21	0.887	2.37	0.818	1.66
$DisSim^{Global}$	0.773	7.90	0.803	3.86	0.833	2.64	0.857	1.49
$DisSim^{Local}$	0.906	<b>0.77</b>	<b>0.931</b>	1.80	<b>0.921</b>	1.50	<b>0.887</b>	0.82
Mutual proximity	<b>0.926</b>	1.38	0.921	<b>1.05</b>	0.911	<b>0.94</b>	0.877	<b>0.77</b>

(b) Lung Cancer (203 samples, 12,600 features, 5 classes)

	Accuracy / Skew ( $k = 1$ )		Accuracy / Skew ( $k = 5$ )		Accuracy / Skew ( $k = 10$ )		Accuracy / Skew ( $k = 20$ )	
Euclidean distance	<b>0.976</b>	1.91	0.981	1.82	0.979	1.60	0.975	1.42
$DisSim^{Global}$	0.956	3.40	0.963	2.38	0.962	1.71	0.953	1.34
$DisSim^{Local}$	0.972	<b>0.83</b>	<b>0.985</b>	<b>0.26</b>	<b>0.983</b>	<b>0.11</b>	<b>0.980</b>	<b>-0.19</b>
Mutual proximity	0.974	1.06	0.980	0.58	0.979	0.44	0.975	0.31

(c) MFeat (2000 samples, 649 features, 10 classes)

	Accuracy / Skew ( $k = 1$ )		Accuracy / Skew ( $k = 5$ )		Accuracy / Skew ( $k = 10$ )		Accuracy / Skew ( $k = 20$ )	
Euclidean distance	0.899	1.70	0.887	2.12	0.915	1.96	0.916	1.73
$DisSim^{Global}$	0.776	18.27	0.772	9.39	0.786	7.16	0.785	5.32
$DisSim^{Local}$	<b>0.915</b>	2.42	<b>0.893</b>	<b>0.78</b>	<b>0.925</b>	<b>0.88</b>	<b>0.932</b>	<b>0.55</b>
Mutual proximity	0.893	<b>1.18</b>	0.879	1.25	0.911	1.11	0.912	0.91

(d) ISOLET (7797 samples, 617 features, 26 classes)

Low, T.; Borgelt, C.; Stober, S.; and Nürnberger, 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 Berlin Heidelberg. 267–278.

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

Radovanović, M. 2015. Hubs in nearest-neighbor graphs: Origins, applications and challenges. NII Shonan Meeting on “Dimensionality and Scalability II: Hands-On Intrinsic Dimensionality”, National Institute of Informatics, Tokyo, Japan, June 28–July 2, 2015.

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.; 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.

Tomašev, N., and Mladenović, D. 2013. Hub co-occurrence modeling for robust high-dimensional kNN classification. In *ECML/PKDD (2)*, volume 8189 of *Lecture Notes in Computer Science*, 643–659. Springer.

Weinberger, K. Q., and Saul, L. K. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research* 10:207–244.

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