On the Surprising Behavior of Distance Metrics in High Dimensional Space

Charu C. Aggarwal[1], Alexander Hinneburg[2], and Daniel A. Keim[2]

[1] IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA.
`charu@watson.ibm.com`
[2] Institute of Computer Science, University of Halle
Kurt-Mothes-Str.1, 06120 Halle (Saale), Germany
{ `hinneburg, keim` }`@informatik.uni-halle.de`

Abstract. In recent years, the effect of the curse of high dimensionality has been studied in great detail on several problems such as clustering, nearest neighbor search, and indexing. In high dimensional space the data becomes sparse, and traditional indexing and algorithmic techniques fail from a efficiency and/or effectiveness perspective. Recent research results show that in high dimensional space, the concept of proximity, distance or nearest neighbor may not even be qualitatively meaningful. In this paper, we view the dimensionality curse from the point of view of the distance metrics which are used to measure the similarity between objects. We specifically examine the behavior of the commonly used L_k norm and show that the problem of meaningfulness in high dimensionality is sensitive to the value of k. For example, this means that the Manhattan distance metric (L_1 norm) is consistently more preferable than the Euclidean distance metric (L_2 norm) for high dimensional data mining applications. Using the intuition derived from our analysis, we introduce and examine a natural extension of the L_k norm to fractional distance metrics. We show that the fractional distance metric provides more meaningful results both from the theoretical and empirical perspective. The results show that fractional distance metrics can significantly improve the effectiveness of standard clustering algorithms such as the k-means algorithm.

1 Introduction

In recent years, high dimensional search and retrieval have become very well studied problems because of the increased importance of data mining applications [1], [2], [3], [4], [5], [8], [10], [11]. Typically, most real applications which require the use of such techniques comprise very high dimensional data. For such applications, the curse of high dimensionality tends to be a major obstacle in the development of data mining techniques in several ways. For example, the performance of similarity indexing structures in high dimensions degrades rapidly, so that each query requires the access of almost all the data [1].

It has been argued in [6], that under certain reasonable assumptions on the data distribution, the ratio of the distances of the nearest and farthest neighbors to a given target in high dimensional space is almost 1 for a wide variety of data distributions and distance functions. In such a case, the nearest neighbor problem becomes ill defined, since the contrast between the distances to different data points does not exist. In such cases, even the concept of proximity may not be meaningful from a qualitative perspective: a problem which is even more fundamental than the performance degradation of high dimensional algorithms.

In most high dimensional applications the choice of the distance metric is not obvious; and the notion for the calculation of similarity is very heuristical. Given the non-contrasting nature of the distribution of distances to a given query point, different measures may provide very different orders of proximity of points to a given query point. There is very little literature on providing guidance for choosing the correct distance measure which results in the most meaningful notion of proximity between two records. Many high dimensional indexing structures and algorithms use the euclidean distance metric as a natural extension of its traditional use in two- or three-dimensional spatial applications. In this paper, we discuss the general behavior of the commonly used L_k norm ($x, y \in \mathcal{R}^d, k \in \mathcal{Z}$, $L_k(x,y) = \sum_{i=1}^{d}(\|x^i - y^i\|^k)^{1/k}$) in high dimensional space. The L_k norm distance function is also susceptible to the dimensionality curse for many classes of data distributions [6]. Our recent results [9] seem to suggest that the L_k-norm may be more relevant for $k = 1$ or 2 than values of $k \geq 3$. In this paper, we provide some surprising theoretical and experimental results in analyzing the dependency of the L_k norm on the value of k. More specifically, we show that the relative contrasts of the distances to a query point depend heavily on the L_k metric used. This provides considerable evidence that the meaningfulness of the L_k norm worsens faster with increasing dimensionality for higher values of k. Thus, for a given problem with a fixed (high) value of the dimensionality d, it may be preferable to use lower values of k. This means that the L_1 distance metric (Manhattan Distance metric) is the most preferable for high dimensional applications, followed by the Euclidean Metric (L_2), then the L_3 metric, and so on. Encouraged by this trend, we examine the behavior of *fractional* distance metrics, in which k is allowed to be a fraction smaller than 1. We show that this metric is even more effective at preserving the meaningfulness of proximity measures. We back up our theoretical results with empirical tests on real and synthetic data showing that the results provided by fractional distance metrics are indeed practically useful. Thus, the results of this paper have strong implications for the choice of distance metrics for high dimensional data mining problems. We specifically show the improvements which can be obtained by applying fractional distance metrics to the standard k-means algorithm.

This paper is organized as follows. In the next section, we provide a theoretical analysis of the behavior of the L_k norm in very high dimensionality. In section 3, we discuss fractional distance metrics and provide a theoretical analysis of their behavior. In section 4, we provide the empirical results, and section 5 provides summary and conclusions.

2 Behavior of the L_k-norm in High Dimensionality

In order to present our convergence results, we first establish some notations and definitions in Table 1.

Table 1. Notations and Basic Definitions

Notation	Definition
d	Dimensionality of the data space
N	Number of data points
$\mathcal{F}$	1-dimensional data distribution in $(0, 1)$
X_d	Data point from $\mathcal{F}^d$ with each coordinate drawn from $\mathcal{F}$
$dist_d^k(x, y)$	Distance between $(x^1, \ldots x^d)$ and $(y^1, \ldots y^d)$ using L_k metric $= \sum_{i=1}^d [(x_1^i - x_2^i)^k]^{1/k}$
$\|\cdot\|_k$	Distance of a vector to the origin $(0, \ldots, 0)$ using the function $dist_d^k(\cdot, \cdot)$
$Dmax_d^k = \max\{\|X_d\|_k\}$	Farthest distance of the N points to the origin using the distance metric L_k
$Dmin_d^k = \min\{\|X_d\|_k\}$	Nearest distance of the N points to the origin using the distance metric L_k
$E[X]$, $var[X]$	Expected value and variance of a random variable X
$Y_d \rightarrow_p c$	A vector sequence $Y_1, \ldots, Y_d$ converges in probability to a constant vector c if: $\forall \epsilon > 0 \; lim_{d\rightarrow\infty} P[dist_d(Y_d, c) \leq \epsilon] = 1$

Theorem 1. Beyer et. al. (Adapted for L_k metric)
If $lim_{d\rightarrow\infty} \; var\left(\frac{\|X_d\|_k}{E[\|X_d\|_k]}\right) = 0$, *then* $\frac{Dmax_d^k - Dmin_d^k}{Dmin_d^k} \rightarrow_p 0$.

Proof. See [6] for proof of a more general version of this result.

The result of the theorem [6] shows that the difference between the maximum and minimum distances to a given query point [1] does not increase as fast as the nearest distance to any point in high dimensional space. This makes a proximity query meaningless and unstable because there is poor discrimination between the nearest and furthest neighbor. Henceforth, we will refer to the ratio $\frac{Dmax_d^k - Dmin_d^k}{Dmin_d^k}$ as *the relative contrast.*

The results in [6] use the value of $\frac{Dmax_d^k - Dmin_d^k}{Dmin_d^k}$ as an interesting criterion for meaningfulness. In order to provide more insight, in the following we analyze the behavior for different distance metrics in high-dimensional space. We first assume a uniform distribution of data points and show our results for $N = 2$

[1] In this paper, we consistently use the origin as the query point. This choice does not affect the generality of our results, though it simplifies our algebra considerably.

points. Then, we generalize the results to an arbitrary number of points and arbitrary distributions.

Lemma 1. *Let $\mathcal{F}$ be uniform distribution of $N = 2$ points. For an L_k metric, $lim_{d\to\infty} E\left[\frac{Dmax_d^k - Dmin_d^k}{d^{1/k-1/2}}\right] = C \cdot \left(\frac{1}{(k+1)^{1/k}}\right)\sqrt{\left(\frac{1}{2\cdot k+1}\right)}$, where C is some constant.*

Proof. Let A_d and B_d be the two points in a d dimensional data distribution such that each coordinate is independently drawn from a 1-dimensional data distribution $\mathcal{F}$ with finite mean and standard deviation. Specifically $A_d = (P_1 \dots P_d)$ and $B_d = (Q_1 \dots Q_d)$ with P_i and Q_i being drawn from $\mathcal{F}$. Let $PA_d = \{\sum_{i=1}^{d}(P_i)^k\}^{1/k}$ be the distance of A_d to the origin using the L_k metric and $PB_d = \{\sum_{i=1}^{d}(Q_i)^k\}^{1/k}$ the distance of B_d. The difference of distances is $PA_d - PB_d = \{\sum_{i=1}^{d}(P_i)^k\}^{1/k} - \{\sum_{i=1}^{d}(Q_i)^k\}^{1/k}$.

It can be shown [2] that the random variable $(P_i)^k$ has mean $\frac{1}{k+1}$ and standard deviation $\left(\frac{k}{k+1}\right)\sqrt{\left(\frac{1}{2\cdot k+1}\right)}$. This means that $\frac{(PA_d)^k}{d} \to_p \frac{1}{(k+1)}$, $\frac{(PB_d)^k}{d} \to_p \frac{1}{(k+1)}$ and therefore

$$\frac{PA_d}{d^{1/k}} \to_p \left(\frac{1}{k+1}\right)^{1/k}, \quad \frac{PB_d}{d^{1/k}} \to_p \left(\frac{1}{k+1}\right)^{1/k} \tag{1}$$

We intend to show that $\frac{|PA_d - PB_d|}{d^{1/k-1/2}} \to_p \left(\frac{1}{(k+1)^{1/k}}\right)\sqrt{\left(\frac{2}{2\cdot k+1}\right)}$. We can express $|PA_d - PB_d|$ in the following numerator/denominator form which we will use in order to examine the convergence behavior of the numerator and denominator individually.

$$|PA_d - PB_d| = \frac{|(PA_d)^k - (PB_d)^k|}{\sum_{r=0}^{k-1}(PA_d)^{k-r-1}(PB_d)^r} \tag{2}$$

Dividing both sides by $d^{1/k-1/2}$ and regrouping the right-hand-side we get:

$$\frac{|PA_d - PB_d|}{d^{1/k-1/2}} = \frac{|((PA_d)^k - (PB_d)^k)|/\sqrt{d}}{\sum_{r=0}^{k-1}\left(\frac{PA_d}{d^{1/k}}\right)^{k-r-1}\left(\frac{PB_d}{d^{1/k}}\right)^r} \tag{3}$$

Consequently, using Slutsky's theorem [3] and the results of Equation 1 we obtain

$$\sum_{r=0}^{k-1}\left(\frac{PA_d}{d^{1/k}}\right)^{k-r-1} \cdot \left(\frac{PB_d}{d^{1/k}}\right)^r \to_p k \cdot \left(\frac{1}{k+1}\right)^{(k-1)/k} \tag{4}$$

Having characterized the convergence behavior of the denominator of the right hand side of Equation 3, let us now examine the behavior of the numerator: $|(PA_d)^k - (PB_d)^k|/\sqrt{d} = |\sum_{i=1}^{d}((P_i)^k - (Q_i)^k)|/\sqrt{d} = |\sum_{i=1}^{d} R_i|/\sqrt{d}$. Here

[2] This is because $E[P_i^k] = 1/(k+1)$ and $E[P_i^{2k}] = 1/(2\cdot k+1)$.

[3] **Slutsky's Theorem:** Let $Y_1 \dots Y_d \dots$ be a sequence of random vectors and $h(\cdot)$ be a continuous function. If $Y_d \to_p c$ then $h(Y_d) \to_p h(c)$.

R_i is the new random variable defined by $((P_i)^k - (Q_i)^k)$ $\forall i \in \{1, \dots d\}$. This random variable has zero mean and standard deviation which is $\sqrt{2} \cdot \sigma$ where σ is the standard deviation of $(P_i)^k$. The sum of different values of R_i over d dimensions will converge to a normal distribution with mean 0 and standard deviation $\sqrt{2} \cdot \sigma \cdot \sqrt{d}$ because of the central limit theorem. Consequently, the mean average deviation of this distribution will be $C \cdot \sigma$ for some constant C. Therefore, we have:

$$\lim_{d \to \infty} E\left[\frac{|(PA_d)^k - (PB_d)^k|}{\sqrt{d}}\right] = C \cdot \frac{k}{k+1}\sqrt{\frac{1}{2 \cdot k + 1}} \tag{5}$$

Since the denominator of Equation 3 shows probabilistic convergence, we can combine the results of Equations 4 and 5 to obtain

$$\lim_{d \to \infty} E\left[\frac{|PA_d - PB_d|}{d^{1/k-1/2}}\right] = C \cdot \frac{1}{(k+1)^{1/k}}\sqrt{\frac{1}{2 \cdot k + 1}} \tag{6}$$

We can easily generalize the result for a database of N uniformly distributed points. The following Corollary provides the result.

Corollary 1. *Let $\mathcal{F}$ be the uniform distribution of $N = n$ points. Then,*
$\left(\frac{C}{(k+1)^{1/k}}\right)\sqrt{\left(\frac{1}{2 \cdot k+1}\right)} \leq lim_{d \to \infty} E\left[\frac{Dmax_d^k - Dmin_d^k}{d^{1/k-1/2}}\right] \leq \left(\frac{C \cdot (n-1)}{(k+1)^{1/k}}\right)\sqrt{\left(\frac{1}{2 \cdot k+1}\right)}$.

Proof. This is because if L is the expected difference between the maximum and minimum of two randomly drawn points, then the same value for n points drawn from the same distribution must be in the range $(L, (n-1) \cdot L)$.

The results can be modified for arbitrary distributions of N points in a database by introducing the constant factor C_k. In that case, the general dependency of $D_{max} - D_{min}$ on $d^{\frac{1}{k} - \frac{1}{2}}$ remains unchanged. A detailed proof is provided in the Appendix; a short outline of the reasoning behind the result is available in [9].

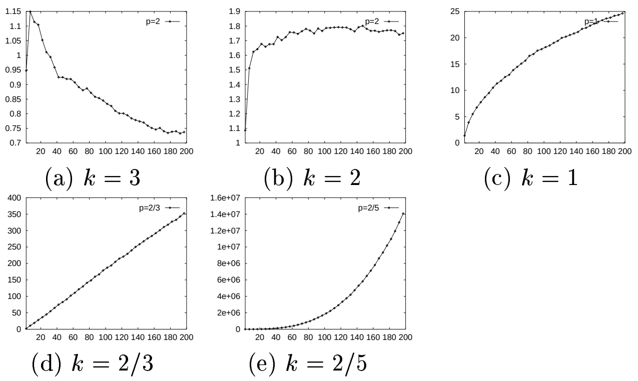
Lemma 2. *[9] Let $\mathcal{F}$ be an arbitrary distribution of $N = 2$ points. Then,*
$lim_{d \to \infty} E\left[\frac{Dmax_d^k - Dmin_d^k}{d^{1/k-1/2}}\right] = C_k$, *where C_k is some constant dependent on k.*

Corollary 2. *Let $\mathcal{F}$ be the arbitrary distribution of $N = n$ points. Then,*

$$C_k \leq lim_{d \to \infty} E\left[\frac{Dmax_d^k - Dmin_d^k}{d^{1/k-1/2}}\right] \leq (n-1) \cdot C_k$$

.

Thus, this result shows that in high dimensional space $Dmax_d^k - Dmin_d^k$ increases at the rate of $d^{1/k-1/2}$, independent of the data distribution. This means that for the manhattan distance metric, the value of this expression diverges to ∞; for the Euclidean distance metric, the expression is bounded by constants whereas for all other distance metrics, it converges to 0 (see Figure 1). Furthermore, the convergence is faster when the value of k of the L_k metric



(a) $k = 3$ (b) $k = 2$ (c) $k = 1$ (d) $k = 2/3$ (e) $k = 2/5$

Fig. 1. $|Dmax - Dmin|$ depending on d for different metrics (uniform data)

Table 2. Effect of dimensionality on relative (L_1 and L_2) behavior of relative contrast

Dimensionality	$P[U_d < T_d]$
1	Both metrics are the same
2	85.0%
3	88.7%
4	91.3%
10	95.6%
15	96.1%
20	97.1%
100	98.2%

increases. This provides the insight that higher norm parameters provide poorer contrast between the furthest and nearest neighbor. Even more insight may be obtained by examining the exact behavior of the relative contrast as opposed to the absolute distance between the furthest and nearest point.

Theorem 2. *Let $\mathcal{F}$ be the uniform distribution of $N = 2$ points. Then,* $lim_{d\to\infty} E\left[\left(\frac{Dmax_d^k - Dmin_d^k}{Dmin_d^k}\right) \cdot \sqrt{d}\right] = C \cdot \sqrt{\frac{1}{2\cdot k+1}}$.

Proof. Let A_d, B_d, $P_1 \ldots P_d$, $Q_1 \ldots Q_d$, PA_d, PB_d be defined as in the proof of Lemma 1. We have shown in the proof of the previous result that $\frac{PA_d}{d^{1/k}} \to \left(\frac{1}{k+1}\right)^{1/k}$. Using Slutsky's theorem we can derive that:

$$\min\{\frac{PA_d}{d^{1/k}}, \frac{PB_d}{d^{1/k}}\} \to \left(\frac{1}{k+1}\right)^{1/k} \tag{7}$$

We have also shown in the previous result that:

$$\lim_{d\to\infty} E\left[\frac{|PA_d - PB_d|}{d^{1/k-1/2}}\right] = C \cdot \left(\frac{1}{(k+1)^{1/k}}\right)\sqrt{\left(\frac{1}{2\cdot k+1}\right)} \tag{8}$$

We can combine the results in Equation 7 and 8 to obtain:

$$\lim_{d\to\infty} E\left[\sqrt{d} \cdot \frac{|PA_d - PB_d|}{\min\{PA_d, PB_d\}}\right] = C \cdot \sqrt{1/(2\cdot k+1)} \tag{9}$$

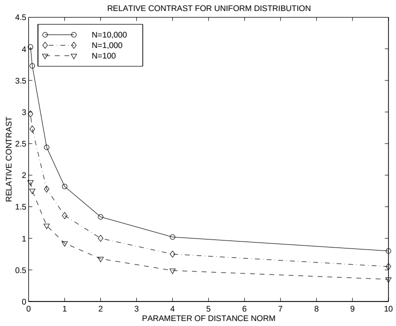


Fig. 2. Relative contrast variation with norm parameter for the uniform distribution

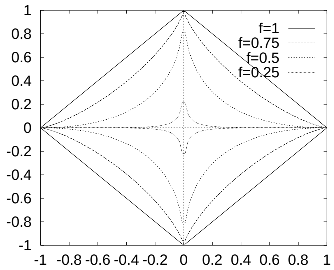


Fig. 3. Unit spheres for different fractional metrics (2D)

Note that the above results confirm of the results in [6] because it shows that the relative contrast degrades as $1/\sqrt{d}$ for the different distance norms. Note that for values of d in the reasonable range of data mining applications, the norm dependent factor of $\sqrt{1/(2 \cdot k + 1)}$ may play a valuable role in affecting the relative contrast. For such cases, even the relative rate of degradation of the different distance metrics for a given data set in the same value of the dimensionality may be important. In the Figure 2 we have illustrated the relative contrast created by an artificially generated data set drawn from a uniform distribution in $d = 20$ dimensions. Clearly, the relative contrast decreases with increasing value of k and also follows the same trend as $\sqrt{1/(2 \cdot k + 1)}$.

Another interesting aspect which can be explored to improve nearest neighbor and clustering algorithms in high-dimensional space is the effect of k on the relative contrast. Even though the expected relative contrast always decreases with increasing dimensionality, this may not necessarily be true for a given data set and different k. To show this, we performed the following experiment on the Manhattan (L_1) and Euclidean (L_2) distance metric: Let $U_d = \left(\frac{Dmax_d^2 - Dmin_d^2}{Dmin_d^2}\right)$ and $T_d = \left(\frac{Dmax_d^1 - Dmin_d^1}{Dmin_d^1}\right)$. We performed some empirical tests to calculate the value of $P[U_d < T_d]$ for the case of the Manhattan (L_1) and Euclidean (L_2) distance metrics for $N = 10$ points drawn from a uniform distribution. In each trial, U_d and T_d were calculated from the same set of $N = 10$ points, and $P[U_d < T_d]$ was calculated by finding the fraction of times U_d was less than T_d in 1000 trials. The results of the experiment are given in Table 2. It is clear that with increasing dimensionality d, the value of $P[U_d < T_d]$ continues to increase. *Thus, for higher dimensionality, the relative contrast provided by a norm with smaller parameter k is more likely to dominate another with a larger parameter.* For dimensionalities of 20 or higher it is clear that the manhattan distance metric provides a significantly higher relative contrast than the Euclidean distance metric with very high probability. Thus, among the distance metrics with integral norms, the manhattan distance metric is the method of choice for providing the best contrast between the different points. This result of our analysis can be directly used in a number of different applications.

3 Fractional Distance Metrics

The result of the previous section that the Manhattan metric ($k = 1$) provides the best discrimination in high-dimensional data spaces is the motivation for looking into distance metrics with $k < 1$. We call these metrics fractional distance metrics. A **fractional distance metric** $dist_d^f$ (L_f norm) for $f \in (0, 1)$ is defined as:

$$dist_d^f(x, y) = \sum_{i=1}^{d} \left[(x^i - y^i)^f\right]^{1/f}.$$

To give a intuition of the behavior of the fractional distance metric we plotted in Figure 3 the unit spheres for different fractional metrics in $\mathcal{R}^2$.

We will prove most of our results in this section assuming that f is of the form $1/l$, where l is some integer. The reason that we show the results for this special case is that we are able to use nice algebraic tricks for the proofs. The natural conjecture from the smooth continuous variation of $dist_d^f$ with f is that the results are also true for arbitrary values of f. [4]. Our results provide considerable insights into the behavior of the fractional distance metric and its relationship with the L_k-norm for integral values of k.

Lemma 3. *Let $\mathcal{F}$ be the uniform distribution of $N = 2$ points and $f = 1/l$ for some integer l. Then,*
$lim_{d\to\infty} E\left[\frac{Dmax_d^f - Dmin_d^f}{d^{1/f-1/2}}\right] = C \cdot \left(\frac{1}{(f+1)^{1/f}}\right)\sqrt{\left(\frac{1}{2\cdot f+1}\right)}.$

Proof. Let $A_d, B_d, P_1 \ldots P_d, Q_1 \ldots Q_d, PA_d, PB_d$ be defined using the L_f metric as they were defined in Lemma 1 for the L_k metric. Let further $QA_d = (PA_d)^f = (PA_d)^{1/l} = \sum_{i=1}^{d}(P_i)^f$ and $QB_d = (PB_d)^f = (PB_d)^{1/l} = \sum_{i=1}^{d}(Q_i)^f$. Analogous to Lemma 1, $\frac{QA_d}{d} \to_p \frac{1}{f+1}$, $\frac{QB_d}{d} \to_p \frac{1}{f+1}$.

We intend to show that $E\left[\frac{|PA_d - PB_d|}{d^{l-1/2}}\right] = C \cdot \left(\frac{1}{(f+1)^{1/f}}\right)\sqrt{\left(\frac{1}{2\cdot f+1}\right)}$. The difference of distances is $|PA_d - PB_d| = \{\sum_{i=1}^{d}(P_i)^f\}^{1/f} - \{\sum_{i=1}^{d}(Q_i)^f\}^{1/f}$ $= \{\sum_{i=1}^{d}(P_i)^f\}^l - \{\sum_{i=1}^{d}(Q_i)^f\}^l$. Note that the above expression is of the form $|a^l - b^l| = |a - b| \cdot (\sum_{r=0}^{l-1} a^r \cdot b^{l-r-1})$. Therefore, $|PA_d - PB_d|$ can be written as $\{\sum_{i=1}^{d} |(P_i)^f - (Q_i)^f|\} \cdot \{\sum_{r=0}^{l-1}(QA_d)^r \cdot (QB_d)^{l-r-1}\}$. By dividing both sides by $d^{1/f-1/2}$ and regrouping the right hand side we get:

$$\frac{|PA_d - PB_d|}{d^{1/f-1/2}} \to_p \left\{\frac{\sum_{i=1}^{d} |(P_i)^f - (Q_i)^f|}{\sqrt{d}}\right\} \cdot \left\{\sum_{r=0}^{l-1}\left(\frac{QA_d}{d}\right)^r \cdot \left(\frac{QB_d}{d}\right)^{l-r-1}\right\} \quad (10)$$

[4] Empirical simulations of the relative contrast show this is indeed the case.

By using the results in Equation 10, we can derive that:

$$\frac{|PA_d - PB_d|}{d^{1/f-1/2}} \rightarrow_p \{\frac{\sum_{i=1}^{d} |(P_i)^f - (Q_i)^f|}{\sqrt{d}}\} \cdot \{l \cdot \frac{1}{(1+f)^{l-1}}\} \tag{11}$$

This random variable $(P_i)^f - (Q_i)^f$ has zero mean and standard deviation which is $\sqrt{2} \cdot \sigma$ where σ is the standard deviation of $(P_i)^f$. The sum of different values of $(P_i)^f - (Q_i)^f$ over d dimensions will converge to normal distribution with mean 0 and standard deviation $2 \cdot \sigma \cdot \sqrt{d}$ because of the central limit theorem. Consequently, the expected mean average deviation of this normal distribution is $C \cdot \sigma \cdot \sqrt{d}$ for some constant C. Therefore, we have:

$$\lim_{d\rightarrow\infty} E\left[\frac{|(PA_d)^f - (PB_d)^f|}{\sqrt{d}}\right] = C \cdot \sigma = C \cdot \left(\frac{f}{f+1}\right)\sqrt{\left(\frac{1}{2 \cdot f + 1}\right)}. \tag{12}$$

Combining the results of Equations 12 and 11, we get:

$$\lim_{d\rightarrow\infty} E\left[\frac{|PA_d - PB_d|}{d^{1/f-1/2}}\right] = \left(\frac{C}{(f+1)^{1/f}}\right)\sqrt{\left(\frac{1}{2 \cdot f + 1}\right)} \tag{13}$$

An direct consequence of the above result is the following generalization to $N = n$ points.

Corollary 3. *When $\mathcal{F}$ is the uniform distribution of $N = n$ points and $f = 1/l$ for some integer l. Then, for some constant C we have:*
$\left(\frac{C}{(f+1)^{1/f}}\right)\sqrt{\left(\frac{1}{2 \cdot f+1}\right)} \leq lim_{d\rightarrow\infty} E\left[\frac{Dmax_d^f - Dmin_d^f}{d^{1/f-1/2}}\right] \leq \left(\frac{C \cdot (n-1)}{(f+1)^{1/f}}\right)\sqrt{\left(\frac{1}{2 \cdot f+1}\right)}.$

Proof. Similar to corollary 1.

The above result shows that the absolute difference between the maximum and minimum for the fractional distance metric increases at the rate of $d^{1/f-1/2}$. Thus, the smaller the fraction, the greater the rate of absolute divergence between the maximum and minimum value. Now, we will examine the relative contrast of the fractional distance metric.

Theorem 3. *Let $\mathcal{F}$ be the uniform distribution of $N = 2$ points and $f = 1/l$ for some integer l. Then,*
$lim_{d\rightarrow\infty} \left(\frac{Dmax_d^f - Dmin_d^f}{Dmin_d^f}\right)\sqrt{d} = C \cdot \sqrt{\frac{1}{2 \cdot f+1}}$ *for some constant C.*

Proof. Analogous to the proof of Theorem 2.

The following is the direct generalization to $N = n$ points.

Corollary 4. *Let $\mathcal{F}$ be the uniform distribution of $N = n$ points, and $f = 1/l$ for some integer l. Then, for some constant C*
$C \cdot \sqrt{\frac{1}{2 \cdot f+1}} \leq lim_{d\rightarrow\infty} E\left[\frac{Dmax_d^f - Dmin_d^f}{Dmin_d^f}\right] \leq C \cdot (n-1) \cdot \sqrt{\frac{1}{2 \cdot f+1}}.$

Proof. Analogous to the proof of Corollary 1.

This result is true for the case of arbitrary values f (not just $f = 1/l$) and N, but the use of these specific values of f helps considerably in simplification of the proof of the result. The empirical simulation in Figure 2, shows the behavior for arbitrary values of f and N. The curve for each value of N is different but all curves fit the general trend of reduced contrast with increased value of f. Note that the value of the relative contrast for both, the case of integral distance metric L_k and fractional distance metric L_f is the same in the boundary case when $f = k = 1$.

The above results show that fractional distance metrics provide better contrast than integral distance metrics both in terms of the absolute distributions of points to a given query point and relative distances. This is a surprising result in light of the fact that the Euclidean distance metric is traditionally used in a large variety of indexing structures and data mining applications. The widespread use of the Euclidean distance metric stems from the natural extension of applicability to spatial database systems (many multidimensional indexing structures were initially proposed in the context of spatial systems). However, from the perspective of high dimensional data mining applications, this natural interpretability in 2 or 3-dimensional spatial systems is completely irrelevant. Whether the theoretical behavior of the relative contrast also translates into practically useful implications for high dimensional data mining applications is an issue which we will examine in greater detail in the next section.

4 Empirical Results

In this section, we show that our surprising findings can be directly applied to improve existing mining techniques for high-dimensional data. For the experiments, we use synthetic and real data. The synthetic data consists of a number of clusters (data inside the clusters follow a normal distribution and the cluster centers are uniformly distributed). The advantage of the synthetic data sets is that the clusters are clearly separated and any clustering algorithm should be able to identify them correctly. For our experiments we used one of the most widely used standard clustering algorithms - the *k-means algorithm*. The data set used in the experiments consists of 6 clusters with 10000 data points each and no noise. The dimensionality was chosen to be 20. The results of our experiments show that the fractional distance metrics provides a much higher classification rate which is about 99% for the fractional distance metric with $f = 0.3$ versus 89% for the Euclidean metric (see figure 4). The detailed results including the confusion matrices obtained are provided in the appendix.

For the experiments with real data sets, we use some of the classification problems from the UCI machine learning repository [5]. All of these problems are classification problems which have a large number of feature variables, and a special variable which is designated as the class label. We used the following

[5] *http://www.cs.uci.edu/~mlearn*

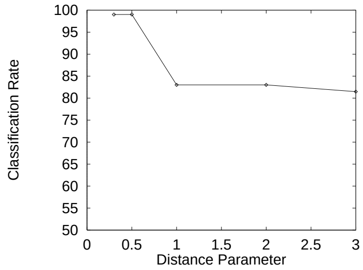


Fig. 4. Effectiveness of k-Means

simple experiment: For each of the cases that we tested on, we *stripped off* the class variable from the data set and considered the feature variables only. The query points were picked from the original database, and the closest l neighbors were found to each target point using different distance metrics. The technique was tested using the following two measures:

1. Class Variable Accuracy: This was the primary measure that we used in order to test the quality of the different distance metrics. Since the class variable is known to depend in some way on the feature variables, the proximity of objects belonging to the same class in feature space is evidence of the meaningfulness of a given distance metric. The specific measure that we used was the total number of the l nearest neighbors that belonged to the same class as the target object over all the different target objects. Needless to say, we do not intend to propose this rudimentary unsupervised technique as an alternative to classification models, but use the classification performance only as an evidence of the meaningfulness (or lack of meaningfulness) of a given distance metric. The class labels may not necessarily always correspond to locality in feature space; therefore the meaningfulness results presented are evidential in nature. However, a consistent effect on the class variable accuracy with increasing norm parameter does tend to be a powerful way of demonstrating qualitative trends.

2. Noise Stability: How does the quality of the distance metric vary with more or less noisy data? We used *noise masking* in order to evaluate this aspect. In noise masking, each entry in the database was replaced by a random entry with masking probability p_c. The random entry was chosen from a uniform distribution centered at the mean of that attribute. Thus, when p_c is 1, the data is completely noisy. We studied how each of the two problems were affected by noise masking.

In Table 3, we have illustrated some examples of the variation in performance for different distance metrics. Except for a few exceptions, the major trend in this table is that the accuracy performance decreases with increasing value of the norm parameter. We have show the table in the range $L_{0.1}$ to L_{10} because it was easiest to calculate the distance values without exceeding the numerical ranges in the computer representation. We have also illustrated the accuracy performance when the L_∞ metric is used. One interesting observation is that the accuracy

Table 3. Number of correct class label matches between nearest neighbor and target

Data Set	$L_{0.1}$	$L_{0.5}$	L_1	L_2	L_4	L_{10}	L_∞	Random
Machine	522	474	449	402	364	353	341	153
Musk	998	893	683	405	301	272	163	140
Breast Cancer (wdbc)	5299	5268	5196	5052	4661	4172	4032	3021
Segmentation	1423	1471	1377	1210	1103	1031	300	323
Ionosphere	2954	3002	2839	2430	2062	1836	1769	1884

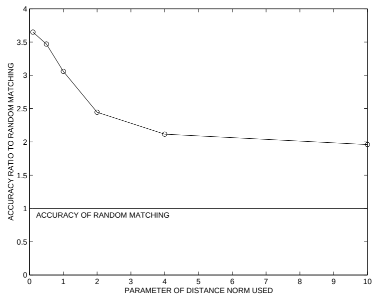


Fig. 5. Accuracy depending on the norm parameter

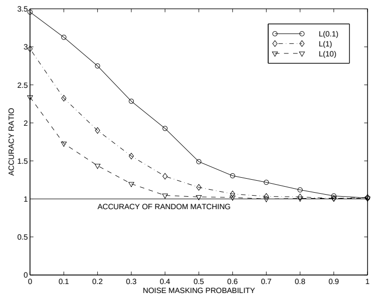


Fig. 6. Accuracy depending on noise masking

with the L_∞ distance metric is often worse than the accuracy value by picking a record from the database at random and reporting the corresponding target value. This trend is observed because of the fact that the L_∞ metric only looks at the dimension at which the target and neighbor are furthest apart. In high dimensional space, this is likely to be a very poor representation of the nearest neighbor. A similar argument is true for L_k distance metrics (for high values of k) which provide undue importance to the distant (sparse/noisy) dimensions. It is precisely this aspect which is reflected in our theoretical analysis of the relative contrast, which results in distance metrics with high norm parameters to be poorly discriminating between the furthest and nearest neighbor.

In Figure 5, we have shown the variation in the accuracy of the class variable matching with k, when the L_k norm is used. The accuracy on the Y-axis is reported as the ratio of the accuracy to that of a completely random matching scheme. The graph is averaged over all the data sets of Table 3. It is easy to see that there is a clear trend of the accuracy worsening with increasing values of the parameter k.

We also studied the robustness of the scheme to the use of noise masking. For this purpose, we have illustrated the performance of three distance metrics in Figure 6: $L_{0.1}$, L_1, and L_{10} for various values of the masking probability on the machine data set. On the X-axis, we have denoted the value of the masking probability, whereas on the Y-axis we have the accuracy ratio to that of a completely random matching scheme. Note that when the masking probability is 1, then any scheme would degrade to a random method. However, it is interesting to see from Figure 6 that the L_{10} distance metric degrades much faster to the random performance (at a masking probability of 0.4), whereas the L_1 degrades to random at 0.6. The $L_{0.1}$ distance metric is most robust to the presence of noise in the data set and degrades to random performance at the slowest rate. These results are closely connected to our theoretical analysis which shows the rapid lack of discrimination between the nearest and furthest distances for high values of the norm-parameter because of undue weighting being given to the noisy dimensions which contribute the most to the distance.

5 Conclusions and Summary

In this paper, we showed some surprising results of the qualitative behavior of the different distance metrics for measuring proximity in high dimensionality. We demonstrated our results in both a theoretical and empirical setting. In the past, not much attention has been paid to the choice of distance metrics used in high dimensional applications. The results of this paper are likely to have a powerful impact on the particular choice of distance metric which is used from problems such as clustering, categorization, and similarity search; all of which depend upon some notion of proximity.

References

1. Weber R., Schek H.-J., Blott S.: A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. *VLDB Conference Proceedings*, 1998.
2. Bennett K. P., Fayyad U., Geiger D.: Density-Based Indexing for Approximate Nearest Neighbor Queries. *ACM SIGKDD Conference Proceedings*, 1999.
3. Berchtold S., Böhm C., Kriegel H.-P.: The Pyramid Technique: Towards Breaking the Curse of Dimensionality. *ACM SIGMOD Conference Proceedings*, June 1998.
4. Berchtold S., Böhm C., Keim D., Kriegel H.-P.: A Cost Model for Nearest Neighbor Search in High Dimensional Space. *ACM PODS Conference Proceedings*, 1997.
5. Berchtold S., Ertl B., Keim D., Kriegel H.-P. Seidl T.: Fast Nearest Neighbor Search in High Dimensional Spaces. *ICDE Conference Proceedings*, 1998.
6. Beyer K., Goldstein J., Ramakrishnan R., Shaft U.: When is Nearest Neighbors Meaningful? *ICDT Conference Proceedings*, 1999.
7. Shaft U., Goldstein J., Beyer K.: Nearest Neighbor Query Performance for Unstable Distributions. Technical Report TR 1388, Department of Computer Science, University of Wisconsin at Madison.
8. Guttman, A.: R-Trees: A Dynamic Index Structure for Spatial Searching. *ACM SIGMOD Conference Proceedings*, 1984.
9. Hinneburg A., Aggarwal C., Keim D.: What is the nearest neighbor in high dimensional spaces? *VLDB Conference Proceedings*, 2000.
10. Katayama N., Satoh S.: The SR-Tree: An Index Structure for High Dimensional Nearest Neighbor Queries. *ACM SIGMOD Conference Proceedings*, 1997.
11. Lin K.-I., Jagadish H. V., Faloutsos C.: The TV-tree: An Index Structure for High Dimensional Data. *VLDB Journal*, Volume 3, Number 4, pages 517–542, 1992.

Appendix

Here we provide a detailed proof of Lemma 2, which proves our modified convergence results for arbitrary distributions of points. This Lemma shows that the asymptotical rate of convergence of the absolute difference of distances between the nearest and furthest points is dependent on the distance norm used. To recap, we restate Lemma 2.

Lemma 2: *Let $\mathcal{F}$ be an arbitrary distribution of $N = 2$ points. Then, $lim_{d\to\infty} E\left[\frac{Dmax_d^k - Dmin_d^k}{d^{1/k-1/2}}\right] = C_k$, where C_k is some constant dependent on k.*

Proof. Let A_d and B_d be the two points in a d dimensional data distribution such that each coordinate is independently drawn from the data distribution $\mathcal{F}$. Specifically $A_d = (P_1 \dots P_d)$ and $B_d = (Q_1 \dots Q_d)$ with P_i and Q_i being drawn from $\mathcal{F}$. Let $PA_d = \{\sum_{i=1}^{d}(P_i)^k\}^{1/k}$ be the distance of A_d to the origin using the L_k metric and $PB_d = \{\sum_{i=1}^{d}(Q_i)^k\}^{1/k}$ the distance of B_d.

We assume that the kth power of a random variable drawn from the distribution $\mathcal{F}$ has mean $\mu_{\mathcal{F},k}$ and standard deviation $\sigma_{\mathcal{F},k}$. This means that: $\frac{PA_d^k}{d} \to_p \mu_{\mathcal{F},k}$, $\frac{PB_d^k}{d} \to_p \mu_{\mathcal{F},k}$ and therefore:

$$PA_d/d^{1/k} \to_p (\mu_{\mathcal{F},k})^{1/k}, \quad PB_d/d^{1/k} \to_p (\mu_{\mathcal{F},k})^{1/k}. \tag{14}$$

We intend to show that $\frac{|PA_d - PB_d|}{d^{1/k-1/2}} \to_p C_k$ for some constant C_k depending on k. We express $|PA_d - PB_d|$ in the following numerator/denominator form which we will use in order to examine the convergence behavior of the numerator and denominator individually.

$$|PA_d - PB_d| = \frac{|(PA_d)^k - (PB_d)^k|}{\sum_{r=0}^{k-1}(PA_d)^{k-r-1}(PB_d)^r} \tag{15}$$

Dividing both sides by $d^{1/k-1/2}$ and regrouping on right-hand-side we get

$$\frac{|PA_d - PB_d|}{d^{1/k-1/2}} = \frac{|(PA_d)^k - (PB_d)^k|/\sqrt{d}}{\sum_{r=0}^{k-1}\left(\frac{PA_d}{d^{1/k}}\right)^{k-r-1}\left(\frac{PB_d}{d^{1/k}}\right)^r} \tag{16}$$

Consequently, using Slutsky's theorem and the results of Equation 14 we have:

$$\sum_{r=0}^{k-1}\left(PA_d/d^{1/k}\right)^{k-r-1} \cdot \left(PB_d/d^{1/k}\right)^r \to_p k \cdot (\mu_{\mathcal{F},k})^{(k-1)/k} \tag{17}$$

Having characterized the convergence behavior of the denominator of the right-hand-side of Equation 16, let us now examine the behavior of the numerator: $|(PA_d)^k - (PB_d)^k|/\sqrt{d} = |\sum_{i=1}^{d}((P_i)^k - (Q_i)^k)|/\sqrt{d} = |\sum_{i=1}^{d} R_i|/\sqrt{d}$.
Here R_i is the new random variable defined by $((P_i)^k - (Q_i)^k)\ \forall i \in \{1, \dots d\}$. This random variable has zero mean and standard deviation which is $\sqrt{2} \cdot \sigma_{\mathcal{F},k}$ where $\sigma_{\mathcal{F},k}$ is the standard deviation of $(P_i)^k$. Then, the sum of different values

of R_i over d dimensions will converge to a normal distribution with mean 0 and standard deviation $\sqrt{2} \cdot \sigma_{\mathcal{F},k} \cdot \sqrt{d}$ because of the central limit theorem. Consequently, the mean average deviation of this distribution will be $C \cdot \sigma_{\mathcal{F},k}$ for some constant C. Therefore, we have:

$$\lim_{d\to\infty} E\left[\frac{|(PA_d)^k - (PB_d)^k|}{\sqrt{d}}\right] = C \cdot \sigma_{\mathcal{F},k} \tag{18}$$

Since the denominator of Equation 16 shows probabilistic convergence, we can combine the results of Equations 17 and 18 to obtain:

$$\lim_{d\to\infty} E\left[\frac{|PA_d - PB_d|}{d^{1/k-1/2}}\right] = C \cdot \frac{\sigma_{\mathcal{F},k}}{k \cdot \mu_{\mathcal{F},k}^{(k-1)/k}} \tag{19}$$

The result follows.

Confusion Matrices We have illustrated the confusion matrices for two different values of p below. As illustrated, the confusion matrix for using the value $p = 0.3$ is significantly better than the one obtained using $p = 2$.

Table 4. Confusion Matrix- p=2, (rows for prototype, colums for cluster)

1208	82	9711	4	10	14
0	2	0	0	6328	4
1	9872	104	32	11	0
8750	8	74	9954	1	18
39	0	10	8	8	9948
2	36	101	2	3642	16

Table 5. Confusion Matrix- p=0.3, (rows for prototype, colums for cluster)

51	115	9773	10	37	15
0	17	24	0	9935	14
15	10	9	9962	0	4
1	9858	66	5	19	1
8	0	9	3	9	9956
9925	0	119	20	0	10