Knowl Inf Syst
DOI 10.1007/s10115-014-0813-4

REGULAR PAPER

Multiplicative distance: a method to alleviate distance instability for high-dimensional data

Jafar Mansouri · Morteza Khademi

Received: 27 January 2014 / Revised: 18 October 2014 / Accepted: 06 December 2014

Abstract Recently, it has been shown that under a broad set of conditions, the commonly used distance functions will become unstable in high-dimensional data space; i.e., the distance to the farthest data point approaches the distance to the nearest data point of a given query point with increasing dimensionality. It has been shown that if dimensions are independently distributed, and normalized to have zero mean and unit variance, instability happens. In this paper, it is shown that the normalization condition is not necessary, but all appropriate moments must be finite. Furthermore, a new distance function, namely multiplicative distance, is introduced. It is theoretically proved that this function is stable for data with independent dimensions (with identical or nonidentical distribution). In contrast to usual distance functions which are based on the summation of distances over all dimensions (distance components), the multiplicative distance is based on the multiplication of distance components. Experimental results show the stability of the multiplicative distance for data with independent and correlated dimensions in the high-dimensional space and the superiority of the multiplicative distance over the norm distances for the high-dimensional data.

Keywords Distance instability · High-dimensional data · Minkowski and fractional norms · Multiplicative and additive distances

1 Introduction

Nowadays, due to the great advances in computers, high-capacity storage disks, multimedia transmission, and compression systems, there are many applications which require high-dimensional data analysis. One important issue in many of these applications is determining distances between data and query points. Beyer et al. [4] have proved that under a broad

J. Mansouri (✉) · M. Khademi
Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
e-mail: jafar.mansouri@gmail.com

M. Khademi
e-mail: khademi@um.ac.ir

Published online: 28 December 2014

set of conditions, distances between data and query points are meaningless or unstable in the high-dimensional space. This means that the ratio of the distances of the nearest and farthest neighbors to a given query point approaches 1 for a wide variety of data distributions and distance functions when dimensionality increases toward infinity, e.g., for data with independently and identically distributed (i.i.d.) dimensions with Minkowski distance. Of course, the instability of Euclidean distance under some conditions has already been stated by Demartines [10]. This instability is also called distance concentration phenomenon [26]. The commonly used distance functions become unstable in the high-dimensional space. Weber et al. [39] have reported that when the dimensionality of data is more than 10, this phenomenon can occur. In such cases, the concepts of proximity and similarity are not meaningful because of poor discrimination between the nearest and furthest neighbors. This instability can greatly affect many applications like information retrieval and data indexing [5,15,20,40], gene analysis [6,27], high-quality image and video analysis [35,37,45,46], clustering [3,31,33, 41], and classification [13,16,17,36], and can result in the poor performance of the systems.

Three solutions have been more addressed for conquering distance instability in high-dimensional data space in the literature. The first solution is space partitioning [8,21,25,28, 47]. In these methods, data are divided to some partitions and the nearest neighbor or distance is obtained over these partitions. These methods have some drawbacks with increasing the dimensionality like significant increase in the number of partitions and complexity, and the performance degradation. Moreover, some of these methods are just applicable for the nearest neighbor search and cannot give the exact value of distance.

The second solution, which is widely used, is to apply some dimension reduction techniques [9,23,32,38,44] to the high-dimensional data space. These methods have this weakness that the reduced-dimension data may be high-dimensional again, for example, in the processing of high-quality images and videos. For example, if singular value decomposition (SVD) is applied on an image in CIF format (352×288), the size of the right and left singular matrices is 352×288. If rows of one of these matrices are concatenated to form a vector, dimensionality of this vector is 101,376. Suppose that this vector is used as a feature vector (e.g., for clustering), and a dimension reduction technique is applied on it. Since the original dimensionality is very high, the reduced-dimension vector is high-dimensional again (e.g., in order of 10^3) and distance functions are unstable for this vector. If we set conditions on the dimension reduction technique to obtain a reduced-dimension vector with low dimensionality, definitely a large amount of important information is lost and this can result in the poor performance of the system. So, dimension reduction techniques are not useful for data with a very high dimensionality in practice.

Since the above two methods have drawbacks and limitations for some applications, other solution should be considered. The third solution is to define a meaningful distance function. In [11] and [19], authors have stated that the sufficient and necessary condition for stability of a distance function is that the ratio of variance of the distance distribution to the square of its expectation, namely Pearson variation, does not tend to zero as the dimensionality grows to infinity (Beyer et al. in [4] proved the sufficient condition of instability). With this criterion, the stability of a distance function can be examined. As will be discussed in Sect. 2, the few presented distance functions for conquering distance instability have some drawbacks. Also, there is no theoretical proof of the stability for those functions. In this paper, the third solution is addressed and a new distance function is proposed.

This paper has the following contributions: In [14], Francois et al. have proved that norm distances are unstable for data with independent dimensions provided that dimensions are normalized, i.e., zero mean and unit variance. This paper proves that normalization condition is not necessary, but dimensions must have finite appropriate moments. Also, we introduce the

Table 1 List of symbols

Notation	Definition
m	Dimensionality of the data space
n	Number of data points
p	A constant
X	Random vectors defined in some probability space
$x \sim F$	A random variable x that takes on values from the distribution F
d_m	A distance function of an m-dimensional data space
$P\{e\}$	Probability of an event e
$E[x]$	Expectation of a random variable x
var$[x]$	Variance of a random variable x
c	Control power
i.i.d.	Independently and identically distributed
DMIN_m	Nearest distance of the points to the query point
DMAX_m	Farthest distance of the points to the query point

multiplicative distance as a new distance function which is theoretically proved for data with independent dimensions (with identical or nonidentical distribution), its Pearson variation does not tend to zero when dimensionality increases to infinity. Simulation results show the stability of this distance function for correlated data, too. As an application, it is shown that this distance function has better performance than the norm distances for clustering. Also, the multiplicative distance can be used in the low-dimensional space in applications like clustering. A list of symbols used in this paper is given in Table 1.

This paper is organized as follows. Section 2 gives an overview on the related works. In Sect. 3, a theorem on the instability of the norm distance functions for data with independent dimensions with identical or nonidentical distribution is presented. Section 4 introduces the multiplicative distance and details its properties and stability for the high-dimensional data. The experimental results are reported in Sect. 5. Finally, some conclusions are drawn in Sect. 6.

2 Related works

In this section, we survey the previous works that studied instability of different distance functions and defined new distance functions. In [10], Demartines has shown that if $X = (x_1, x_2, \ldots, x_m)$ is an m-dimensional random vector with i.i.d. dimensions with any distribution, Euclidean norm is unstable. Beyer et al. in a seminal paper [4] have stated that under certain conditions on the distance distribution (induced by data distribution and distance function), as dimensionality increases, the distance to the nearest data point approaches the distance to the farthest data point of a given query point. In other words, the ratio $\frac{\text{DMAX}_m - \text{DMIN}_m}{\text{DMIN}_m}$, named relative contrast, converges to 0. One important case given in Beyer et al. [4] is for data and query points with i.i.d. dimensions for Minkowski norm.

Aggarwal and Yu [2] have proposed a grid-based approach to determine the similarity function based on Minkowski norms. In this method, each dimension is divided into k equi-depth ranges. The jth range for dimension i is denoted by $R[i, j]$. Let $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_m)$ be two records and m be number of dimensions. For dimension i, if both

x_i and y_i belong to the same range of $R[i, j]$, then the two records are said to be in proximity on dimension i. The entire set of dimensions in which the two records lie in the same range is referred to as the proximity set. Let $S[X, Y, k]$ be the proximity set for two records X and Y for a given level of discretization. Furthermore, for each dimension $i \in S[X, Y, k]$, let u_i and l_i be the upper and lower bounds for the corresponding range in the dimension i in which the records X and Y are in proximity to each another. Then, for a given pair of records X and Y, and a level of discretization k, the similarity between the records is given by:

$$\text{PIDist}(X, Y, k) = \left[\sum_{i \in S[X,Y,k]} \left(1 - \frac{|x_i - y_i|}{u_i - l_i} \right)^p \right]^{1/p} \tag{1}$$

The presence of $u_i - l_i$ in the denominator normalizes the distances with respect to the different ranges of the coordinates. The limitation of the above function is that this function leads to ignoring the exact value of dissimilarity on the dimensions in which two records are not in the same range. Moreover, since the algorithm is based on equi-depth ranges, if some new data are added to or removed from dataset, ranges can change significantly. In other words, u_i and l_i are very sensitive to the number of data. Also, the stability has not been theoretically proved for this function.

In [18], Hinneburg et al. have obtained the difference of maximum and minimum of the L_p-norm for i.i.d. random vectors and for an arbitrary distribution when dimensionality grows to infinity. Aggarwal et al. [1] introduced fractional norms, an extension of Minkowski norm with a positive exponent less than one, to alleviate the instability effect. They detailed the Hinneburg's work for the relative contrast for uniform distribution and then extended the results to the fractional norm. They also concluded that fractional norms are usually more suitable than Minkowski norms in terms of the relative contrast for the uniform distribution. It should be noted that fractional norms are not metric since the triangle inequality does not hold in general.

In [14], Francois et al. have shown that the instability of the norms in high-dimensional spaces is really an intrinsic instability property of the norms and not a side effect of the finite sample size. In addition, all fractional distances become unstable in high-dimensional spaces. They also have proved that the root of Pearson variation is a strictly decreasing function of p ($0 < p < \infty$) when the dimensions are distributed in accordance with an i.i.d. uniform distribution over the interval [0, 1]. However, there exist some distributions for which fractional norms are not always less unstable than higher-order norms. Moreover, Francois et al. have proved that if dimensions are independent (but not necessarily identically distributed), Minkowski and fractional norms are unstable provided that the dimensions are normalized, i.e., zero mean and unit variance. They have experimentally shown that usually high-dimensional data that present a lot of correlation or dependencies between dimensions will be much less unstable than that if all dimensions are independent. Furthermore, they have shown with some examples that it cannot be said which one of the Minkowski or fractional norms is more stable, except in special cases.

As stated before, in [11] and [19], authors have proved the necessary and sufficient condition for the stability of distances in the high-dimensional data space. In [19], Hsu and Chen have introduced shrinkage-divergence proximity (SDP) function for alleviating distance instability. For any m-dimensional data points $X = (x_1, \ldots, x_m)$ and $Y = (y_1, \ldots, y_m)$, the general form of SDP is defined as:

$$\text{SDP}(X, Y) = \sum_{i=1}^{m} w_i f_{s_{i1}, s_{i2}}(d_1(x_i, y_i)) \tag{2}$$

where

$$f_{a,b}(x) = \begin{cases} 0 & \text{if } 0 \le x < a \\ x & \text{if } a \le x < b \\ e^x & \text{otherwise} \end{cases} \tag{3}$$

where d_1 is a distance function for one-dimensional data. The weighting parameter w_i, which reflects particular characteristics of the dataset, is determined by the domain knowledge or statistical characteristics among dimensions, subject to the importance of dimension i to the application's needs. The parameters s_{i1} and s_{i2} are called the shrinkage threshold and divergence threshold of dimension i, respectively. These parameters are obtained empirically in terms of standard deviation of the dimension i. The SDP function is not metric since the triangle inequality does not hold in general. In addition, it can be verified that $\mathrm{SDP}_{s1,s2}(X, Y) = 0$ does not imply $X = Y$. One limitation of this distance function is that SDP is sensitive to its chosen parameters, and they are obtained empirically. The other limitation of the above distance function is that it cannot give the exact value of distance on the dimensions with $d_1(x_i, y_i) < s_{i1}$. Also, the stability of this function has not been theoretically proved.

In [11], Durrant and Kaban have shown that for a class of data distributions having non-i.i.d. dimensions, namely the family of linear latent variable models, the Euclidean distance will not be unstable as long as the amount of relevant dimensions grows no slower than the overall data dimensions. An irrelevant dimension in linear latent variable model will only contain noise. In linear latent variable models, each dimension is defined based on a linear combination of the latent factors, and there is a special kind of correlation between dimensions. The limitation is that this condition is not often met in practice. It should be mentioned that the existence of correlation between dimensions does not guarantee the stability of norm distances; e.g., Beyer et al. have given an example in [4] that there is a correlation between dimensions, but the norm distance is unstable.

In [26], Ledoux has stated that the Hamming distance will be concentrated around a mean value on the high-dimensional Boolean cube equipped with counting probability measure. Also, the instability of cosine similarity has been explored by Radovanovi'c et al. in [34] for i.i.d. data. In [24], Kaban has investigated the instability phenomenon for high but finite-dimensional data in a distribution-free manner and in this case has obtained a lower bound on the probability that distances become unstable.

3 The instability of norms

Many additive distance functions become unstable in high-dimensional data space for a wide range of data and query distributions. Francois et al. have proved one case of instability in the below theorem.

Theorem 1 (Francois et al. [14]) *If dimensions of the data points are independent (but not necessarily identically distributed) and normalized, then fractional and Minkowski norms are unstable when dimensionality tends to infinity.*

Normalization here means that for each random variable x_k, subtract the variable from its mean and divide it by its standard deviation so that $E[x_k] = 0$ and $\mathrm{var}[x_k] = 1$, $k = 1, 2, \ldots, m$. In this section, it is proved that the condition of normalization is not necessary if dimensions have finite appropriate moments and number of dimensions with nonzero variances tends to infinity as dimensionality tends to infinity. First, Theorem 2 is stated which is used in the proof of Theorem 3:

Theorem 2 ([11] and [19]) *Sufficient and necessary condition of instability: Let p be a constant ($0 < p < \infty$), and $X_i \sim F$, $i = 1, 2, \ldots, n$ be a set of random vectors chosen independently from the query point of $Q \sim F_q$, and d_m be a distance function. The distance $d_m(X_i, Q)$ is unstable if and only if its Pearson variation tends to zero when dimensionality tends to infinity, i.e.,*

$$\lim_{m\to\infty} P\left[\frac{DMAX_m - DMIN_m}{DMIN_m} \leq \varepsilon\right] = 1 \text{ if and only if } \lim_{m\to\infty} \operatorname{var}\left(\frac{(d_m(X_i, Q))^p}{E[(d_m(X_i, Q))^p]}\right) = 0. \tag{4}$$

$\operatorname{var}\left(\frac{(d_m(X_i,Q))}{E[(d_m(X_i,Q))]}\right)$ *is defined as Pearson variation. According to this theorem, a distance function is unstable in high-dimensional space if its Pearson variation of distance distribution approaches 0 with increasing dimensionality.*

Theorem 3 *In the high-dimensional space if dimensions are independent (identically or not identically distributed), fractional and Minkowski norms are unstable when dimensionality grows to infinity provided that all the appropriate moments are finite (i.e., up to the $\lceil 2p \rceil$'th moment, $0 < p < \infty$) and number of dimensions with nonzero variance ($\sigma^2 = \operatorname{var}[|x|^p]$) tends to infinity.*

Proof Let $0 < p < \infty$ and $X_i = (x_{i,1}, x_{i,2} \ldots, x_{i,m})$, $i = 1, \ldots, n$ be a set of random vectors with m independent dimensions. Each dimension has a specific distribution, i.e., $\forall i : x_{i,k} \sim F_k \quad k = 1, 2, \ldots, m$. As in Francois et al. [14], we use the origin as the query point. This choice does not affect the generality of our results, though it simplifies our algebra considerably. Thus, we have:

$$d_m(X_i, \text{origin}) = \left(\sum_{k=1}^{m} |x_{i,k}|^p\right)^{1/p} \tag{5}$$

In accordance with Theorem 2, for checking the instability, the Pearson variation of the distance distribution must be calculated. As in Beyer et al. [4], we can write:

$$\frac{\operatorname{var}[d_m^p]}{(E[d_m^p])^2} = \frac{\sum_{k=1}^{m} \sigma_k^2}{\left(\sum_{k=1}^{m} \mu_k\right)^2} = \frac{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_m^2}{(\mu_1 + \mu_2 + \cdots + \mu_m)^2} \tag{6}$$

where $\mu_k = E[|x_{i,k}|^p]$ and $\sigma_k^2 = \operatorname{var}[|x_{i,k}|^p]$ (note that μ_k and σ_k^2 are calculated for the "*absolute values*"). Assume that number of dimensions with $\sigma_k^2 \neq 0$ would be $\tilde{m}$. For these dimensions, set $\mu_{\min} = \min\{\mu_1, \mu_2, \ldots, \mu_{\tilde{m}}\}$ and $\sigma_{\max}^2 = \max\{\sigma_1^2, \sigma_2^2, \ldots, \sigma_{\tilde{m}}^2\}$. $\mu_{\min}$ cannot be zero, since if one of the $\mu_1, \mu_2, \ldots, \mu_{\tilde{m}}$ is zero, its respective variance must be zero, and this is in contradiction to our assumption. So, we have:

$$\frac{\operatorname{var}[d_m^p]}{(E[d_m^p])^2} \leq \frac{\tilde{m}\sigma_{\max}^2}{(\tilde{m}\mu_{\min})^2} = \frac{\sigma_{\max}^2}{\tilde{m}\mu_{\min}^2} \tag{7}$$

if $m \to \infty$, then, according to the assumption, $\tilde{m} \to \infty$. Therefore, $\frac{\operatorname{var}[d_m^p]}{(E[d_m^p])^2} \to 0$ as $m \to \infty$. It is observed that this proof is free of the normalization condition. □

Note that in fact, Theorem 1 is a special case of Theorem 3. In other words, this theorem is true for any kind of normalization, too. Notice that the condition of zero mean in Theorem 1 is prior to the calculating distance; e.g., a random variable can have a distribution $N(0,1)$, but

for obtaining Pearson variation, μ_k is calculated for the absolute value of the normalized data. With Theorem 3, the set of conditions under which the additive distances, like Minkowski and fractional norms, become unstable is extended in comparison with Theorem 1.

4 Multiplicative distance

From Theorem 3 and the related works in Sect. 2, it can be concluded that many distance functions, like Minkowski and fractional norms, are unstable in the high-dimensional space for many data distributions. As stated before, the similarity and dissimilarity functions defined in Aggarwal and Yu [2] and Hsu and Chen [19] have their own limitations. Therefore, defining a new distance function that can resist to instability or reduce it is necessary. In this section, we propose a new distance function, named multiplicative distance, which can resist to instability in high-dimensional space for a wide range of data distributions. Compared to the additive distance, which comprises of the addition of elements, the multiplicative distance contains the product of elements. In this section, the definition of the multiplicative distance, its similarities and differences with additive distances, its stabilities and some other characteristics of it are considered.

4.1 Multiplicative distance versus additive distances

Definition Let $X = (x_1, x_2 \ldots, x_m)$ be a random vector with $x_k \sim F_k$, $k = 1, \ldots, m$ and $Q = (q_1, q_2, \ldots, q_m)$ be a query point with $q_k \sim \tilde{F}_k$. Set $z_k = 1 + |x_k - q_k|$. The general form of the multiplicative distance of X from Q is defined as:

$$\mathrm{MD}(X, Q) = \left(\prod_{k=1}^{m} z_k^{c_k} \right) - 1 \tag{8}$$

where $c_k > 0$ is named "control power," which controls the effect of each z_k on the distance. $z_k^{c_k}$ is defined as the distance component. If $\forall k : c_k = c$, each dimension has the equal effect on the distance. In the simple and usual form of the multiplicative distance, we have $\forall k : c_k = 1$. For example, if $X = (5, -1, 0.5, -7.5, 6)$, $Y = (3, 2, 0, -9, 6)$, and $c_k = 1 \quad k = 1 : 5$, then $\mathrm{MD}(X, Y) = (1+|5-3|)(1+|-1-2|)(1+|0.5-0|)(1+|-7.5-(-9)|)(1+|6-6|)-1 = 44$.

Now, multiplicative distances can be compared with additive distances. In additive distances like Minkowski norms $\left(\sum_{i=1}^{m} |t_i|^p\right)^{1/p}$, the distance comprises of the addition of distance components. Here, distance components are t_k^p where $t_k = |x_k - q_k|$. But the multiplicative distance comprises of the product of distance components. The reason of adding "1" to $|x_k - q_k|$ is that a distance function, like a norm distance function, must be non-descending; i.e., when a new dimension is added, the distance must be increased or remained unchanged. In the similar way, in the multiplicative distance when a new dimension is added, its distance component must not decrease the distance. If we just set $|x_k - q_k|$ as the distance component, in some cases with adding a new dimension, the distance can be decreased. To avoid this situation, each $|x_k - q_k|$ must be larger than 1, and for this reason, 1 is added to $|x_k - q_k|$. In the summation, the neutral element is 0, while in the multiplication, the neutral element is 1. These facts are seen in the norm and multiplicative distance formulas. Since in the multiplicative distance, distance components are multiplied, the value of the multiplicative distance is much larger than the value of the additive distance in which distance

components are added. In other words, the multiplication intensifies the value of the distance when distance components are larger than 1.

There is weighted Minkowski or fractional norm $\left(\sum_{i=1}^{m} w_i |t_i|^p\right)^{1/p}$, in which each w_k controls the effect of the corresponding distance component on the distance. Similarly, in the multiplicative distance "control power, c" performs this task but with a difference. In Minkowski or fractional norms, if $\forall k$: $w_k = w$, then, different values of w do not affect the Pearson variation, while in the multiplicative distance (as will be described later) if $\forall k$: $c_k = c$, different values of c directly affects the Pearson variation.

It can be thought that by taking logarithm or exponent of the multiplicative distance, it is similar to L_1-norm or SDP function without threshold, respectively. However, these are not true since logarithm and exponent are nonlinear operators, and they can change the expected value and variance. In other words, $E[\log(x)] \neq \log(E[x])$ and $\text{var}[\log(x)] \neq \log(\text{var}[x])$. So, we "cannot" apply logarithm or exponent to the multiplicative distance. Furthermore, the previous distances are based on summation of components, but the multiplicative distance is based on the product, and the nature of the product is different from the nature of the summation. Moreover, the multiplicative distance function gives the exact value of distance in contrast to the functions defined in Aggarwal and Yu [2] and Hsu and Chen [9].

4.2 Multiplicative distance and its stability in high-dimensional space

Now, in the following theorem, it is proved that for data with independent dimensions (but not necessarily identically distributed), the multiplicative distance is stable in the high-dimensional data space, in contrast to additive distances which are unstable (Theorem 3).

Theorem 4 *If the set of random vectors $X_i = (x_{i,1}, \ldots, x_{i,m})$, $i = 1, \ldots, n$ are data points with independent but not necessarily identically distributed entries $(x_{i,k} \sim F_k)$, $k = 1, \ldots, m$, and $Q = (q_1, q_2, \ldots, q_m)$ with $q_k \sim \tilde{F}_k$ is the query point chosen independently from all X_i, then the multiplicative distance of X_i from Q is stable when dimensionality tends to infinity.*

Proof For stability, the Pearson variation of the multiplicative distance must be examined as the dimensionality rises to infinity. Under the independent assumption of $x_{i,k}$, $z_{i,k} = 1 + |x_{i,k} - q_k|$ are independent variables.

$$\frac{\text{var}[\text{MD}(X_i, Q)]}{(E[\text{MD}(X_i, Q)])^2} = \frac{\text{var}\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]}{\left(E\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right] - 1\right)^2} \geq \frac{\text{var}\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]}{\left(E\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]\right)^2} \tag{9}$$

since $E\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]$ is a positive value larger than 1. If $a \geq b$ and b does not tend to zero when dimensionality tends to infinity, definitely, a does not tends to zero. Therefore, it is sufficient to show that $\frac{\text{var}\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]}{\left(E\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]\right)^2}$ does not tend to zero.

$$\frac{\text{var}\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]}{\left(E\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]\right)^2} = \frac{E\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)^2\right] - \left(E\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]\right)^2}{\left(E\left[\left(\prod_{k=1}^{m} z_{i,k}^{c_k}\right)\right]\right)^2} \tag{10}$$

If random variables $z_{i,k}$ are independent, then the random variables $y_{i,k} = g_k(z_{i,k})$ are also independent, where $g_k(.)$ is a deterministic function. Thus, the Pearson variation can be written:

$$\frac{E\left[\left(z_{i,1}^{c_1}\right)^2\right] \times E\left[\left(z_{i,2}^{c_2}\right)^2\right] \times \cdots \times E\left[\left(z_{i,m}^{c_m}\right)^2\right] - \left(E\left[z_{i,1}^{c_1}\right] \times E\left[z_{i,2}^{c_2}\right] \times \cdots \times E\left[z_{i,m}^{c_m}\right]\right)^2}{\left(E\left[z_{i,1}^{c_1}\right] \times E\left[z_{i,2}^{c_2}\right] \times \cdots \times E\left[z_{i,m}^{c_m}\right]\right)^2} \tag{11}$$

Set $\sigma_k^2 = \text{var}[z_{i,k}^{c_k}]$ and $\mu_k = E[z_{i,k}^{c_k}]$; hence $E[(z_{i,k}^{c_k})^2] = \sigma_k^2 + \mu_k^2$. It is clear that $\mu_k \geq 1$. Therefore, the Pearson variation can be written:

$$\begin{aligned} &= \frac{\left(\sigma_1^2 + \mu_1^2\right)\left(\sigma_2^2 + \mu_2^2\right) \cdots \left(\sigma_m^2 + \mu_m^2\right) - \mu_1^2 \mu_2^2 \cdots \mu_m^2}{\mu_1^2 \mu_2^2 \cdots \mu_m^2} \\ &= \left(\frac{\sigma_1^2}{\mu_1^2} + 1\right)\left(\frac{\sigma_2^2}{\mu_2^2} + 1\right) \cdots \left(\frac{\sigma_m^2}{\mu_m^2} + 1\right) - 1 \end{aligned} \tag{12}$$

It is shown that when $m \to \infty$, if all σ_i^2 are not very close to zero, the Pearson variation does not tend to zero, so the multiplicative distance is stable. □

Corollary 1 *If the set of random vectors $X_i = (x_{i,1}, \ldots, x_{i,m})$ $i = 1, \ldots, n$ are data points with i.i.d. entries from some distribution $x_{i,k} \sim F$, $k = 1, \ldots, m$, and $Q = (q_1, q_2, \ldots, q_m)$ with $q_k \sim \tilde{F}$ is the query point chosen independently from all X_i, then the multiplicative distance of any random vector to the query point is stable when dimensionality grows to infinity.*

4.3 More on the multiplicative distance

Figure 1 depicts the L_1 norm, L_2 norm, $L_{0.79}$ norm, and the multiplicative distance with different values of the control power for $X = (x_1, x_2)$ when $d_2(X, \text{origin}) = 1$. For better representing the curves, they are shown in two sub-figures. The sub-figures are depicted just for the first region of Cartesian coordination. The other three regions can be depicted by symmetry. The reason for choosing the $L_{0.79}$ norm is that its curve is similar to the

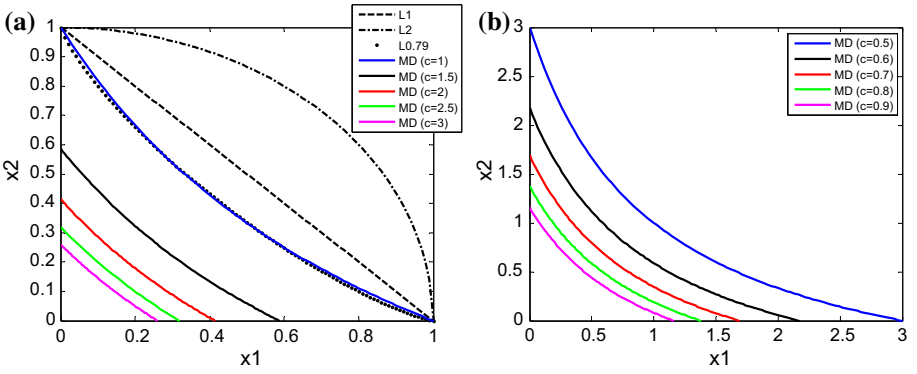

Fig. 1 Unit spheres (just for the first region of Cartesian coordination) for some distances and the multiplicative distances

multiplicative distance for $c = 1$ (simple form). In these sub-figures, the multiplicative distance is denoted by MD. As shown in Fig. 1, with decreasing c, the range of x_1 and x_2 is increased.

As an important point, since the multiplicative distance is the product of distance components, when the number of dimensions is increased, there can be a case in which for calculating the distance, overflow occurs. To remove this problem, the small c can be used. The overflow depends on the programming language, number of dimensions, and distribution of data. So, usually, the value of c cannot be determined in advance. Moreover, the multiplicative distance is not metric since the triangle inequality does not hold in general. However, the influence of the triangle inequality may be insignificant in many applications such as clustering, especially those for high-dimensional space [12,22]. For a metric distance, number of distance calculations can be reduced in the fast nearest neighbor search using the triangle inequality. But since the triangle inequality does not hold for the multiplicative distance, all distances from data points to the query point should be calculated for the multiplicative distance. In the following, we consider the effect of c on the Pearson variation.

Theorem 5 *If* $MD(X, Q) = \left(\prod_{k=1}^{m} z_k^c\right) - 1$, *i.e.,* $\forall k : c_k = c$, *and* $z_k = 1 + |x_k - q_k|$, *then, the Pearson variation of the multiplicative distance is an increasing function of* c.

Proof For simplicity, we set $\mathrm{MD}(X, Q) = \left(\prod_{k=1}^{m} z_k^c\right) - 1 = Z^c - 1 = Z^{c_1}$. So, the Pearson variation of the multiplicative distance equals to $\frac{E\left[Z^{2c_1}\right] - E^2\left[Z^{c_1}\right]}{E^2\left[Z^{c_1}\right]}$. Therefore, we consider the following equation:

$$f(c_1) = \frac{E\left[Z^{2c_1}\right]}{E^2\left[Z^{c_1}\right]} = \frac{E\left[e^{2c_1 Y}\right]}{E^2\left[e^{c_1 Y}\right]} = \frac{m_Y(2c_1)}{m_Y^2(c_1)} \tag{13}$$

where $Y = \log(Z)$ and $m_Y(c_1) = E\left[e^{c_1 Y}\right]$ is the moment generating function of Y. A well-known fact is that a moment generating function is log-convex on the interval where it exists. Thus, $r(c_1) = \log m_Y(c_1)$ defines a convex function of c_1. Now, $\log f(c_1) = r(2c_1) - 2r(c_1)$, and

$$\frac{\mathrm{d}}{\mathrm{d}c_1} \log f(c_1) = 2r'(2c_1) - 2r'(c_1) \geq 0 \tag{14}$$

since the derivative $r'(c_1)$ is increasing. Thus, $\log f(c_1)$ is increasing function of c_1, and so is the Pearson variation of Z^{c_1}. Thus, the Pearson variation of the multiplicative distance is an increasing function of c. □

To show the effect of c on the Pearson variation, the following experiment is performed. The three synthetic 300-dimensional sample datasets were generated as follows: For the first dataset, the jth entry of ith data point $X_i = (x_{i,1}, \ldots, x_{i,m})$ was randomly sampled from a uniform distribution U(0,1). Similarly, for the second and third datasets, entries were randomly sampled from a normal distribution N(0,1) and an exponential distribution Exp(0.5), respectively. Number of data points for each dataset is 1,000. Figure 2 shows the effect of c on the Pearson variation. As it can be observed, the Pearson variation is increased with increasing c.

Since Pearson variation of the multiplicative distance is an increasing function of c, there is no global or local optimal value for the Pearson variation with respect to c. The larger c is, the larger Pearson variation. But large c can make overflow. Also, when there is no superiority between dimensions, the "c" is selected the same for all dimensions. So, usually, $c = 1$ is selected for simplicity.

Major properties of the multiplicative distance function are listed as follows:

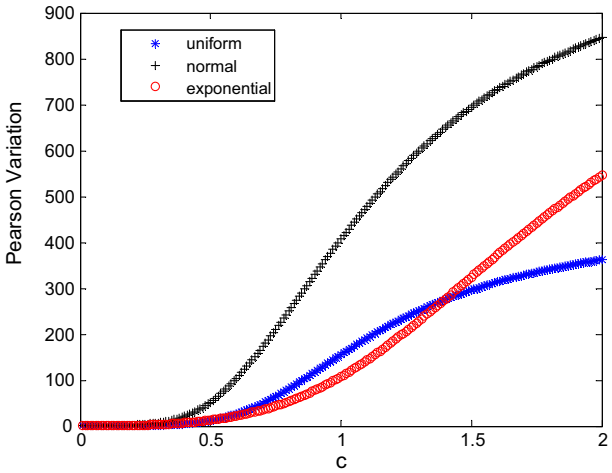

Fig. 2 Effect of c on the Pearson variation of the multiplicative distance for different data distributions

- Stability of the multiplicative distance is proved theoretically for data with independent dimensions and experimentally for data with correlated dimensions.
- Pearson variation of the multiplicative distance is an increasing function of c.
- Multiplicative distance is not a metric. A nonnegative function $d : \chi \times \chi \rightarrow R$ (R is the set of real numbers) is a metric for data space χ if $\forall X, Y, Z \in \chi$ it satisfies the following properties:

  (1) $d(X, Y) \geq 0$
  (2) $d(X, Y) = 0$ if and only if $X = Y$
  (3) $d(X, Y) = d(Y, X)$
  (4) $d(X, Y) + d(Y, Z) \geq d(X, Z)$

  Multiplicative distance function satisfies the first three properties but does not satisfies the fourth property (triangle property) in general.
- Multiplicative distance is not homogeneous, i.e., $\text{MD}(\alpha X, \alpha Q) \neq \alpha \text{MD}(X, Q)$.

5 Experimental evaluations

To provide a practical perspective and check the stability/instability of the multiplicative and norm distances for data with independent and correlated dimensions, some simulations are conducted on both synthetic (independent) and real (correlated) datasets. Section 5.1 illustrates some characteristics of datasets and experimental setup. In Sect. 5.2, the Pearson variation and the relative contrast of datasets are considered for the multiplicative distance and some norm distance functions. A statistical test is also used for assessing the stability of distances. In Sect. 5.3, these distance functions are applied to the k-means clustering algorithm to have a comparison between these functions for a real application. The effect of noise on distances is also considered on the clustering application.

5.1 Experimental setup

To compare the performance and stability of the multiplicative distance with some well-known distances, various datasets with different properties in diverse domains were selected.

Table 2 Characteristics of datasets in the experiments

Dataset name	Number of features	Number of instances	Number of classes
Iris	4	150	3
Liver	10	579	2
Satellite	36	6,435	6
Sonar	60	208	2
Usps	256	4,000	10
Madelon	500	2,000	2
Gisette	5,000	1,000	2
Leu	7,129	72	2
Farm	54,877	200	2

The datasets were chosen from UCI machine learning repository[1] and LIBSVM dataset.[2] These domains (in order of their appearance in Table 2) are iris plant prediction, liver patient prediction, multispectral satellite image detection, sonar signal detection, handwritten text recognition, synthetic dataset, confusable digit handwritten detection, gene monitoring for diseases, and farm animal advertisement detection. Except Madelon (synthetic dataset), other datasets are real. Datasets have different number of classes and instances. Table 2 summarizes these datasets and some of their characteristics. Specifically, some of datasets have few features to compare the multiplicative distance with other distances even for the low-dimensional data.

Features of data can be measured in different units. Therefore, normalization should be performed before calculating distances. Generally speaking, there are two kinds of normalizations. In the first kind, features are subtracted from their mean (μ_i) and divided by their standard deviation (σ_i) so that the normalized variables have zero mean and unit variance; i.e.,

$$x_{i,\text{new}} = \frac{x_i - \mu_i}{\sigma_i} \tag{15}$$

In the second kind of normalization, each variable ranges from zero to one; i.e.,

$$x_{i,\text{new}} = \frac{x_i - \min\{x_i\}}{\max\{x_i\} - \min\{x_i\}} \tag{16}$$

Both of these normalizations are considered in this paper for evaluating the effect of normalization on the results.

5.2 Stability checking

To asses the stability of the multiplicative distance and compare it with norm distances, first, the Pearson variation and relative contrast criteria are addressed [4]. The query point is the origin. The Pearson variation and the respective relative contrast for datasets with two types of normalizations for L_1, L_2, $L_{0.79}$, and the multiplicative distance functions have been shown in Tables 3, 4, 5, and 6. Generally, in the simulation the simple form of the multiplicative distance ($c = 1$) is used. But for some datasets, selecting $c = 1$ leads to the overflow. Hence,

[1] http://archive.ics.uci.edu/ml/.

[2] http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/.

Table 3 Pearson variation for type 1 of normalization

Dataset name	L_1	L_2	$L_{0.79}$	Multiplicative distance
Iris	0.1770	0.1515	0.1870	0.6575
Liver	0.2290	0.3520	0.2060	160.1374
Satellite	0.2360	0.1825	0.2551	185.1407
Sonar	0.0633	0.0839	0.0605	90.3465
Usps	0.0591	0.0940	0.0529	3.4719e+3
Madelon	0.0013	0.0012	0.0014	991.5387
Gisette	0.0184	0.0445	0.0140	957.0122
Leu	0.0610	0.0512	0.0632	71.9977
Farm	0.3331	0.2569	0.2193	200.0000

Table 4 Pearson variation for type 2 of normalization

Dataset name	L_1	L_2	$L_{0.79}$	Multiplicative distance
Iris	0.1876	0.1249	0.2082	0.4474
Liver	0.0752	0.0582	0.0877	0.4363
Satellite	0.0692	0.0626	0.0727	8.9320
Sonar	0.0313	0.0171	0.0364	91.8973
Usps	0.1371	0.0617	0.1845	2.3401e+3
Madelon	0.1698e−3	0.1517e−3	0.1782e−3	40.8736
Gisette	0.0658	0.0179	0.1052	999.9363
Leu	0.0145	0.0114	0.0158	3.3364
Farm	0.9978	0.2278	2.0194	196.8006

Table 5 Relative contrast for type 1 of normalization

Dataset name	L_1	L_2	$L_{0.79}$	Multiplicative distance
Iris	11.6327	9.7512	12.5964	80.8808
Liver	15.3273	21.0733	14.4831	4.4082e+4
Satellite	12.0620	8.5138	13.3936	5.8276e+16
Sonar	1.9920	2.5198	2.0169	4.6922e+11
Usps	2.2544	2.8587	2.2268	9.9172e+56
Madelon	0.2726	0.2785	0.2799	1.8221e+21
Gisette	1.3655	2.5197	1.1384	2.5391e+38
Leu	2.2523	1.9365	2.3286	8.2188e+15
Farm	12.0238	21.1615	8.2057	7.0172e+14

for these datasets and for simplicity, the largest value of the negative integer powers of 10, i.e., 0.1, 0.01, . . . , that overflow does not happen, is selected as the control power. For Iris, Liver, Satellite, Sonar, Usps, and Madelon, $c = 1$, for Gissete, $c = 0.1$, and for Leu and Farm, $c = 0.01$, were selected based on the experiments.

Tables 3, 4, 5, and 6 show the Pearson variation and the relative contrast for the norm and multiplicative distances for two kinds of normalizations. It is shown that Madelon, which is an artificial dataset with independent dimensions, has very small Pearson variation and relative

Table 6 Relative contrast for type 2 of normalization

Dataset name	L_1	L_2	$L_{0.79}$	Multiplicative distance
Iris	10.2334	9.5670	10.3815	31.8078
Liver	6.2574	4.3088	7.4704	51.6420
Satellite	3.5220	3.3442	3.7137	4.2137e+6
Sonar	1.9964	1.2414	2.2035	1.1991e+7
Usps	10.0762	4.3341	14.6986	1.8968e+48
Madelon	0.0989	0.0916	0.1013	6.9734e+6
Gisette	3.1028	1.1234	4.8050	1.3430e+28
Leu	0.9105	0.5016	1.1259	5.3867e+5
Farm	0.2484e+3	0.0148e+3	1.0807e+3	3.6264e+6

contrast for both types of normalizations for norm distances. This shows the instability of norm distances for independent data. Other datasets with high dimensionality usually have small Pearson variation and relative contrast for norm distance functions. Since other datasets are real, there is correlation between their dimensions. This correlation causes that datasets with higher dimensionality, and their Pearson variations and relative contrasts are larger than the case in which dimensions are independent (Madelon dataset). Consequently, norm distances are less unstable for other (real) datasets.

Furthermore, it is observed that Pearson variation and relative contrast for Madelon is large for both types of normalizations for the multiplicative distance. These results show the stability of the multiplicative distance for independent data in the high-dimensional data space. For other datasets with high dimensionality, which there exists correlation between their dimensions, the multiplicative distance has large Pearson variation and relative contrast for both types of normalizations. These results demonstrate the stability of the multiplicative distance in the high-dimensional data space for the correlation case. Of course, as it is observed in the proof of Theorem 4, if there are data that all σ_i^2 are very close to zero, the multiplicative distance becomes unstable.

All datasets have large Pearson variation and relative contrast for the multiplicative distance in comparison with the norm distances. In contrast to the norm distances which are usually unstable for high-dimensional data, the multiplicative distance is stable. The multiplicative distance is also stable for low-dimensional data (Iris dataset) as the norm distance is stable.

The type 2 of normalization does not always have smaller Pearson variation and relative rather than the type 1 of normalization, although the range of values in the type 2 is usually smaller than the type 1. Regarding Pearson variation, there is no superiority for two kinds of normalizations over each other for distance functions. From the relative contrast point of view, this matter is true for the norm distances. However, for the multiplicative distance, the first type of normalization gives larger relative contrasts. Therefore, the first type of normalization is better than the second type. Nevertheless, both of these normalizations have a large Pearson variation and relative contrast for the multiplicative distance and can be used in high-dimensional space.

The reason that the Pearson variation is large for the multiplicative distance is that, as it is seen in the proof of Theorem 4 (e.g., for data with independent dimensions), the Pearson variation is approximately equal to the product of values larger than one. For this reason, when dimensionality is increased, the Pearson variation is increased. Of course, the Pearson variation is also dependent on c and large c makes large Pearson variation. For example, if

for Farm dataset for $c = 1$ overflow did not occur, its Pearson variation would be larger than in the case of $c = 0.01$.

Another way for assessing distance instability/stability is the statistical test proposed in Kaban [24] that gives a finite-dimensional characterization of the distance instability phenomenon based on the following theorem.

Theorem 6 [24] *Let X_i, $i = 1, \ldots, n$, be independently drawn m-dimensional sample points from some distribution F, and Q be the query point. Denote $DMIN_m(n) = \min_{1 \le i \le n} d_m(X_i, Q)$ and $DMAX_m(n) = \max_{1 \le i \le n} d_m(X_i, Q)$. Then,*

$$P\{\mathrm{DMAX}_m(n) < (1+\varepsilon)\mathrm{DMIN}_m(n)\} \ge \left\{\left(1 - \left(\frac{2}{(1+\varepsilon)^p - 1} + 1\right)^2 \mathrm{PV}_m(p)\right)_+\right\}^n \tag{17}$$

where $(u)_+ = \max(0, u)$ and

$$\mathrm{PV}_m(p) = \mathrm{var}\left[\frac{(d_m(X_i, Q))^p}{E\left[(d_m(X_i, Q))^p\right]}\right] \tag{18}$$

It follows that this bound (17) is tight in the neighborhood of probability 1 when the probability of distance instability $P\{\mathrm{DMAX}_m(n) < (1+\varepsilon)\mathrm{DMIN}_m(n)\}$ is high. In other words, it means that for a specific value of ε, what is the lower bound on the probability that instability happens? This makes this bound appropriate as a statistical test to identify whether a given distance function suffers from the instability problem in some unknown data distribution. This probability bound requires that the true Pearson variation of the distance distribution induced by F is known. However, F is most often unknown in practice. Instead, we have data samples drawn from F. For this reason, the following theorem is used in which the true Pearson variation is replaced with its sample estimate.

Theorem 7 [24] *Let $X_1, \ldots, X_n \sim F$ be random samples, $n \ge 2$, and Q be the query point, $DMIN_m(n) = \min_{1 \le i \le n} d_m(X_i, Q)$, and $DMAX_m(n) = \max_{1 \le i \le n} d_m(X_i, Q)$, and let $Y_1, \ldots, Y_r \sim F$ be observed data samples, $r \ge 2$. Assume $P\{d_m(X_1, Q) = \ldots = d_m(X_n, Q) = d_m(Y_1, Q) = \ldots = d_m(Y_r, Q)\} = 0$. Then,*

$$\begin{aligned} &P\{\mathrm{DMAX}_m(n) < (1+\varepsilon)\mathrm{DMIN}_m(n)\} \\ &\ge \left\{\left(1 - \left(\frac{2}{(1+\varepsilon)^p - 1} + 1\right)^2 \overline{\mathrm{PV}}_{m,r}(p) \times \frac{r^2 - 1}{r^2} - \frac{1}{r}\right)_+\right\}^n \end{aligned} \tag{19}$$

where $\overline{\mathrm{PV}}_{m,r}(p)$ is the estimated Pearson variation from the dataset $Y_1, \ldots, Y_r$.

$$\begin{aligned} \overline{\mathrm{PV}}_{m,r}(p) &= \frac{\overline{var[(d_m(Y, Q))^p]}}{\left(\overline{E[(d_m(Y, Q))^p]}\right)^2} \\ \overline{E[d_m(Y, Q)^p]} &= \frac{1}{r}\sum_{i=1}^{r} (d_m(Y_i, Q))^p \\ \overline{var[(d_m(Y, Q))^p]} &= \frac{1}{r-1}\sum_{i=1}^{r} ((d_m(Y_i, Q))^p - \overline{E[(d_m(Y, Q)^p)]})^2 \end{aligned} \tag{20}$$

Figure 3 shows the lower bounds on the probability that norm and multiplicative distances become unstable for datasets in the underlying unknown data distributions, for $n = 2$, $p = 1$ [p in the formula (17)], plotted against a range of deviation ε. In Fig 3, N1 and N2 represent first and second types of normalizations, respectively. The reference query point is the origin. Again, it is observed that for norm distances, even for small values of ε, the probability of instability is usually high. This shows that norm distances are unstable for the dataset with independent dimensions and usually for the datasets with correlated dimensions. However, the lower bounds on the probability of instability remain zero for the considerable range of ε for the multiplicative distance. This indicates the stability of the multiplicative distance for all datasets (with independent and correlated dimensions) and for two kinds of normalizations.

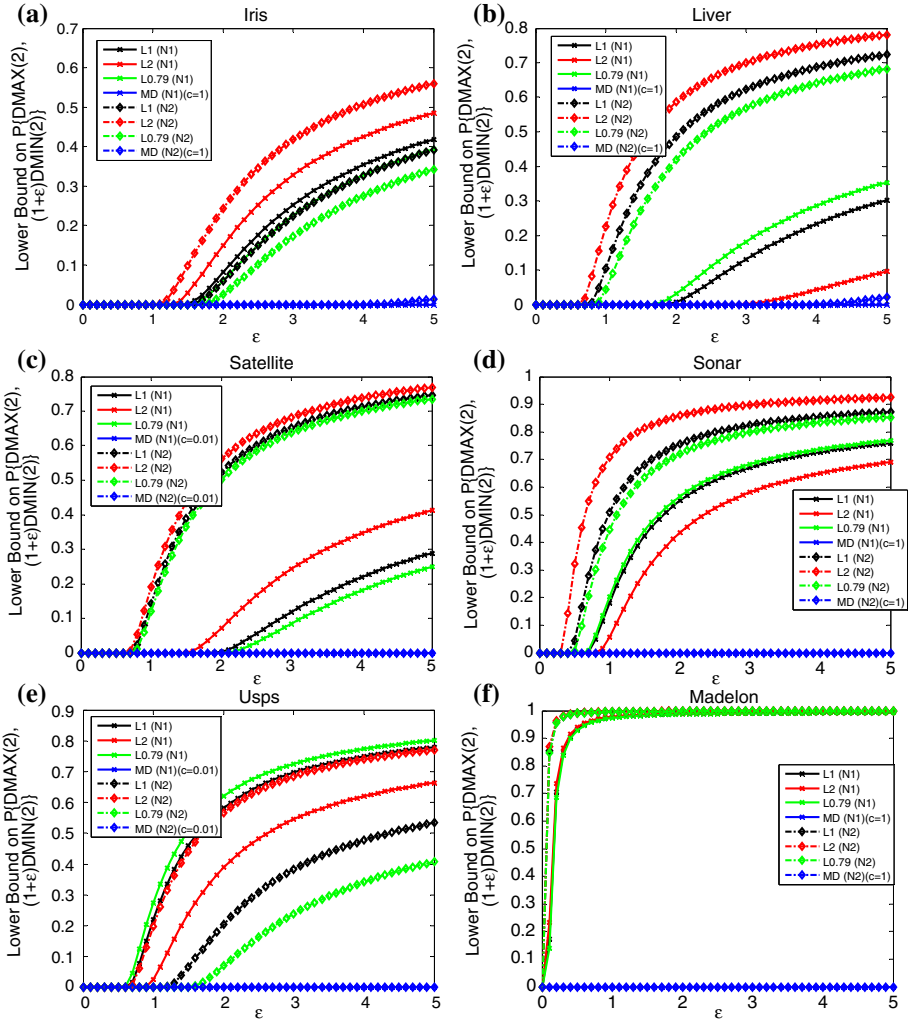

Fig. 3 Lower bounds on the probability that norm and multiplicative distances become unstable for two kinds of normalizations for different datasets: **a** Iris, **b** Live, **c** Satellite, **d** Sonar, **e** Usps, **f** Madelon, **g** Gisette, **h** Leu, **i** Farm. It is observed that multiplicative distance is stable for all datasets and all kinds of normalizations

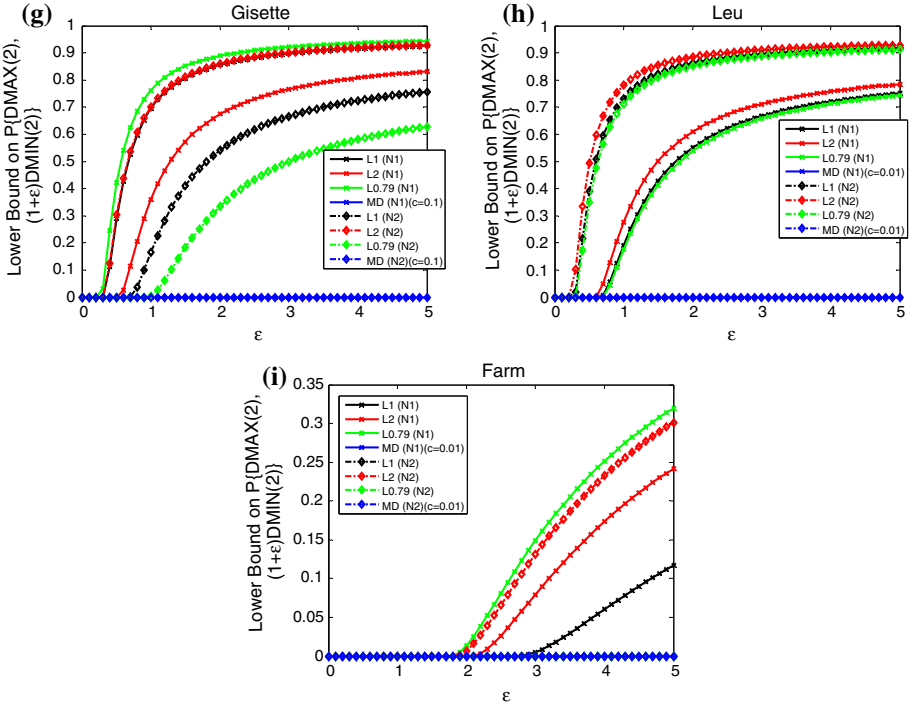

Fig. 3 continued

Notice that, usually, the curve for the multiplicative distance for type 1 of normalization is not visible because it coincides with the curve of type 2 of normalization for this distance. Also, from the figures it is clear that the most unstable case is for Madelon for norm distances.

5.3 Clustering application

For the comparison of multiplicative distance with other distance functions, the performance of the aforementioned distances is investigated on a real application. For this purpose, we choose clustering. There are many clustering algorithms but maybe the most popular one is k-means [42]. If the dataset contains n points of $X_1, X_2, \ldots, X_n$, then the k-means clustering is the optimization process of grouping them into k clusters so that the global criterion function:

$$\sum_{j=1}^{k}\sum_{l=1}^{n} f(X_l, C_j)$$

is either minimized or maximized. C_j represents the centroid of the cluster j, for $j = 1, ..., k$, and $f(X_l, C_j)$ is the clustering criterion function for a point X_l and a centroid C_j. If f is a distance function, the global criterion must be minimized. The steps of k-means algorithm are as follows:

(a) Select k initial cluster centroids.
(b) For each point of the dataset, compute the clustering criterion function with each cluster centroid. Assign each point to its best choice, i.e., the nearest centroid.

(c) Recalculate k centroids based on the documents assigned to them.
(d) Repeat Steps (b) and (c) until convergence.

A number of clustering evaluation techniques exist in the literature. The most commonly used metrics are cluster recall and cluster precision [7,29,30,43]. Each cluster can be considered as the result of a query, whereas each preclassified set of points can be considered as the desired set of points for that query. Thus, the precision $P(i, j)$ and recall $R(i, j)$ of each cluster j for each class i can be calculated. If n_i is the number of the members of the class i, m_j is the number of the members of the cluster j, and k_{ij} is the number of the members of the class i in the cluster j, then $P(i, j)$ and $R(i, j)$ can be defined as:

$$P(i, j) = \frac{k_{ij}}{m_j}$$
$$R(i, j) = \frac{k_{ij}}{n_i} \tag{21}$$

The overall clustering performance is measured by F-measure value. In other words, the F-measure is a harmonic combination of the precision and recall values. The F-measure is a standard evaluation metric in the field of the information retrieval [7,29,30,43]. The corresponding F-measure $F(i, j)$ for $P(i, j)$ and $R(i, j)$ is defined as:

$$F(i, j) = \frac{\left(\beta^2 + 1\right) \times P(i, j) \times R(i, j)}{\left(\beta^2 \times P(i, j)\right) + R(i, j)} \tag{22}$$

where β is the relative importance of clustering precision versus clustering recall. In this paper, β is chosen to be 1, assigning equal importance to cluster precision and recall. Then, the F-measure for the whole clustering result is defined as [29,43]:

$$F = \sum_i \frac{n_i}{n} \max_j (F(i, j)) \tag{23}$$

The F-measure values are in the interval [0, 1]. The larger F-measure value indicates the higher clustering quality. In Tables 7 and 8, the F-measures have been shown for the datasets for two types of normalizations.

Based on Tables 7 and 8, the multiplicative distance outperforms norm distances for high-dimensional clustering. Furthermore, it can be observed that for norm distances, usually the

Table 7 F-measure for type 1 of normalization

Dataset name	L_1	L_2	$L_{0.79}$	Multiplicative distance
Iris	0.8589	0.8332	**0.8662**	0.8647
Liver	0.6093	0.6808	0.6510	**0.7005**
Satellite	0.7240	0.7159	0.7237	**0.7265**
Sonar	0.5372	0.6072	0.5473	**0.6234**
Usps	0.6982	0.6811	0.6841	**0.7143**
Madelon	0.5773	0.5765	0.5784	**0.5882**
Gisette	0.6125	0.6119	0.6233	**0.6659**
Leu	0.5594	0.5558	0.5642	**0.6799**
Farm	0.6645	0.6656	0.6656	**0.7565**

Best results are indicated in bold

Table 8 F-measure for type 2 of normalization

Dataset name	L_1	L_2	$L_{0.79}$	Multiplicative distance
Iris	**0.8988**	0.8853	0.8847	0.8847
Liver	**0.6510**	**0.6510**	**0.6510**	**0.6510**
Satellite	0.7248	0.7197	0.7240	**0.7252**
Sonar	0.5675	0.5579	0.5326	**0.6129**
Usps	0.7015	0.6892	0.7053	**0.7126**
Madelon	0.5784	0.5815	0.5815	**0.5908**
Gisette	0.6119	**0.6791**	0.6056	0.6607
Leu	0.5594	0.5558	0.5514	**0.6897**
Farm	**0.6656**	**0.6656**	0.6360	0.6424

Best results are indicated in bold

Table 9 F-measure for Madelon dataset for different values of c

Normalization	$c = 1$	$c = 0.5$	$c = 0.25$	$c = 0.05$	$c = 0.025$
Type 1	**0.5882**	**0.5882**	**0.5882**	0.5861	0.5724
Type 2	**0.5908**	**0.5908**	0.5891	0.5683	0.5307

Best results are indicated in bold

Table 10 F-measure for Farm dataset for different values of c

Normalization	$c = 0.01$	$c = 0.005$	$c = 0.0025$	$c = 0.0001$	$c = 0.00005$
Type 1	**0.7565**	**0.7565**	**0.7565**	0.7325	0.7317
Type2	**0.6424**	**0.6424**	**0.6424**	0.6210	0.6193

Best results are indicated in bold

second normalization has better results; but for the multiplicative distance, the first kind of normalization has superiority to the second type. Moreover, the multiplicative distance has a good performance even for the low-dimensional data.

It is useful to note that although changing the control power affects the Pearson variation and relative contrast, it does not have an effect on the F-measure for the multiplicative distance provided that with changing the control power, the aforementioned stability criteria (Pearson variation and relative contrast) have large values. In this situation, slightly changing the centroids does not change the points belonging to each centroid unless the changes of the query points are considerable; or the relative contrast or the Pearson variation are near to zero. For considering the effect of the control power on the clustering performance, different values of c have been applied on the multiplicative distance for Madelon (synthetic) and Farm (real) datasets. The results are displayed in Tables 9 and 10. As it is seen, when the control power changes, until Pearson variation is large, the F-measure remains constant. When small c is used such that Pearson variation is small, the performance decreases.

For considering the effect of noise on these distance functions, a white Gaussian noise with zero mean and unit variance is added to each feature of the un-normalized data and then normalization is performed. The F-measure results for noisy datasets are displayed in Tables 11 and 12. These results have been obtained by averaging on 100 experiments. The values in the parentheses represent standard errors in percentage.

Table 11 F-measure for noisy datasets for type 1 of normalization

Dataset name	L_1	L_2	$L_{0.79}$	Multiplicative distance
Iris	0.6473 (5.2 %)	**0.6915 (5.3 %)**	0.6786 (4.9 %)	0.6742 (4.6 %)
Liver	0.5837 (4.0 %)	0.6792 (3.1 %)	0.6361 (4.3 %)	**0.6975 (2.5 %)**
Satellite	0.7114 (3.9 %)	0.7062 (4.8 %)	0.7153 (3.9 %)	**0.7174 (1.7 %)**
Sonar	0.5371 (7.5 %)	0.5377 (6.4 %)	0.5308 (7.8 %)	**0.5665 (3.8 %)**
Usps	0.6525 (7.9 %)	0.6571 (6.2 %)	0.6428 (8.2 %)	**0.6832 (0.7 %)**
Madelon	**0.5840 (11.6 %)**	0.5796 (11.8 %)	0.5770 (11.6 %)	0.5731 (0.9 %)
Gisette	0.5994 (10.4 %)	0.6004 (8.5 %)	0.6155 (10.6 %)	**0.6548 (1.6 %)**
Leu	0.5407 (7.6 %)	0.5416 (8.3 %)	0.5249 (7.3 %)	**0.6360 (2.3 %)**
Farm	0.5649 (3.2 %)	0.5362 (3.7 %)	0.5563 (4.2 %)	**0.5835 (2.0 %)**

Best results are indicated in bold

Table 12 F-measure for noisy datasets for type 2 of normalization

Dataset name	L_1	L_2	$L_{0.79}$	Multiplicative distance
Iris	0.6386 (4.8 %)	0.6647 (5.5 %)	0.6711 (4.1 %)	**0.6965 (5.2 %)**
Liver	0.5667 (7.0 %)	**0.5739 (7.9 %)**	0.5686 (6.2 %)	0.5636 (5.0 %)
Satellite	0.6993 (7.2 %)	0.6984 (7.4 %)	0.7006 (6.9 %)	**0.7064 (2.9 %)**
Sonar	0.5309 (9.6 %)	0.5256 (10.3 %)	0.5260 (9.4 %)	**0.5527 (3.6 %)**
Usps	0.6624 (5.4 %)	0.6605 (7.6 %)	0.6620 (5.1 %)	**0.6891 (0.8 %)**
Madelon	**0.5826 (12.5 %)**	0.5817 (12.6 %)	0.5774 (12.3 %)	0.5780 (2.4 %)
Gisette	0.6018 (7.4 %)	**0.6671 (10.6 %)**	0.6029 (5.9 %)	0.6502 (1.1 %)
Leu	0.5485 (10.7 %)	0.5375 (10.8 %)	0.5246 (10.5 %)	**0.6247 (3.1 %)**
Farm	0.5216 (2.2 %)	0.5183 (4.1 %)	0.5438 (1.6 %)	**0.5761 (1.8 %)**

Best results are indicated in bold

In the noisy case, usually, the multiplicative distance function has better F-measures in comparison with norm distance functions. Based on the experiments, it is concluded that both kinds of normalizations yield good results for the multiplicative distance in terms of clustering performance, although the first type of normalization slightly outweighs.

6 Conclusion

In the high-dimensional data space, many distance functions become unstable under a broad set of conditions. In this paper, it was proved that for independent dimensions with finite appropriate moments, norm distance functions become unstable without any condition on the normalization of dimensions. Moreover, a stable distance function in the high-dimensional data, named multiplicative distance, was introduced. This distance function is based on the multiplication of distance components, contrarily to the usual distances functions which are based on the summation of distance components. It was theoretically proved that the multiplicative distance function was stable for data with independent dimensions (with identical or nonidentical distribution) in high-dimensional data space. Experimental results showed the superiority of the multiplicative distance over the norm distance functions in terms of

stability and clustering results for data with independent and correlated dimensions in the high-dimensional space. Furthermore, the multiplicative distance functions could be used in low-dimensional space in applications such as clustering.

References

1. Aggarwal CC, Hinneburg A, Keim DA (2001) Parametric detection of meaningless distances in high dimensional data. In: Proceedings international conference on database theory (ICDT), pp 420–434
2. Aggarwal CC, Yu PS (2000) The IGrid index: reversing the dimensionality curse for similarity indexing in high dimensional space. In: Proceedings sixth ACM SIGKDD international conference on knowledge discovery and data mining (KDD '00), pp 119–129
3. Bassani HF, Araujo AFR (2012) Dimension selective self-organizing maps for clustering high dimensional data. In: Proceedings international conference on neural networks (IJCNN), pp 1–8
4. Beyer K, Goldstein J, Ramakrishnan R, Shaft U (1999) When is nearest neighbors meaningful? In: Proceedings seventh international conference on database theory (ICDT '99), vol. 1540, pp 217–235
5. Chang J, Lee A (2008) Parallel high-dimensional index structure for content-based information retrieval. In: Proceedings 8th IEEE international conference on computer and information technology, pp 101–106
6. Cheng Q (2010) A sparse learning machine for high-dimensional data with application to microarray gene analysis. IEEE/ACM Trans Comput Biol Bioinform 7(4):636–646
7. Chu Y-H, Huang J-W, Chuang K-T, Yang D-N, Chen M-S (2010) Density conscious subspace clustering for high-dimensional data. IEEE Trans Knowl Data Eng 22(1):16–30
8. Cui J, Xiao B, Yin Z (2010) Speed up linear scan in high-dimensions using extended B+-tree. In: Proceedings 2nd international workshop on database technology and applications (DBTA), pp 1–4
9. Deegalla S, Bostrom H (2006) Reducing high-dimensional data by principal component analysis vs. random projection for nearest neighbor classification. In: Proceedings international conference on machine learning and applications, pp 245–250
10. Demartines P (1994) Analyse de donne'es par re'seaux de neurones auto-organise's. PhD dissertation, Institut Nat'l Polytechnique de Grenoble, Grenoble, France (in French)
11. Durrant RJ, Kaban A (2009) When is 'nearest neighbour' meaningful: a converse theorem and implications. J Complex 25(4):385–397
12. Ertoz L, Steinbach M, Kumar V (2003) Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data. In: Proceedings SIAM international conference on data mining
13. Fauvel M, Chanussot J, Benediktsson JA, Villa A (2013) Parsimonious Mahalanobis kernel for the classification of high dimensional data. Pattern Recognit 46(3):845–854
14. Francois D, Wertz M-V, Verleysen SM-M (2007) The concentration of fractional distances. IEEE Trans Knowl Data Eng 19(7):873–886
15. Gu X, Zhang Y, Zhang L, Zhang D, Li J (2013) An improved method of locality sensitive hashing for indexing large-scale and high-dimensional features. Signal Process 93(8):2244–2255
16. Hasan A, Adnan MA (2012) High dimensional microarray data classification using correlation based feature selection. In: Proceedings international conference on biomedical engineering (ICoBE), pp 319–321
17. He Q, Wang Q, Du C-Y, Ma X-D, Shi Z-Z (2010) A parallel hyper-surface classifier for high dimensional data. In: Proceedings international symposium on knowledge acquisition and modeling, pp 338–343
18. Hinneburg A, Aggarwal CC, Keim DA (2000) What is the nearest neighbor in high dimensional spaces? In: Proceedings 26th international conference on very large data bases, pp 506–515
19. Hsu C-M, Chen M-S (2009) On the design and applicability of distance functions in high-dimensional data space. IEEE Trans Knowl Data Eng 21(4):523–536
20. Huang S-C, Wu, T-K (2012) Robust semi-supervised SVM on kernel partial least discriminant space for high dimensional data mining. In: Proceedings international conference on information science and applications (ICISA), pp 1–6
21. Jagadish HV, Ooi BC, Tan K-L, Yu C, Zhang R (2005) iDistance: an adaptive B+-tree based indexing method for nearest neighbor search. ACM Trans Database Syst 30(2):364–397
22. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323
23. Kabán A (2011) On the distance concentration awareness of certain data reduction techniques. Pattern Recognit 44(2):265–277
24. Kabán A (2012) Non-parametric detection of meaningless distances in high dimensional data. Stat Comput 22:375–385

25. Koudas N, Ooi BC, Shen HT, Tung AKH (2004) LDC: enabling search by partial distance in a hyper-dimensional space. In: Proceedings international conference on data engineering, pp 6–17
26. Ledoux M (2001) The concentration of measure phenomenon. Mathematical Surveys and Monographs, vol 89. American Mathematical Society
27. Lee G, Rodriguez C, Madabhushi A (2008) Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies. IEEE/ACM Trans Comput Biol Bioinform 5(3):368–384
28. Lejsek H, Asmundsson FH, Jonsson BT, Amsaleg L (2009) NV-Tree: an efficient disk-based index for approximate search in very large high-dimensional collections. IEEE Trans Pattern Anal Mach Intell 31(5):869–883
29. Li Y, Luo C, Chung SM (2008) Text clustering with feature selection by using statistical data. IEEE Trans Knowl Data Eng 20(5):641–652
30. Liang J, Vaishnavi VK, Vandenberg A (2006) Clustering of LDAP directory schemas to facilitate information resources interoperability across organizations. IEEE Trans Syst Man Cybern A Syst Hum 36(4):631–642
31. Liu H, Wei R-X, Jiang G-P (2012) Similarity measurement for data with high-dimensional and mixed feature values through fuzzy clustering. In: Proceedings international conference on computer science and automation engineering (CSAE), vol 3, pp 617–62
32. Mo D, Huang SH (2012) Fractal-based intrinsic dimension estimation and its application in dimensionality reduction. IEEE Transa Knowl Data Eng 24(1):59–71
33. Okada S, Nishida T (2011) Online incremental clustering with distance metric learning for high dimensional data. In: Proceedings international joint conference on neural networks, pp 2047–2054
34. Radovanovi'c M, Nanopoulos A, Ivanovi'c M (2010) On the existence of obstinate results in vector space models. In: Proceedings 33rd international ACM SIGIR conference on research and development in information retrieval, New York, pp 186–193
35. Samiappan S, Prasad S, Bruce LM (2013) Non-uniform random feature selection and kernel density scoring with SVM based ensemble classification for hyperspectral image analysis. IEEE J Sel Top Appl Earth Obs Remote Sens 6(2):792–800
36. Triguero I, Derrac J, Herrera F (2012) A taxonomy and experimental study on prototype generation for nearest neighbor classification. IEEE Trans Syst Man Cybern C Appl Rev 42(1):86–100
37. Turkay C, Lundervold A, Lundervold AJ, Hauser H (2012) Representative factor generation for the interactive visual analysis of high-dimensional data. IEEE Trans Vis Comput Gr 18(12):2621–2630
38. Varshney KR, Willsky AS (2011) Linear dimensionality reduction for margin-based classification: high-dimensional data and sensor networks. IEEE Trans Signal Process 59(6):2496–2512
39. Weber R, Schek H-J, Blott S (1998) A quantitative analysis and performance study for similarity search methods in high-dimensional spaces. In: Proceedings very large data base conference (VLDB'98), pp 194–205
40. Wei B, Guan T, Yu J (2014) Projected residual vector quantization for ANN search. IEEE Multimed 21(3):41–51
41. Wu C, Yang H, Zhu J, Zhang J, King I, Lyu RM (2013) Sparse Poisson coding for high dimensional document clustering. In: Proceedings IEEE international conference on big data, pp 512–517
42. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D (2008) Top 10 algorithms in data mining. Knowl Inf Syst 14(1):1–37
43. Xiong H, Wu J, Chen J (2009) K-Means clustering versus validation measures: a data-distribution perspective. IEEE Trans Syst Man Cybern B Cybern 39(2):318–331
44. Xu M, Chen H, Varshney PK (2013) Dimensionality reduction for registration of high-dimensional data sets. IEEE Trans Image Process 22(8):3041–3049
45. Yasen Z, Xinwei Z, Ge L, Xian S, Hongqi W, Kun F (2014) Semi-supervised manifold learning based multigraph fusion for high-resolution remote sensing image classification. IEEE Lett Geosci Remote Sens 11(2):464–468
46. Yu J, Wang M, Tao D (2012) Semisupervised multiview distance metric learning for cartoon synthesis. IEEE Trans Image Process 21(11):4636–4648
47. Zhang H, Li N (2012) K-Dvd-Tree: a high dimensional data index applying to SIFT feature matching. In: Proceedings fifth international symposium on computational intelligence and design (ISCID), vol 2, pp 14–17

Jafar Mansouri is a Ph.D. candidate at Ferdowsi University of Mashhad, Iran. His main research interests include image and video processing and analysis, multimedia information retrieval, and machine learning.

Morteza Khademi was born in Iran in 1958. He received the B.Sc. and M.S. degrees from Isfahan University of Technology, Isfahan, Iran, in 1985 and 1987, respectively, and the Ph.D. degree from the University of Wollongong, Wollongong, Australia, on video communications in 1995, all in Electrical Engineering. He joined Ferdowsi University of Mashhad, Iran, in 1987. He is currently professor at the Department of Electrical Engineering, Ferdowsi University of Mashhad. Since then, he has co-chaired two conferences on "Electrical Engineering (ICEE2004)" and "Machine Vision and Image Processing (MVIP2006)" in Iran. He has received two awards including: Outstanding Graduate Student Award in 1999 and The Best Translation Award for translation of "Digital Image Processing by Gonzales" from Amir Kabir University, Iran, in 2005. His current research interests are in the areas of video communications, biomedical signal processing, and data analysis. He has published over 90 articles in these research fields.