

Effect of Data Standardization on the Result of k -Means Clustering

Kensuke Tanioka and Hiroshi Yadohisa

Abstract In applying clustering to multivariate data, in which there are some large-scale variables, clustering results depend on the variables more than the user's needs. In such cases, we should standardize the data to control the dependency.

For high-dimensional data, Doherty et al. (Appl Soft Comput 7:203–210, 2007) argued numerically that data standardization by variable range leads to almost the same results regardless of the kinds of norms, although Aggarwal et al. (Lect Notes Comput Sci 1973:420–434, 2001) showed theoretically that a fraction norm reduces the effect of the curse of high dimensionality for k -means result more than the Euclidean norm does. However, they have not considered the effects of standardization and factors properly.

In this paper, we verify the effects of six data standardization methods with various norms and examine factors that affect the clustering results for high-dimensional data. As a result, we show that data standardization with the fraction norm reduces the effect of the curse of high dimensionality and gives a more effective result than data standardization with the Euclidean norm and not applying data standardization with the fraction norm.

K. Tanioka (✉)

Graduate School of Culture and Information Science, Doshisha University Kyoto, 610-0313, Japan

e-mail: dik0012@mail4.doshisha.ac.jp

H. Yadohisa

Department of Culture and Information Science, Doshisha University Kyoto 610-0313, Japan

e-mail: hyadohis@mail.doshisha.ac.jp

1 Introduction

K -means clustering is a non-hierarchical clustering method that classifies objects into groups with multivariate data. If there exist some variables, that have a large scale or great variability, these variables strongly affect the clustering result. In these cases, data standardization would be used to control the scale or variability.

Nonetheless, no single approach to data standardization has been shown to be the best. [Milligan and Cooper \(1988\)](#) showed the effect of standardization methods on hierarchical clustering results and concluded that standardization by variable range is the most effective. For k -means clustering, [Steinley \(2004\)](#) indicated that standardization by the maximum of variables is also effective.

For high-dimensional data, [Aggarwal et al. \(2001\)](#) showed theoretically that clustering by using the fraction norm provides more distinct results than by using the Euclidean norm. In contrast, [Doherty et al. \(2007\)](#) argued numerically that data standardization resulted in the k -means clustering with the Euclidean norm outperforming k -means with the fraction norm. Moreover, despite their differences, these papers have not considered the effects of:

- (a) Factors such as noise dimensions, outlier conditions and cluster size.
- (b) Other standardization methods on the results of k -means clustering by using a non-Euclidean norm.

As a result, in this paper, we examine the effect of six data standardization methods on the result of k -means clustering through Monte Carlo simulation using a non-Euclidean norm because for real data, we can't verify whether the factor affects the clustering result or not. The purpose of this simulation is to compare how data standardizations with the fraction norm affect various factors such as error conditions, variance of variables, cluster configurations and so on.

2 Fraction Norm

Here we introduce the fraction norm for k -means clustering. The Minkowski norm, L_p , which is a family of distance measures, is described as:

$$L_p(\mathbf{x}, \mathbf{y}) = \left\{ \sum_{i=1}^d |x_i - y_i|^p \right\}^{1/p}, \quad (p \geq 1) \quad (1)$$

where d is the number of dimension of vectors. [Aggarwal et al. \(2001\)](#) extended this concept to $p \in (0, 1)$ for reducing the effect of the curse of high dimensionality. [Figure 1](#) represents a unit length from the origin in the Euclidean plane with L_p norms ($p = 0.3, 0.5, 1, 2$). $L_{0.3}$ shows an inwards-curved loci, although L_2 traces an outwards-curved loci. The properties of the fraction norm are as follows:

Fig. 1 Unit length loci from the origin with L_p ($p = 2, 1, 0.5, 0.3$) norms

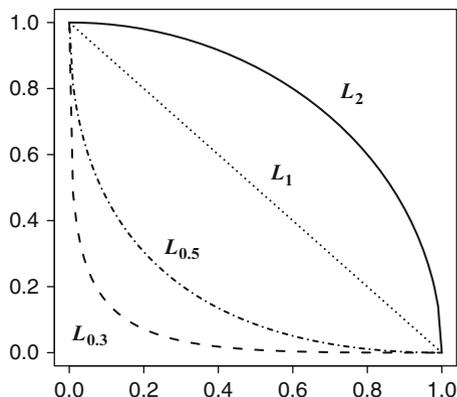


Table 1 Norms of vectors \mathbf{a} and \mathbf{b} from origin measured with L_p

Norm	$L_p(\mathbf{a})$	$L_p(\mathbf{b})$	$L_p(\mathbf{b})/L_p(\mathbf{a})$
$p = 2$	8.83	50.29	5.69
$p = 1/3$	238.62	506.95	2.12

(1) the fraction measures defined by L_p with $p \in (0, 1)$ do not satisfy the triangle inequality; (2) it reduces the effect of outlier values compared with Euclidean measures.

For example, consider the two vectors $\mathbf{a} = (2, 3, 4, 7)$ and $\mathbf{b} = (2, 3, 4, 50)$, and let $L_p(\mathbf{x})$ be the norm between vector \mathbf{x} and the origin. Table 1 shows the length of vectors \mathbf{a} , \mathbf{b} and the ratio \mathbf{a} to \mathbf{b} measured with the L_2 and $L_{1/3}$ norm. The ratio of the $L_{1/3}(\mathbf{a})$ to $L_{1/3}(\mathbf{b})$ is less than the ratio of the $L_2(\mathbf{a})$ to $L_2(\mathbf{b})$.

3 Methods of Standardization

In this section, we present the six standardization methods and describe their properties using Monte Carlo simulation. First, we define

$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$, the $\mathbf{N} \times \mathbf{n}$ data matrix, where \mathbf{x}_j is the column vector;

$\mathbf{1} =$ an $\mathbf{N} \times 1$ vector of ones;

$\bar{\mathbf{x}}_j =$ mean of variable \mathbf{x}_j ($\mathbf{j} = 1, 2, \dots, \mathbf{n}$);

$\sigma_j =$ standard deviation of variable \mathbf{x}_j ($\mathbf{j} = 1, 2, \dots, \mathbf{n}$);

$\max(\mathbf{x}_j) =$ maximum of variable \mathbf{x}_j ($\mathbf{j} = 1, 2, \dots, \mathbf{n}$);

$\min(\mathbf{x}_j) =$ minimum of variable \mathbf{x}_j ($\mathbf{j} = 1, 2, \dots, \mathbf{n}$);

$\text{Rank}(\mathbf{x}_j) =$ within – variable ranking of variable \mathbf{x}_j ($\mathbf{j} = 1, 2, \dots, \mathbf{n}$);

$$x_j^q = x_{(a)j} + b(x_{(a+1)j} - x_{(a)j}) \quad (0 \leq q \leq 1; j = 1, 2, \dots, n);$$

where $x_{(c)}$ is an order statistic and \mathbf{a} and \mathbf{b} are defined by:

$$a = [q(n + 1)], \quad b = q(n + 1) - a.$$

In short, x_j^q is the 100 q th percentile.

The original data and the six standardization functions are as follows:

$$Z_0(\mathbf{x}_j) = \mathbf{x}_j, \tag{2}$$

$$Z_1(\mathbf{x}_j) = (\mathbf{x}_j - \bar{x}_j \mathbf{1}') / \sigma_j, \tag{3}$$

$$Z_2(\mathbf{x}_j) = \mathbf{x}_j / \max(\mathbf{x}_j), \tag{4}$$

$$Z_3(\mathbf{x}_j) = \mathbf{x}_j / (\max(\mathbf{x}_j) - \min(\mathbf{x}_j)), \tag{5}$$

$$Z_4(\mathbf{x}_j) = \mathbf{x}_j / N \bar{x}_j, \tag{6}$$

$$Z_5(\mathbf{x}_j) = \text{Rank}(\mathbf{x}_j), \tag{7}$$

and

$$Z_6(\mathbf{x}_j) = \mathbf{x}_j / (x_j^{0.975} - x_j^{0.025}). \tag{8}$$

Z_0 represents the original data. Z_1 , Z_2 , Z_3 , Z_4 and Z_5 were examined by [Milligan and Cooper \(1988\)](#) and Z_6 was recommended by [Mirkin \(2005\)](#).

Although [Milligan and Cooper \(1988\)](#) concluded that Z_3 is the most effective compared to the other standardizations, Z_3 is affected by outliers. In particular, under the outlier condition, applying Z_3 to each variable in the data matrix causes the scales of those variables, that contain outliers to become smaller than the scale of those variables that do not. However, Z_6 provides a more equal scale for the variables than Z_3 .

4 Simulation Design

We use a Monte Carlo simulation to verify the effects of data standardization on the result of k -means clustering. The method for evaluating clustering results is to generate data with a known cluster structure and compare the recovered cluster structure with the known cluster structure on several data structures. The design of this simulation is based on [Milligan \(1985\)](#), [Milligan and Cooper \(1988\)](#), [Steinley \(2004\)](#), and [Steinley and Henson \(2005\)](#), with some modifications. For the data generation procedure, overlap between clusters is adjusted on the first dimension of the variable space ([Steinley and Henson 2005](#)). For all other dimensions, clusters

Table 2 Factors in the simulation

Factor No.	Factor name	Factor No.	Factor name
Factor 1	Number of clusters	Factor 6	Distribution of variables
Factor 2	Number of variables	Factor 7	Variance of variables
Factor 3	Number of observations	Factor 8	Error conditions
Factor 4	Cluster densities	Factor 9	Dissimilarities
Factor 5	Initial seeds	Factor 10	Probability that clusters overlap

are allowed to either overlap or not and the maximum range of the data is limited to be two-thirds of the range of the first dimension. In this simulation, we use 10 factors to retain the validity of the simulation and use the adjusted Rand index (ARI) to evaluate the clustering results (Hubert and Arabie 1985).

Next, we describe the levels of the 10 factors shown in Table 2.

Number of clusters: The number of clusters has three levels, 5, 10 and 20 (Steinley and Brusco 2007).

Number of variables: The number of variables has two levels, 25 and 50 (Steinley and Brusco 2007).

Number of observations: The number of observations is 200 because the number of observations hardly influences the recovery of the known cluster structure (Steinley and Brusco 2007).

Cluster densities: This factor has three levels:

- (a) All clusters have the same number of observations.
- (b) Three clusters have 50% of the observations while the remaining observations are evenly divided among the remaining clusters.
- (c) Three clusters have 20% of the observations while the remaining observations are evenly divided among the remaining clusters.

Initial seeds: K -means clustering depends on the initial seeds. Thus, we perform k -means clustering 10 times, which is the number of cluster updates, and on each implementation randomly generate the initial seeds. The number of times the initial seeds are generated is 200. The classification result is selected according to the criterion that minimizes the overall within cluster variance.

Distribution of variables: The distribution of each variable is selected randomly from a normal distribution, uniform distribution and triangular distribution (Steinley and Henson 2005).

Variance of variables: This factor has two levels:

- (a) The variance of variables is not manipulated.
- (b) Half of the number of variables are multiplied by random numbers between 5 and 10 (Steinley 2004).

Error conditions: This factor has three levels:

- (a) Error-free: Neither outliers nor noise dimensions are added to the generated data.
- (b) Outliers: Outliers (20% of the observations) are added to the error-free data.
- (c) Noise dimensions: Half of the number of dimensions of the error-free data is replaced with noise dimensions where the noise dimensions are distributed from uniform distributions. The range of the noise dimensions is equal to the first dimension.

Dissimilarities: This factor has two levels: Euclidean norm, $p = 2$ and fraction norm, $p = 0.3$.

Probability that clusters overlap: On the first dimension, the probability that two clusters overlap is manipulated by five levels, 0.01, 0.1, 0.2, 0.3 and 0.4 (Steinley and Henson 2005).

5 Results

In this section, we show the results of the Monte Carlo simulation from two perspectives, error conditions and probability that clusters overlap.

Table 3 shows the recovery results for the data standardization methods under various error conditions and dissimilarities. Each cell in the table represents the average value of ARI. Under any error condition, the data standardizations with the fraction norm provide more effective results on the k -means methods than data standardizations with the Euclidean norm. Using the fraction norm reduces the difference among the recoveries from the various data standardization methods more than the Euclidean norm.

In the error-free conditions, Z_5 with a Euclidean norm produces a higher recovery than any other standardization method with a Euclidean norm. When using the fraction norm, Z_2 and Z_3 are also effective for the result of k -means clustering.

As is well known, clustering results of k -means methods are affected by outliers. In the outlier conditions, Z_5 shows the highest recovery of any other standardization

Table 3 Effect of the error conditions and kinds of norm on the recovery for the six standardization methods

Error conditions Norm	Error-free		Outliers		Noise dimensions	
	L_2	$L_{0.3}$	L_2	$L_{0.3}$	L_2	$L_{0.3}$
Z_0	0.25	0.45	0.12	0.32	0.11	0.51
Z_1	0.29	0.53	0.11	0.39	0.43	0.64
Z_2	0.37	0.59	0.25	0.48	0.46	0.68
Z_3	0.37	0.58	0.24	0.49	0.47	0.67
Z_4	0.32	0.52	0.19	0.38	0.40	0.66
Z_5	0.55	0.66	0.51	0.64	0.56	0.64
Z_6	0.34	0.56	0.12	0.38	0.47	0.66

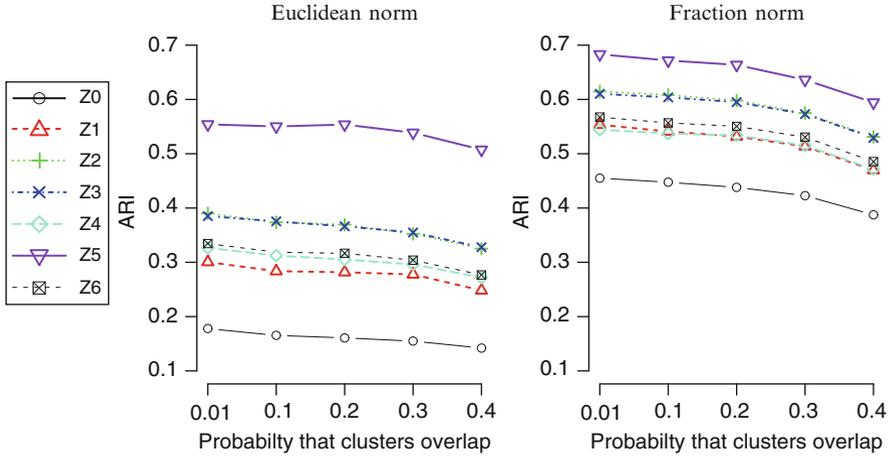


Fig. 2 Effect of probability that clusters overlap on the recovery for the standardization methods

methods regardless of norm, although the recoveries of the other standardization methods are lower than those under any error conditions as a whole. Z_6 gives a lower recovery than Z_2 and Z_3 , even though Z_6 provides more equal scale of variables than Z_2 and Z_3 in the error condition.

High-dimensional data usually contains many noise variables. In the noise dimension conditions, Z_0 with the Euclidean norm is affected by noise dimensions, although the other methods gave stable recoveries. The reason is that the data values in the noise dimension is uniformly-distributed in the range and does not affect the k -means results, although the scale affects the k -means result.

Figure 2 represents the recoveries with respect to the probability that clusters will overlap on the first dimension. Data standardizations with the fraction norm provide higher recoveries than those with a Euclidean norm regardless of the probability. As a result, Z_5 provides the most effective recovery on the k -means method regardless of norm. The reason is quite understandable: Z_5 gives the highest recovery of any method in the outliers condition.

6 Conclusion

We can obtain two conclusions from this simulation. Firstly, as a whole, the recoveries of data standardization methods with the fraction norm reduce the effect of the curse of high dimensionality more than with the Euclidean norm, although [Doherty et al. \(2007\)](#) argued that data standardization leads to the same clustering results regardless of norm. However, this simulation partly supports the results of [Doherty et al. \(2007\)](#) that for the fraction norm, the differences of the recoveries between standardization methods are smaller than those for the Euclidean norm.

Secondarily, the result in our simulation indicates that Z_5 is the most effective for k -means clustering of all standardization methods we tested, whereas Steinley (2004) concluded that Z_2 and Z_3 are the most effective. The difference occurs because Steinley (2004) only considered variables distributed from normal distributions. In particular, k -means clustering postulates that each cluster detected is distributed from a normal distribution. However, real data does not always follow a normal distribution. In this simulation, we also employed two kinds of distributions that are not normal distributions. Thus, the recoveries in this simulation are lower than those in Steinley (2004) and Z_2 and Z_3 are affected by the effect of the variables that follow non-normal distributions. For example, when variables are distributed as a triangular distribution, which represents a left-or right-skewed distribution, and the probability that clusters overlap is high, k -means clustering tends to consider the data around the highest probability of the distribution with the data generated from the neighboring distribution as a cluster. However, Z_5 does not consider the data around the highest probability of the distribution, but considers the ranking of the variable.

The results of this study indicate that we should apply Z_5 to multivariate data with the fraction norm when we classify the observations by the k -means method for high-dimensional data. The reason is quite understandable. The fraction norm is more effective for reducing the curse of dimensionality than the Euclidean norm and Z_5 is less affected by the distributions of variables than the other standardization methods.

Although we provide several results in this simulation, the simulation design still requires some modifications. For example, we did not consider the properties of high-dimensional data such as sparsities and masking variables. Thus, the effects of these factors on the clustering results need to be verified through Monte Carlo simulations. Finally, wherever feasible in the future, we would like to prove that data standardization with the fraction norm reduces the effect of the curse of high dimensionality on k -means clustering more effectively than data standardization without the fraction norm.

References

- Aggarwal CC, Hinneburg A, Keim DA (2001) On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science* 1973:420–434
- Doherty KAJ, Adams RG, Davey N (2007) Unsupervised learning with normalised data and non-euclidean norms. *Applied Soft Computing* 7:203–210
- Hubert L, Arabie P (1985) Comparing partitions. *Journal of Classifications* 2:193–218
- Milligan GW (1985) An algorithm for generating artificial test clusters. *Psychometrika* 50:123–127
- Milligan GW, Cooper MC (1988) A study of standardization of variables in cluster analysis. *J Classification* 5:181–204
- Mirkin B (2005) *Clustering for data mining: A data recovery approach*. Chapman and Hall, Boca Raton, FL

- Steinley D (2004) Standardizing variables in k-means clustering. In: Banks D, House L, McMorris FR, Arabie P, Gaul W (eds) Classification, clustering, and data mining application. Springer, Berlin, pp 53–60
- Steinley D, Brusco MJ (2007) Initializing k-means batch clustering: A critical evaluation of several techniques. *J Classification* 24:99–121
- Steinley D, Henson R (2005) OCLUS: An analytic method for generating clusters with known overlap. *J Classification* 22:221–250