Topology Analysis of Data Space Using Self-Organizing Feature Map

Kazuhiro MINAMIMOTO Kazushi IKEDA Kenji NAKAYAMA
Department of Electrical and Computer Eng., Faculty of Eng., Kanazawa Univ.
2-40-20, Kodatsuno, Kanazawa 921, Japan
E-mail nakayama@ec.t.kanazawa-u.ac.jp

ABSTRACT

In order to analyze the topological structure of the data space using Kohonen's self-organizing feature map (SOFM), a criterion is discussed. The Euclidian distance between the reference vector and the data, the number of the reference vectors and the topology preserving measure are taken into account, and are combined in a unified criterion. Through computer simulation, it is confirmed that goodness of the different reference topologies, that is dimensions, can be clearly discriminated regardless the parameters. Thus, the unified criterion makes it possible to analyze the essential data space topology.

1. Introduction

The self-organizing feature map (SOFM) proposed by Kohonen [1] is useful to map the environment features into the reference vector (RV) space, in which they are arranged in some order. The RV space has its own topology, that is structural constraint. This topology tends to be preserved during the SOFM learning. If this topology matches to that of the feature space, the features can be mapped onto the suitable RVs, at the same time, they are arranged in the plausible order.

On the other hand, the N-dimensional features, simply called "data" in the following, observed in some environment, are not usually distributed in a full N-dimensional space, rather they are partially distributed in limited sub-spaces. In data analysis, it will be important to analyze the essential dimension and the shape of the region, where the data are distributed.

In this paper, we discuss analysis of topological structure of the data space using Kohonen's SOFM. Especially, the essential dimension of the data distribution region is taken into account. For this purpose, criteria for evaluating the SOFM results are investigated.

A first problem is what is a correct topology for the data distribution. For example, N-dimensional data, distributed on a 2-dimensional plane or along a thin tube, has a 2-dimensional topology or a 1-dimensional topology. The next problem is what kinds of measures can be applied to evaluating the SOFM results. For this purpose, we employ the mean-square distance, the topology preservation measures [2],[3] and the number of the RVs. Their efficiency and relations are discussed.

2. Self-Organizing Feature Map

The SOFM, taken into account in this paper, is the original one proposed by Kohonen [1]. Let $\boldsymbol{x}(i), i = 1, \cdots, N_X$, and $\boldsymbol{r}(j), j = 1, \cdots, N_R$, be N-dimensional input vectors and M-dimensional reference vectors (RVs), respectively. Usually, $M \leq N$ and $N_R \ll N_X$. The initial location of $\boldsymbol{r}(i)$ is set to be random. In the SOFM procedure, $\boldsymbol{x}(i)$ is given, and the nearest $\boldsymbol{r}(j')$ is selected as the winner. $\boldsymbol{r}(j')$ and its neighbors are shifted toward the input $\boldsymbol{x}(i)$.

$$\boldsymbol{r}(j', n) = \min_j \{ \| \boldsymbol{r}(j, n) - \boldsymbol{x}(i) \|^2 \} \quad (1)$$

$$\boldsymbol{r}(j, n) = \boldsymbol{r}(j, n) + \mu(n)(\boldsymbol{x}(i) - \boldsymbol{r}(j, n)) \quad (2)$$
$$\text{for } j \in NB(j', n)$$

where $\| \cdot \|$ is an l_2-norm, $\mu(n)$ is a learning rate, n is an iteration number, and $NB(j', n)$ indicates the neighbor of $\boldsymbol{r}(j')$ at the nth-iteration. $\mu(n)$ and $NB(j', n)$ are gradually decreased as the SOFM makes progress.

3. Performance Measure for Optimun Topology

Goodness of the SOFM result is evaluated based on the followings. First, the distance between the input data and the selected RVs, that is the winner is necessary. This means how well the region, where the input data are distributed, is covered by the RVs. Even though the mean-square distance is small, the topology of the RVs is not always match to that of the data distribution. If the resulting topology is twisted from the original one, it cannot exactly represent that of the data distribution. The

0-7803-2768-3/95/$4.00

number of the RVs is also important. When it is small, the given topology can be expected to be the optimum one. Finally, relations among the above three criteria is also important.

3.1. Euclidian Distance

The RVs are used to represent the input vectors. Usually, the number of the RVs is less than that of the input data, $N_X \leq N_R$. Thus, one RV represents a plural number of the input data. This representation is usually evaluated by the Euclidian distance between the winner RV and the corresponding input data. So, we employ the root-mean-square error given by

$$\mathrm{RMSE} = \frac{1}{N_X} \sum ||\boldsymbol{r}(\mathrm{j}) - \boldsymbol{x}(\mathrm{i})|| \qquad (3)$$

3.2. Topology Preservation Measure

First, we must define goodness of the resulting topology. Since the topologies are categorized based on their dimension, the topology preservation measure may be suitable for this purpose. For example, in the case of the 1-dimensional topology, it may be preferred that the RVs are locally arranged on a straight line and are located at equally spaced points. In the same manner, in the case of 2-dimensional topology, it is desired that the RVs are located at the square lattice. These structures can be defined as follows: The receptive field covered by $r(j)$, denoted $R(j)$, borders on $R(j')$, where j' indicates the next neighbors along the given topology. Several kinds of measures based on the above criterion have been proposed [2],[3]. One of them is the topographic function [2], defined by

$$f(j', k) = Num\{j | td[j' - j] > k; R(j') \cap R(j) \neq 0\} \qquad (4)$$

$$\Phi(k) = \sum_j f(j, k) \qquad (5)$$

$f(j', k)$ means the number of $r(j)$, whose receptive field $R(j)$ borders on $R(j')$, and the topological distance $td[j' - j]$ between j' and j is greater than k. $td[\cdot]$ means the distance along the topology of the references. k is selected so as to detect the twisted topology. The large $\Phi(k)$ indicates that its topology is highly deviated from the optimum one, and the small $\Phi(k)$ means its topology is close to the optimum one. Therefore, the topology with a small $\Phi(k)$ can represent that of the data sub-space.

3.3. Number of Reference Vectors

The number of RVs is also an important factor, which can be used to evaluate goodness of the topology. If the given topology is well suited to the input data distribution, then it can be expected that the number of the RVs is small.

4. Combination of Three Measures

4.1. Relations among Three Measures

The next problem is how to combine the above measures, the root-mean-square error (RMSE), the topology preserving measure (TPM) and the number of the RVs (NRV). For this purpose, relation among them is first discussed.

In this paper, it is assumed that the input data are uniformly distributed in some region, which takes an arbitrary shape. Furthermore, we assume the use of the adaptive growing topology method [4]

In this method, the RVs, which are rarely used, are deleted. Thus, the shape is changed from the original one by deleting the unnecessary RVs while maintaining its dimension. In this case, the topology preserving measure becomes very important. If the topology is perfectly preserved, namely $\Phi(k) = 0$, the RMSE may be uniquely determined by the NRV. The topology, with which the RMSE takes the minimum under the same NRV, or the NRV takes the minimum under the same RMSE, can represent the minimum structure of the data distribution region. That is what we want to extract.

4.2. Unified Criterion for Data Topology Analysis

Based on the above discussions, we combine the above three measures to generate a Data-Topology-Measure(DTM).

$$\mathrm{DTM(N, M, N_X, N_R)} = \mathrm{F(RMSE, NRV, TPM)} \qquad (6)$$

In this paper, we employ a linear combination given by

$$\mathrm{DTM(N, M, N_X, N_R)} = \alpha \mathrm{RMSE} + \beta \mathrm{NRV} + \gamma \mathrm{TPM} \qquad (7)$$

α, β and γ are scaling factors, which will be determined so that three criteria perform almost the same contribution.

In searching the optimum topology, it is required that the DTM of the optimum topology (M-dimension) always takes the minimum value regardless the other parameters, N, N_X, N_R. This property will be investigated in the next section through computer simulation.

5. Simulation Results and Discussions

5.1. Simulation Parameters

The input data are basically 3-dimensional vectors with the coordinate (x, y, z). The data are distributed in a box, whose size is $d_x \times d_y \times d_z$. In the simulation, d_x and d_y are fixed to unity, and d_z takes several values, from 0 to 1. Thus the input data are distributed in a cube for $d_z = 1$, and on a plane, that is a 2-dimensional space, for $d_z = 0$.

The number of the data is $N_X = 1000$. 1-, 2- and 3-dimensional topologies are taken into account for the RVs. 2- and 3-dimensional topologies have a square and a cube, respectively. All RVs are used in the SOFM procedure, that is no RV was eliminated.

5.2. Simulation Results

Figure 1 shows examples of the data, which are distributed on a 2-dimensional space, and RV distribution obtained by the SOFM using 1-, 2- and 3-dimensional topologies. Figure 2 shows examples of the RMSE. The learning almost converged around 1000iterations (epoches). Figure 3 (a) shows the RMSE in terms of NRV. The DTM given by Eq.(7) in terms of NRV is shown in Figs.3 (b) and 3 (c), using $(\alpha, \beta, \gamma) = (1.2.5 \times 10^{-4}, 0)$ and $(1.2.5 \times 10^{-4}, 1)$, respectively. These scaling factors are selected so that three terms in Eq.(7) have almost the same level. In calculating TPM, k is set to 1.

Furthermore, the simulations were carried out for different d_z, that is $d_z = 0.3, 0.5, 0.8$. Figure 4 shows the data distributions and the corresponding reference vector distributions. Figure 5 shows the DTM in terms of NRV.

5.3. Discussions

Figure 3 (a) shows the RMSE gradually decrease as the NRV increase. If we have some limitation on the network size, that is NRV, it is not good to increase the NRV. It may have some optimum value. The data topology measure DTM, shown in Fig. 3 (b), includes both the RMSE and the NRV, and have the minimum point. If we can determine the optimum scaling factor for some practical use, it is possible to find the optimum NRV.

In this simulation, the data are distributed on the 2-dimensional plane. Therefore, the data topology analysis should be 2-dimension. In Figs.3 (a) and 3 (b), however, the RMSE and the DTM without the TPM of the 2-dimension are not always the minimum among three dimensions. By adding the TPM further, it is possible to distinguish them as shown in Fig. 3 (c). The DTM of the 2-dimension can take the minimum value regardless the NRV. This means the combination of three measures is important in order to analyze the essential topology of the data distributed space.

In the next simulations, the region of the data distribution approaches to a cube. When $d_z = 0.3$, the DTM for the 2- and 3-dimensional topologies are almost the same. However as d_z is increased more, the DTM for the 3-dimensional topology is decreased compared with the other. Thus, the data distribution with d_z, being larger than 0.5, can be analyzed to have a 3-dimensional structure.

6. Conclusion

In order to analyze the topological structure of the data space using Kohonen's SOFM, a criterion has been discussed in this paper. The root-mean-square error, the number of the reference vectors and the topology preserving measure are taken into account, and are combined in a unified criterion. Through computer simulation, it has been confirmed that difference of goodness of the reference topology, that is dimension, can be made clear regardless the pa rameters. Thus, by using the unified criterion, it is possible to analyze the essential data space topol ogy.

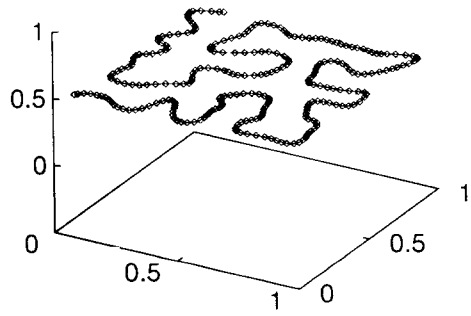
References

[1] T.Kohonen, Self-Organization and Associative Memory, 3rd Ed., Springer-Verlag, 1989.

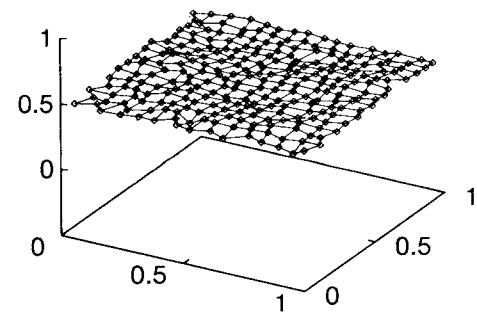
[2] T.Villmann, R.Der and T.Martinetz, "A new qualitative measure of topology preservation in Kohonen's Feature Map", *Proc. IEEE ICNN'94 Florida*,pp.646-648,july 1994.

[3] A.Hamalainen, "A measure of disorder for the self-organizing map", *Proc. IEEE ICNN'94 Florida*,pp.659-664,July 1994.

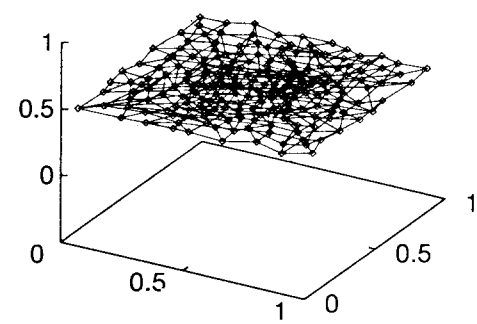
[4] B.Fritzke, "Unsupervised clustering with growing cell structures", *Proc. IEEE IJCNN'91 Seattle*,pp.531-536,July 1991



(a) 1-dimensional topology.
NRV = 250

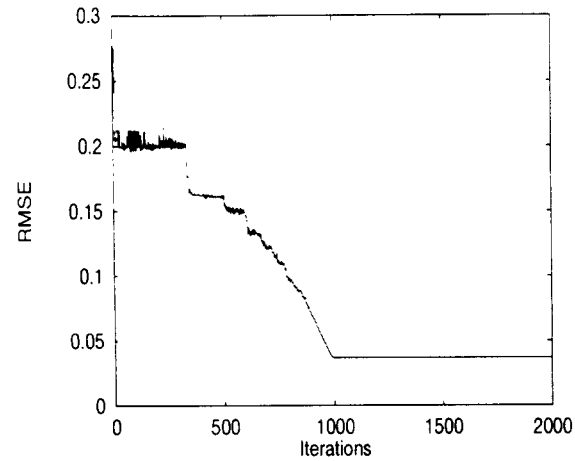


(b) 2-dimensional topology
NRV = 23×23

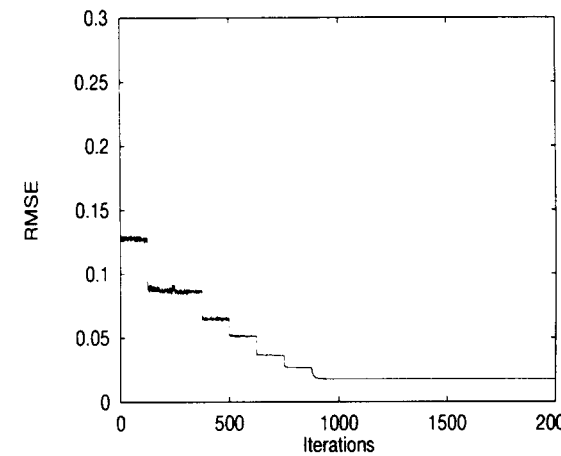


(c) 3-dimensional topology
NRV = 8×8×8

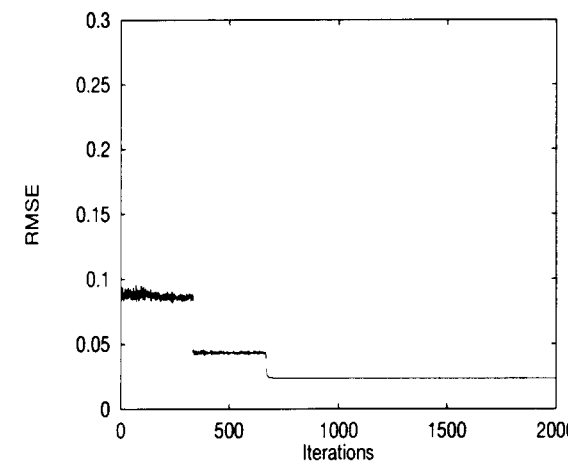
Fig. 1: Distribution of data(·) and reference vectors(◇).



(a) 1-dimensional topology
NRV = 250

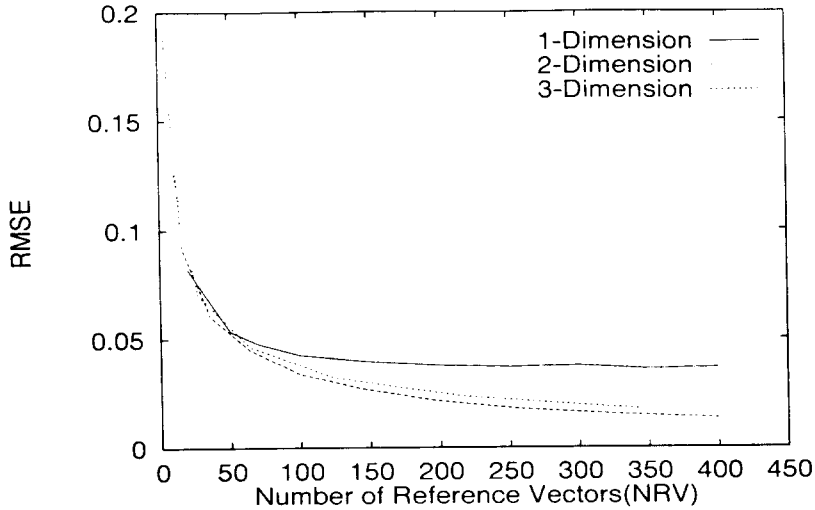


(b) 2-dimensional topology.
NRV = 23×23

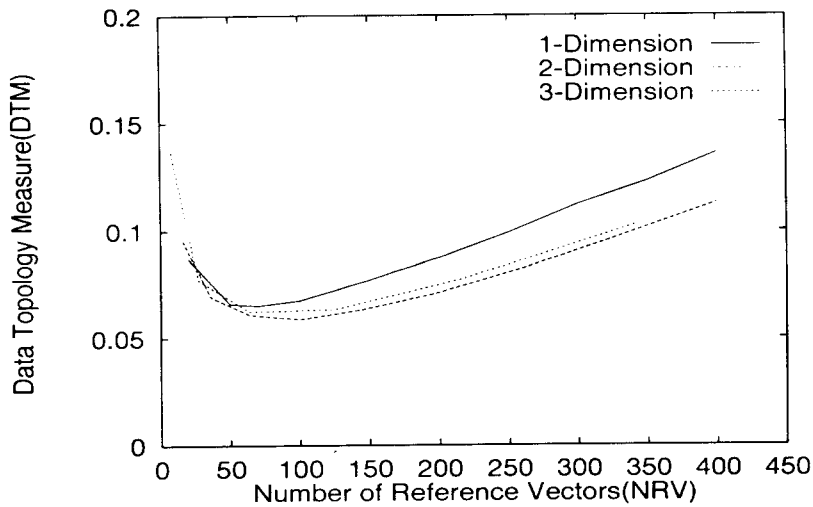


(c) 3-dimensional topology.
NRV = 8×8×8

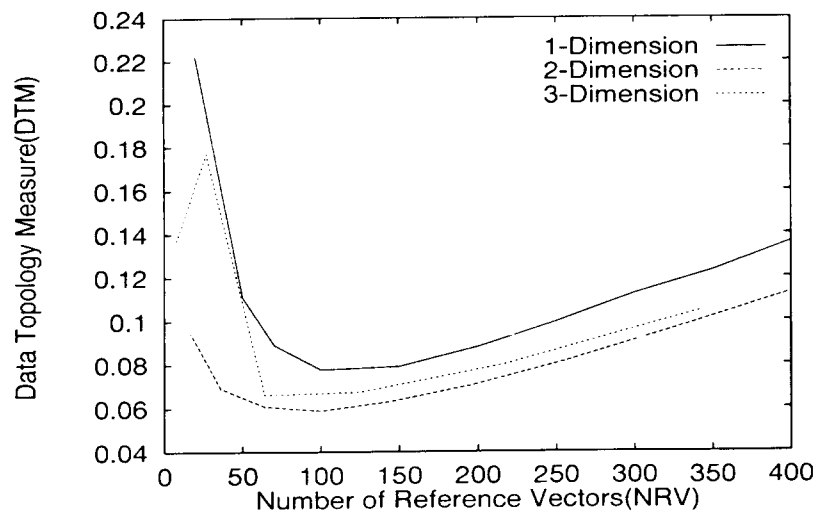
Fig. 2: Learning curves of RMSE.



(a) Relation between RMSE and NRV.

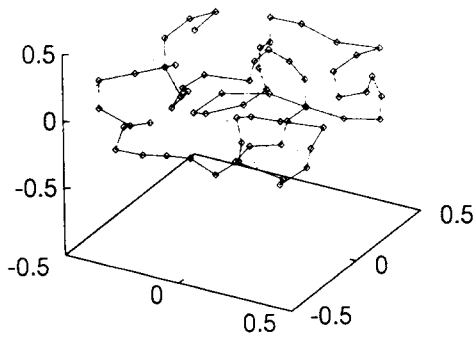


(b) Data topology measure (DTM) in terms of the number of reference vectors(NRV). $(\alpha, \beta, \gamma) = (1.2.5\times10^{-4},0)$.

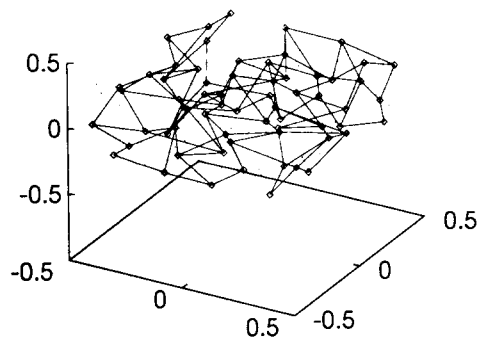


(c) Data topology measure (DTM) in terms of the number of reference vectors(NRV). $(\alpha, \beta, \gamma) = (1.2.5\times10^{-4},1)$

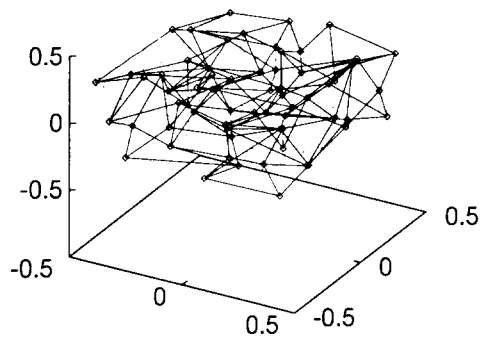
Fig. 3: Relations among three criteria RMSE, NRV and DTM



(a) 1-dimensional topology.
NRV = 64

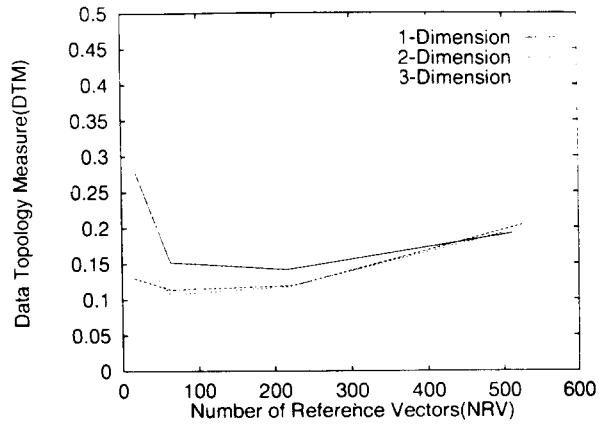


(b) 2-dimensional topology.
NRV = 8×8

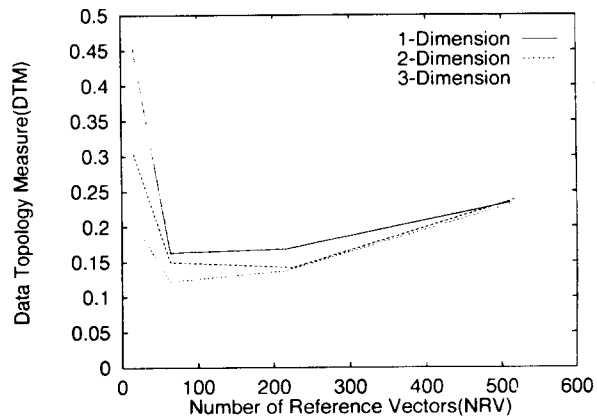


(c) 3-dimensional topology.
NRV = 4×4×4

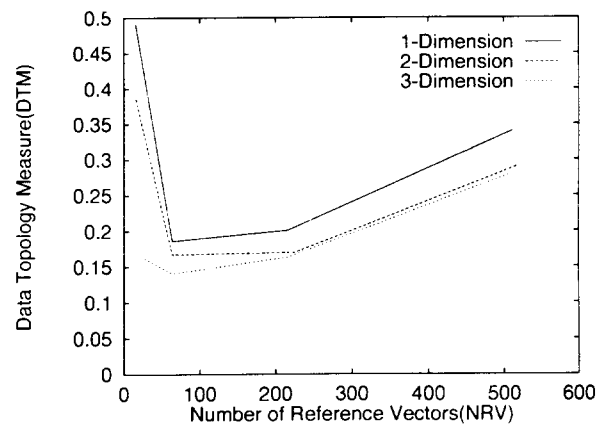
Fig. 4: Distribution of data(·) and reference vectors(◇).



(a) $d_z = 0.3$



(b) $d_z = 0.5$



(c) $d_z = 0.8$

Fig. 5: Data topology measure (DTM) in terms of the number of reference vectors(NRV)