Optimization of Distances for a Stochastic Embedding and Clustering of High-Dimensional Data

Naoto Nishikawa[1] and Shinji Doi[2]
Division of Electrical, Electronic and Information Engineering, Graduate School of Engineering, Osaka University
Yamadaoka 2–1, Suita, Osaka, 565–0871 Japan
E-mail : [1]nishikawa@is.eei.eng.osaka-u.ac.jp, [2]doi@eei.eng.osaka-u.ac.jp

Abstract: The stochastic proximity embedding (SPE) is a method of data visualization in research area of data clustering and mining. The SPE can visualize high-dimensional data by embedding them in a low-dimensional space according to a given similarity among input data. This paper extends the SPE by applying a simple iterative learning process. Without any knowledge on data, the extended SPE can automatically optimize the similarity of data and can produce low-dimensional embeddings more accurately than the original SPE.

1. Introduction

Data mining aims to search hidden knowledges, unexpected patterns and rules in large volumes of data. The process of data mining includes data selection, cleaning, clustering and prediction [1]. The stochastic proximity embedding (SPE) is a method for data clustering and visualization [2]. The SPE can embed high-dimensional data in a low-dimensional space according to a given similarity (or distance) among input data. The algorithm of SPE is so simple that the SPE can be applied to large volumes of high-dimensional data in various scientific problems. For example, the SPE can solve the protein structure determination problem which is one of optimization problems [4]. If the similarity used for SPE was appropriately defined for the input data, the low-dimensional embedding can show a clear cluster structure and characteristic feature of the input data. Thus, the definition of the similarity is very important. Usually, the similarity is empirically defined according to our knowledge on the data.

In this paper, we extend the SPE and propose a method which automatically optimizes the similarity by a simple iterative learning without any knowledge. The extended SPE can embed the input data more accurately than the original SPE. Using relatively low-dimensional artificial data and high-dimensional practical data, we demonstrate the effectiveness of our proposed method for data clustering and mining.

2. Stochastic Proximity Embedding

The SPE allocates large volumes of high-dimensional data in a low-dimensional space according to their proximities or similarities.

2.1 Algorithm

The algorithm of the SPE is as follows:

1) Assign initial D-dimensional coordinates

$$\boldsymbol{y}^i = (y_1^i, y_2^i, \ldots, y_D^i)^T \quad (i = 1, 2, \ldots, N)$$

to given input data, where N is the number of input data.

2) Set appropriate values of the neighborhood radius r_c and the learning rate λ and calculate similarities r_{ij} $(j = 1, 2, \ldots, N)$ among input data. (The smaller the similarity of data is, the larger the value of r_{ij} becomes. Therefore, we should call r_{ij} a dissimilarity. However we use the word "proximity" or "similarity" for a sake of convenience.)

3) Select two data, say i and j, at random and calculate the Euclidean distance $d_{ij}(\equiv ||\boldsymbol{y}^i - \boldsymbol{y}^j||)$.

4) If

$$(r_{ij} \leq r_c) \ \vee \ (\ (r_{ij} > r_c) \ \wedge \ (d_{ij} < r_{ij})\ ), \tag{1}$$

then update the coordinates y_d^i and y_d^j $(d = 1, 2, \ldots, D)$ as follows:

$$y_d^i \longleftarrow y_d^i + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \epsilon} (y_d^i - y_d^j) \tag{2}$$

$$y_d^j \longleftarrow y_d^j + \lambda \frac{1}{2} \frac{r_{ij} - d_{ij}}{d_{ij} + \epsilon} (y_d^j - y_d^i) \tag{3}$$

where ϵ is a small number used to avoid division by zero. For accurate embeddings of the data which are close to each other, if $r_{ij} \leq r_c$, or if $r_{ij} > r_c$ and $d_{ij} < r_{ij}$, the coordinates y_d^i and y_d^j are updated so that the distance d_{ij} matches more closely the similarity r_{ij}.

5) Repeat 3) – 4) for a prescribed number S of learning steps.

6) Reduce λ by a decrement $\Delta\lambda$.

7) Repeat 3) – 6) for some cycles C.

In general, it is desirable to set $\lambda = 2$, $\Delta\lambda = 0.02$, $C = 100$, and the value of r_c needs to be chosen enough larger than the similarities r_{ij} [2].

2.2 Error Function

To evaluate the performance of the SPE, we use the following error function E:

$$E = \sum_{i>j} f(d_{ij}, r_{ij}) / \sum_{i>j} r_{ij}^2 \tag{4}$$

$$f(d_{ij}, r_{ij}) = \begin{cases} (d_{ij} - r_{ij})^2 & \text{if } (r_{ij} \leq r_c) \vee (d_{ij} < r_{ij}) \\ 0 & \text{otherwise} \end{cases}$$

The error function is expected to be minimized in the above iterative embedding [3].

3. Optimization of Similarity

The SPE can embed high-dimensional data in a low-dimensional space according to a similarity among input data. Let us denote the input data as $\boldsymbol{x}^i = (x_1^i, x_2^i, \ldots, x_M^i)$ (M is the dimension of the input data and is supposed to be much larger than the dimension D of the embedding space). Let the similarity r_{ij} between $\boldsymbol{x}^i$ and $\boldsymbol{x}^j$ be expressed by

$$r_{ij} = \frac{1}{M}\sqrt{\sum_{m=1}^{M}\{w_m(x_m^i - x_m^j)\}^2} \tag{5}$$

where w_m is the weight for the m-th component of the M-dimensional input data.

3.1 Optimization of Similarity by the Gradient Descent Method

In this paper, we seek the optimal value of w_m so that the embedding error E could be minimized:

$$\begin{array}{ll} \min_{w_1,w_2,\ldots,w_M} & E \\ s.t. & w_m \geq 0 \ \ (m = 1, 2, ..., M) \end{array}$$

The optimization of w_m is solved by a simple iterative learning of the gradient descent method:

$$w_m \longleftarrow w_m - \eta\frac{\partial E}{\partial w_m} \tag{6}$$

where η is a learning rate. $\partial E/\partial w_m$ is determined by

$$\begin{aligned} \frac{\partial E}{\partial w_m} &= \frac{\partial}{\partial w_m}\left(\sum_{i>j}(d_{ij} - r_{ij})^2 / \sum_{i>j} r_{ij}^2\right) \\ &= -2\left[\sum_{i>j}(d_{ij} - r_{ij})(x_m^i - x_m^j)\cdot\sum_{i>j} r_{ij}^2 \right. \\ &\quad \left. + \sum_{i>j} r_{ij}(x_m^i - x_m^j)\cdot\sum_{i>j}(d_{ij} - r_{ij})^2\right] \\ &\quad \Big/ N\left(\sum_{i>j} r_{ij}^2\right)^2 \end{aligned} \tag{7}$$

Further explanation on $\partial E/\partial w_m$ is given in Section 4.

3.2 Algorithm of the Extended SPE

The algorithm of the extended SPE is as follows:

1) Assign initial D-dimensional coordinates

$$\boldsymbol{y}^i = (y_1^i, y_2^i, \ldots, y_D^i)^T \ \ (i = 1, 2, \ldots, N)$$

to given M-dimensional input data, where N is the number of input data.

2) Set appropriate values of the neighborhood radius r_c, the learning rate λ and the initial values of the weights w_m ($m = 1, 2, \ldots, M$).

3) Calculate similarities r_{ij} $(j = 1, 2, \ldots, N)$ among input data according to Eq. (5).

4) Select i and j, at random and calculate the Euclidean distance d_{ij}.

5) If $(r_{ij} \leq r_c) \vee ((r_{ij} > r_c) \wedge (d_{ij} < r_{ij}))$, then update the coordinates y_d^i and y_d^j $(d = 1, 2, \ldots, D)$ according to Eqs. (2) and (3).

6) Repeat 4) – 5) for a prescribed number S of learning steps.

7) Reduce λ by a decrement $\Delta\lambda$.

8) Update the weights w_m as follows:

$$w_m \longleftarrow w_m - \eta\frac{\partial E}{\partial w_m}$$

9) Repeat 8) for a prescribed number R of weight update.

10) Repeat 3) – 9) for some cycles C.

4. Performance comparison of the original SPE with the extended SPE

We use three kinds of input data and embed them into a low-dimensional space by three methods: the original SPE, the extended SPE and the principle component analysis (PCA) which is one of basic methods for dimensionality reduction. The first set of data is two-dimensional artificial data, the second set is five-dimensional artificial data and the third set is high-dimensional practical data in UCI Repository [6]. The values of parameters are set as $C = 100$, $S = 1000$, $\lambda = 2.1$, $\Delta\lambda = 0.2$, $r_c = 10$. By embedding these data, we compare the performance of the three methods.

Experiment 1

Firstly, we embed two-dimensional artificial data in a one-dimensional space. The input data are denoted as $\boldsymbol{x}^i (= (x_1^i, x_2^i)$ $(i = 1, 2, ..., 10))$. The first component x_1^i of the two-dimensional input data $\boldsymbol{x}^i$ is equally spaced, while the second component x_2^i is randomly spaced as shown in Fig. 1. By embedding these input data in a one-dimensional space, we examine whether the regularity of the input data is preserved by the embedding or not. In this experiment, we set the iteration number R of weight update to 1 or 10 and all initial values of w_m to 1.

Table 1 and Fig. 2 show the results of one-dimensional embeddings of two-dimensional artificial data. In Table 1, we can see the errors of the original SPE and the extended SPE. The error of the extended SPE is less than that of the original SPE, and the error of the case of $R = 10$ is much less than that of $R = 1$. Figure 2 shows the one-dimensional

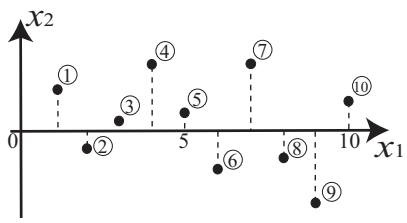

Figure 1. Input data in Experiment 1. Each point corresponds to each input datum, and the circled number labeled to each point shows the datum number i.

Table 1. Error E in Experiment 1

method	E
the original SPE	9.61×10^{-4}
the extended SPE ($R = 1$)	1.19×10^{-4}
the extended SPE ($R = 10$)	1.37×10^{-6}

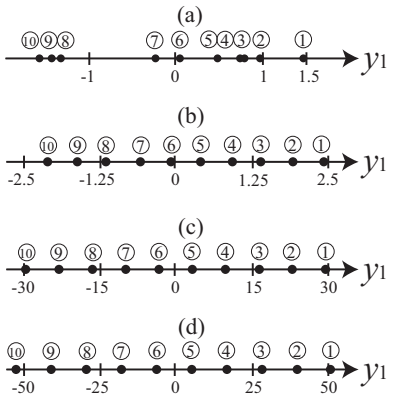

Figure 2. One-dimensional embeddings of two-dimensional artificial data generated by (a) PCA, (b) the original SPE, (c) the extended SPE ($R = 1$) and (d) the extended SPE ($R = 10$).

embeddings generated by the PCA, the original SPE and the extended SPE. We can see that the embedded data in Fig. 2(a) is unequally spaced. On the other hand, in Figs. 2(b), (c) and (d), the embedded data is equally spaced. These results show that the iterative learning of weight w_m is very effective to reproduce or extract the regularity of the input data.

Here, we explain why the iterative learning of the extended SPE is effective for the reproduction of the regularity. Let us focus on Eq. (7). As described before, the SPE updates the distance d_{ij} as $d_{ij} \simeq r_{ij}$. From Eq. (7) , we can find that if $(x_m^i - x_m^j)$ is positive, $\partial E/\partial w_m$ is negative, and vice versa. We show the detailed explanation about $(x_m^i - x_m^j)$ as follows:

(i) the case that x_m^i ($i = 1, 2, \ldots, N$) is equally spaced.
If x_m^i is equally spaced e.g., $(x_m^1, x_m^2, x_m^3, \cdots) = (1, 3, 5, \ldots)$, $(x_m^i - x_m^j)$ is positive about any pair of i and j ($i > j$). Therefore, $\partial E/\partial w_m$ is negative under this condition. This means that the weight w_m for the regular (equally spaced) component x_m^i increases.

(ii) the case that x_m^i is unequally spaced.
There is no rule which determines whether $(x_m^i - x_m^j)$ is positive or negative. Therefore, the weight w_m for the irregular (unequally spaced) component x_m^i remains near its initial values.

Experiment 2

Next, we embed five-dimensional artificial data in a two-dimensional space. The input data are denoted as $\boldsymbol{x}^i (= (x_1^i, x_2^i, x_3^i, x_4^i, x_5^i)$ $(i = 1, 2, \ldots, 100))$. The first two components x_1^i and x_2^i of input data are equally spaced according to a 10×10 orthogonal lattice shown in Fig. 3. The other components x_3^i, x_4^i and x_5^i are given randomly between 1 and 10.

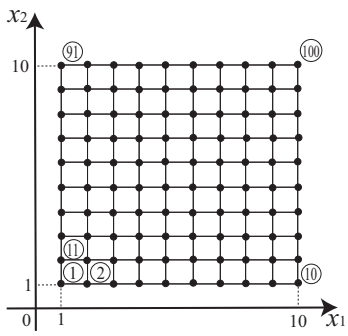

Figure 3. The first two components x_1^i and x_2^i $(i = 1, 2, ..., 100)$ of five-dimensional input data $\boldsymbol{x}^i$. Each point corresponds to each input datum, and the circled number labeled to each point shows the datum number i.

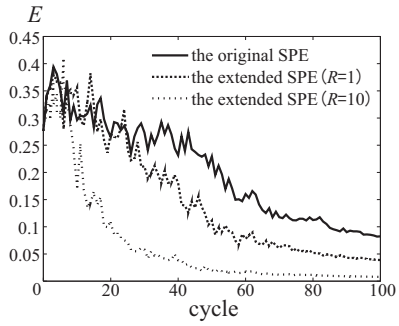

Figure 4. Error E vs. cycle number C in the embedding of five-dimensional artificial data.

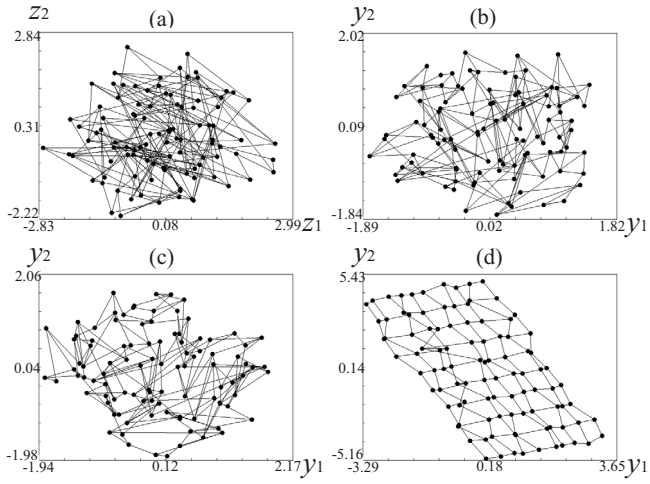

Figure 5. Two-dimensional embeddings of five-dimensional artificial data generated by (a) PCA, (b) the original SPE, (c) the extended SPE ($R = 1$) and (d) the extended SPE ($R = 10$). The data points are connected according to the reticular pattern in Fig. 3.

Other conditions are the same as Experiment 1.

Figures 4 and 5 show the results of two-dimensional embeddings of five-dimensional artificial data. In Fig. 4, we can

Table 2. Datasets used in Experiment 3 (N: the number of data, M: the dimension of each datum, K: the number of classes)

	N	M	K
Wine	178	13	3
Breast Cancer	400	9	2

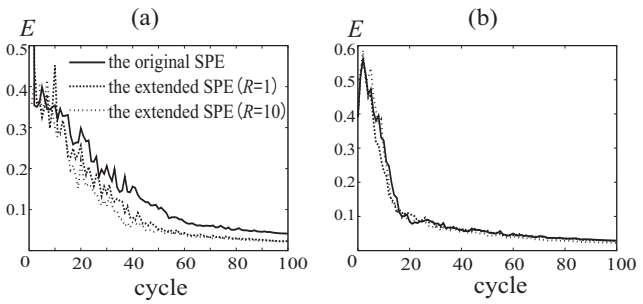

Figure 6. Error E vs. cycle number C in the embedding of high-dimensional practical data. (a) "Wine". (b) "Breast Cancer".

see how the errors change with the cycle numbers. The errors of the extended SPE are less than those of the original SPE, and the error of the case of $R = 10$ is much less than that of $R = 1$. Figure 5 shows the two-dimensional embeddings generated by the PCA, the original SPE and the extended SPE. We can find that only the extended SPE with $R = 10$ has succeeded in generating the lattice structure of the input data. These numerical results show that the optimization of the similarity of the input data by an iterative learning is very effective to embed and extract the regularity of the input data.

Experiment 3

In this experiment, we embed high-dimensional practical data in a two-dimensional space. We use two datasets in UCI repository [6]. One is named "Wine" and the other is "Breast Cancer". Table 2 summarizes the properties of these datasets: the number of data N, the dimension of each datum M, the number of classes (clusters) K. We set the number R of weight update to 1. Other conditions are the same as Experiment 1.

Figures 6 and 7 show the results of two-dimensional embeddings of the above data. In Fig. 6, we can see how the errors change with the cycle numbers. The errors of the extended SPE are less than those of the original SPE. Figure 7 shows the two-dimensional embeddings generated by the PCA, the original SPE and the extended SPE. The embedded data are clustered accurately in all embeddings. In this experiment, the input data are so simple that there is no difference between the two-dimensional representations by the extended SPE and those by other two methods.

The three experiments show that the extended SPE has at least the same or better capability than the original SPE and the PCA in embedding of high-dimensional data.

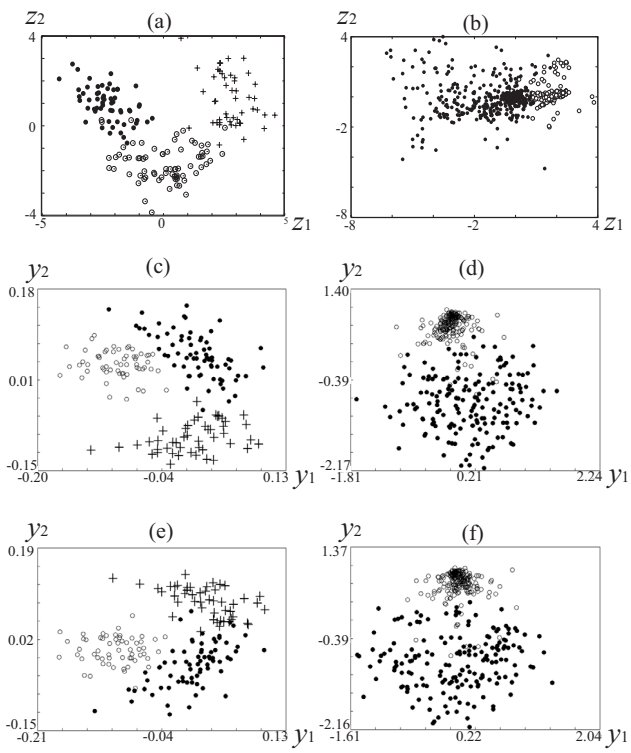

Figure 7. Two-dimensional embeddings of high-dimensional practical data. (a)(c)(e): "Wine". (b)(d)(f): "Breast Cancer". (a)(b): The PCA. (c)(d): The original SPE. (e)(f): The extended SPE. The data points which belong in a same cluster is represented by a same symbol.

5. Conclusions

The SPE is a simple and fast algorithm for producing low-dimensional representations or embedding of high-dimensional data according to a similarity or metric. In this paper, we extended the SPE to optimize the data similarity. According to numerical experiments, we have shown that the extended SPE is a better method than the original SPE and the PCA in data embedding and visualization.

References

[1] P. Adriaans and D. Zantinge, *Data Mining*, 1st ed. Addison-Wesley, England, 1996.

[2] D. K. Agrafiotis and H. Xu, "A self-organizing principle for learning nonlinear manifolds," *Proc. Network Academy of Sciences*, pp. 15869-15872, 2002.

[3] H. Kashima, S. Doi, and S. Kumagai, "Application of stochastic proximity embedding to distance geometry problems," Proc. SICE-ICASE International Joint Conference, pp. 18-21, 2006.

[4] D. K. Agrafiotis, "Stochastic proximity embedding," *J. Comp. Chem.*, vol. 24, pp. 1215-1221, 2003.

[5] University of California, Irving ftp site: http://ftp.ics.uci.edu/pub/machine-learning-databases