Journal of the Society of Korea Industrial and Systems Engineering
Vol. 34, No. 2, pp.60－67, June 2011.

새로운 적합도 함수를 사용한 비계량형 다차원 척도법에 대한 연구

이동주† · 이창용

공주대학교 산업시스템공학과

A Study on Non-Metric Multidimensional Scaling Using A New Fitness Function

Dong-Ju Lee† · Chang-Yong Lee

Dept. of Industrial and Systems Engineering, Kongju National University

Since the non-metric Multidimensional scaling (nMDS), a data visualization technique, provides with insights about engineering, economic, and scientific applications, it is widely used for analyzing large non-metric multidimensional data sets. The nMDS requires a fitness function to measure fit of the proximity data by the distances among n objects. Most commonly used fitness functions are nonlinear and have a difficulty to find a good configuration. In this paper, we propose a new fitness function, an absolute value type, and show its advantages.

Keyword : Non-Metric Multidimensional Scaling, Linear Programming, Fitness Function, Data Visualization

1. 서 론

비계량형 다차원 척도법(nMDS, non-metric multidimensional scaling)은 대량의 데이터들을 유클리드 공간(Euclidean space)에 동상사상(同相寫像, embedding)하여 개체들 간의 상호 관련성을 시각적으로 표현함으로써 개체들 간의 관련성을 연구하는 방법으로 공학, 경제학, 생물정보학(bio-informatics), 순수과학 등에 널리 적용되는 기법이다. 즉, nMDS는 N개의 개체들 간의 비유사성(dissimilarity) $\delta_{ij}(i, j = 1, 2, \cdots, N)$이 비계량(non-metric)으로 주어져 있어 유클리드 관점에서 개체들 간의 "거리"(distance) 개념을 설정하기 어려운 경우, 개체들 간의 비유사성과 유클리드 공간 상으로 사상(mapping, 寫像)된 거리간의 오차가 최소가 되도록 개체들을 유클리드 공간 상으로 동상사상하는 방법이다. 따라서 nMDS는 개체들 간의 비유사성과 거리 사이의 오차를 최소로 하는 유클리드 공간 상의 개체 위치(혹은 좌표)를 결정하는 일종의 최적화 알고리즘으로 간주할 수 있다.

기존의 nMDS 알고리즘의 목적은 비유사성과 적합된 거리간의 단조 관계의 정도를 나타내는 척도(measure)인 적합도(fitness)[8]의 값을 최소로 하는 해를 찾는 것이다. 적합도(fitness)는 실제 비유사성(dissimilarity, δ_{ij})과 적합된 거리 사이의 오차(d_{ij})로 d_{ij}와 δ_{ij}의 관계를 최적화시킬 $f(\cdot)$를 찾는 것이다.

적합도 함수들은 모두 비선형으로 다수의 국부최적해가 존재하므로 최적해를 보장하지 못할 뿐 아니라, 우수해를 찾는 데에도 많은 시간을 소모한다. 그러므로 기존

논문접수일 : 2011년 05월 16일　　게재확정일 : 2011년 06월 01일
† 교신저자 djlee@kongju.ac.kr
※ 본 논문은 2010년도 공주대학교 자체학술연구비의 지원에 의하여 연구되었음.

의 연구들은 대부분 빠른 시간 내에 우수한 해를 찾기 위해 다차원 척도법에 메타휴리스틱을 적용하여 해를 개선하는 연구들이었다.

Leung et al.[9]는 d_{ij}를 계산할 때 유클리디언 거리(Euclidean distance)대신 city block 거리를 사용하고 Simulated Annealing(SA)[10]을 이용하여 해를 구하였다. city block 거리란 직교좌표계에 일정한 좌표축의 점 위에 투영한 선분길이의 합으로 제 2.3절에서 소개한다. Zilinskas[17]는 유클리디언 거리와 city-block 거리 두 경우에 SA를 적용하여 해를 구하였다. Klock et al.[9]은 국부최적해로 빨리 수렴하는 확정적 기법(deterministic algorithm)과 SA와 같이 시간을 더 사용하더라도 국부최적해를 벗어나 좀 더 나은 국부최적해로 탐색해가는 확률적 기법(stochastic technique)의 장점을 혼합한 deterministic annealing이란 기법을 개발하고 다차원척도법에 적용하였다. Abbiw-Jackson et al.[5]은 큰 문제를 작은 문제로 나누고 각 작은 문제를 국부탐색기법으로 푸는 divide-and-conquer 알고리듬을 이용하여 다차원척도법에서 효율적으로 해를 구하는 방법을 연구하였다.

그러나 이러한 노력들에도 불구하고 목적식인 적합도함수가 복잡하여 계산량이 많고, SA 등의 메타휴리스틱을 적용하더라고 최적해를 보장하지는 못 한다. 한편, Malone et al.[13]은 다차원척도법에서 초기해(initial configuration)의 질이 좋으면 더 나은 해를 탐색해낸다는 것에 착안하여 더 나은 초기해를 찾는 기법을 제안하였다.

국내의 다차원척도법에 대한 연구동향을 살펴보면 이창용 외[4]는 비계량형 다차원 척도법 문제에 simulated annealing기법을 적용하여 우수한 해를 탐색하고자 하였다. 이 외의 국내 연구는 방법론에 대한 연구는 없었으며 주로 다양한 분야에 다차원척도법을 적용하여 여러 가지 요인들의 전체관계를 시각적으로 알아볼 수 있도록 위치화하고 이를 이용하여 분석한 연구들이었다.

김미향 외[1]는 원자력발전소 주변의 해역의 동물 플랑크톤의 군집특성을 조사하기 위해 비계량형 다차원척도법을 이용하였다. 박기용 외[3]은 외식기업에 대한 유사성 자료로 다차원척도법을 적용하여 외식기업간의 경쟁관계, 브랜드별 외식기업의 이미지 유사성 등을 살펴보았다. 노의경 외[2]는 동일한 파란색으로 염색된 면직물을 시각만으로 감각과 감성, 시지각과의 관계를 다차원척도법을 적용하여 분석하였다.

기존의 적합도 함수들은 이상치(outlier)의 영향을 많이 받으므로 본 연구에서는 적합 될 수 없는 자료의 영향을 줄이기 위해 절대값을 이용한 새로운 적합도 함수를 제안하고 이에 따른 새로운 해법을 제시하고자 한다.

이어지는 제 2장에서는 기존의 적합도 함수들과 본 논문에서 제시하는 적합도 함수에 대해 살펴보고 제 3장에서는 제시된 적합도 함수를 풀기 위한 해법을 소개한다. 제 4장에서는 간단한 예제를 통해 해법의 이해를 도우며 실험을 통해 개체수별 목적식 값의 변화와 계산시간의 변화에 대해 살펴본다. 마지막 제 5장에서는 결론과 미래의 연구방향을 제시한다.

2. 적합도 함수

개체들의 쌍별 "비유사성 값" δ_{ij}와 이 비유사성으로 구한 "형상 공간상의 개체들의 위치간 거리" d_{ij}의 차이를 적합도라 한다. 이러한 적합도를 계산하는 기존의 함수들과 본 연구에서 제안하는 적합도 함수를 살펴보면 다음과 같다.

2.1 Kruscal의 스트레스 함수

Kruscal[10]이 제시한 스트레스 함수를 살펴보면 식 (1)과 같다.

$$ST = \left[\frac{\sum_{i} \sum_{<j} (d_{ij} - f(\delta_{ij}))^2}{\sum_{i} \sum_{<j} d_{ij}^2} \right]^{1/2} \quad (1)$$

ST : 스트레스
δ_{ij} : 관찰대상 i와 j의 실제 비유사성(dissimilarity)
d_{ij} : 관찰대상 i와 j의 거리
$f(\delta_{ij})$: 프로그램에 의해 추정된 거리

위의 스트레스 함수는 비선형이며 국소최적치가 많은 울퉁불퉁한 모양인데, 최대 경사법(gradient algorithm)을 사용하여 해를 구한다. 그러나, city block 거리의 경우에서 자주 발생하는 개체가 적합된 위치가 동일한 두 점이 있는 경우에 문제가 발생할 수 있으며 그레디언트(gradient)가 존재하지 않는 경우가 발생할 수도 있다는 단점이 존재한다[6, 12]. 기존의 스트레스 함수인 식 (1)은 제곱합의 형태이기에 형상공간상에 적합 될 수 없는 자료가 있을 때, 이 자료가 스트레스 함수의 값을 좌우하게 된다는 단점이 있다. 즉, 몇 개의 이상치들이 전체 결과를 좌우할 우려가 있다.

2.2 잠재 에너지

Kruscal의 스트레스함수는 임의의 중간단계로 쓰이는 쌍별 거리인 $f(\delta_{ij})$를 도입하여 d_{ij}와 $f(\delta_{ij})$의 차이(스트레스)를 최소화하는 형상공간에서의 개체들의 위치를 구

하려 하였다. 그러나 $f(\delta_{ij})$는 유일하지 않고, $f(\delta_{ij})$가 결과에 영향을 미치므로 Taguchi et al.[16]는 purely non-metric multidimensional scaling(purely nMDS) 알고리듬을 제안하여 $f(\delta_{ij})$를 제거하고 d_{ij}와 δ_{ij}를 직접 비교하여 그 차이를 최소화하고자 하였다.

이 방법은 우선 개체들의 쌍별 거리인 d_{ij}와 δ_{ij}를 크기순으로 정렬하여 정렬된 d_{ij}와 δ_{ij}의 순위를 각각 T_n과 n으로 나타낸다. 그 후에 모든 개체들의 쌍 i, j에 대하여 그 차이의 제곱인 "잠재 에너지"(potential energy)

$$\Delta = \Sigma(T_n - n)^2 \tag{2}$$

가 최소가 되는 유클리드 공간상의 개체들의 배열(configuration)을 찾는 방법이다. 이 방법은 개체의 수가 많은 경우에도 적용할 수 있다는 장점이 있다. 비선형계획법 문제에 해당되므로 이 방법 역시 최대 경사법(gradient search)을 사용하여 해를 구한다. 특히, 이 방법은 이상치가 존재할 경우 이상치가 해에 큰 영향을 미친다.

2.3 City Block 거리를 이용한 적합도 함수

Leung and Lau[11]는 개체들 사이의 거리로 city block 거리를 사용하고 실제 비유사성과 city block 거리의 차이를 제곱하는 식 (3)과 같은 적합도 함수를 사용하였다.

$$\sum_i \sum_{j<i} (\delta_{ij} - \sum_k |x_{ik} - x_{jk}|)^2 \tag{3}$$

Leung and Lau[11]는 식 (3)의 적합도 함수에서 2차원인 경우에 한하여 해를 찾는 기법을 제안하였다. 식 (3)의 경우에는 차이의 제곱을 취하므로 형상공간 상에 적합 될 수 없는 자료가 큰 영향을 미칠 수 있다는 단점이 있다.

2.4 제시하는 적합도 함수

본 연구에서는 절대값을 사용한 적합도함수를 도입하였다. x_{ik}를 개체 i의 k번째 좌표라 하고 $d_{ij}(x) = \sum_k |x_{ik} - x_{jk}|$라 할 때 적합도 함수는 식 (4)와 같다.

$$\frac{\sum_i \sum_{j>i} |\delta_{ij} - d_{ij}(x)|}{n(n-1)/2} = \frac{\sum_i \sum_{j>i} |\delta_{ij} - \sum_k |x_{ik} - x_{jk}||}{n(n-1)/2} \tag{4}$$

식 (4)의 적합도 함수는 city block 거리를 사용하였고 d_{ij}와 δ_{ij}의 차이를 계산하기 위해서도 절대값을 취하였다. 또한, 개체 쌍별 d_{ij}와 δ_{ij}의 차이의 평균을 구하기 위해 개체 쌍 수로 나누었다.

예측(forecasting)에서 예측의 정확성(accuracy)을 나타내는 척도에는 MSE(mean squared error), RMSE(root mean absolute error) 등이 있다. MSE는 실제값과 예측값의 차이인 오차를 제곱하고 이들의 합을 구하는데 RMSE는 MSE의 제곱근을 구한 것이며, MAE(mean absolute error)는 실제값과 예측값의 차이의 절대값을 구하고 이를 더한 것이다. 일반적으로 오차의 절대값을 구한 MAE가 오차의 제곱을 한 MSE와 RMSE보다 이상치에 덜 민감한 것으로 알려져 있다[14]. 즉, 오차의 절대값으로 계산하는 경우에 이상치에 덜 민감하므로 본 연구에서는 식 (4)와 같이 오차의 절대값을 이용하는 적합도 함수를 제안하였다. 절대값을 취하므로 인해 이상치의 영향을 기존의 적합도 함수들에 비해 줄일 수 있다.

그러나 미분을 할 수 없어 최대경사법을 쓸 수 없으므로 새로운 해법이 필요하다.

3. 진입제한규칙 기반 해법

식 (4)에서 분모의 개체 쌍의 수는 결과에 영향을 미치지 않으므로 이후부터는 생략하여 설명하기로 한다. 식 (4)를 적합도 함수로 도입한 경우의 최적화 모형은 다음과 같다.

$$\text{P1} : \underset{x}{Min} \sum_i \sum_{j>i} \left| \delta_{ij} - \sum_k |x_{ik} - x_{jk}| \right| \tag{5}$$

식 (5)의 최적화모형은 다음과 같이 변형될 수 있다. $x_{ik} - x_{jk} = h_{ijk} - l_{ijk}$라 두자. 여기서, $x_{ik} \geq x_{jk}$이면 $h_{ijk} \geq 0$이고 $l_{ijk} = 0$이며, $x_{ik} \leq x_{jk}$이면 $l_{ijk} \geq 0$이고 $h_{ijk} = 0$이다.

$x_{ik} - x_{jk} = h_{ijk} - l_{ij}$이므로 식 (5)는 $\sum_i \sum_{j>i} \left| \delta_{ij} - \sum_k (h_{ikj} + l_{ikj}) \right|$라 할 수 있다.

같은 방법으로 $\delta_{ij} - \sum_k (h_{ijk} + l_{ijk}) = H_{ij} - L_{ij}$라 두면 $\left| \delta_{ij} - \sum_k (h_{ikj} + l_{ijk}) \right| = H_{ij} + L_{ij}$이다. 여기서, $\delta_{ij} - \sum_k (h_{ijk} + l_{ijk}) = \geq 0$일 때 $H_{ij} \geq 0$이고 $L_{ij} = 0$이다. 또한, $\delta_{ij} - \sum_k (h_{ijk} + l_{ijk}) \leq 0$일 때 $L_{ij} \geq 0$이고 $H_{ij} = 0$이다.

이상을 요약하여 P1을 혼합이진정수모형으로 나타내면 다음과 같다.

$$\text{P2} : \text{Min} \sum_i \sum_{j>i} (H_{ij} + L_{ij}) \tag{6}$$

s.t.

$$x_{ik}-x_{jk}=h_{ijk}-l_{ijk}, \quad \forall i,k,j>i \tag{7}$$

$$\delta_{ij}-\sum_{k}(h_{ijk}+l_{ijk})=H_{ij}-L_{ij}, \quad \forall i,k,j>i \tag{8}$$

$$h_{ijk}-Ma_{ijk}\le 0, \quad \forall i,k,j>i \tag{9}$$

$$l_{ijk}-M(1-a_{ijk})\le 0, \quad \forall i,k,j>i \tag{10}$$

$$H_{ij}-Mb_{ij}\le 0, \quad \forall i,k,j>i \tag{11}$$

$$L_{ij}-M(1-b_{ij})\le 0, \quad \forall i,k,j>i \tag{12}$$

$$\forall x_{ik}, l_{ijk}, h_{ijk}, L_{ij}, H_{ij}\ge 0 \tag{13}$$

$$\forall a_{ijk}, b_{ij}\in\{0,1\} \tag{14}$$

여기서 M은 임의의 큰 수이다.

식 (6)는 목적식으로 선형이며 식 (7)과 식 (8)은 절대값을 구하는 목적식을 선형화하기 위해 사용되었다. 즉, 식 (7)의 경우 $|x_{ij}-x_{jk}|$를 선형의 제약식으로 나타내기 위해 사용되었다. $x_{ij}-x_{jk}$가 양의 값을 가지는 경우 h_{ijk}가 비음의 값을 가지고 $l_{ijk}=0$이며, $x_{ij}-x_{jk}$가 음의 값을 가지는 경우 l_{ijk},가 비음의 값을 가지고 $h_{ijk}=0$이다. 그러나 식 (7)만 있는 경우 h_{ijk}, l_{ijk} 둘 다 양의 값을 가지는 경우가 발생하므로 식 (9)와 식 (10)를 추가하여 h_{ijk}, l_{ijk} 둘 중 하나는 비음을 나머지는 0을 가지도록 하였다.

이와 마찬가지로, $\sum_{i}\sum_{j>i}\left|\delta_{ij}-\sum_{k}(h_{ikj}+l_{ikj})\right|$을 선형화하기 위해 식 (8)이 사용되었으며, 식 (11)과 식 (12)은 L_{ij}, H_{ij}의 둘 중 하나는 비음의 값을 가지고 나머지는 0을 가지게끔 하는 제약식이다. 식 (13)은 x_{ijk}를 제외한 각 변수가 비음인 것을 나타내며 식 (14)는 a_{ijk}, b_{ij}가 이진정수임을 나타낸다. 그러므로 이 수학 모형은 이진정수를 가진 혼합정수계획모형이다. h_{ijk}와 l_{ijk}, H_{ij}와 L_{ij}는 각각 둘 중의 하나만 값을 가질 수 있는데 이는 이진정수를 도입하면 모형화 할 수 있다.

수학 모형 P2의 혼합이진정수모형은 특수한 형태로 좀 더 쉬운 수학모형으로 변환이 가능하이다. 즉, h_{ijk}와 l_{ijk}, H_{ij}와 L_{ij}는 각각 둘 중 하나가 값을 가지면 나머지 하나는 0이 되도록 수학모형 P2에서는 이진정수 a_{ijk}, b_{ij}를 도입하였다. 그런데 다음과 같은 비선형 제약식을 도입한다면 이진정수 a_{ijk}, b_{ij}는 필요가 없게 된다.

$$\sum_{i}\sum_{j>i}\sum_{k}h_{ijk}\times l_{ijk}+\sum_{i}\sum_{j>i}H_{ij}\times L_{ij}=0 \tag{15}$$

식 (15)는 h_{ijk}와 l_{ijk}, H_{ij}와 L_{ij}중 각각 하나만 값을 가지도록 한다. 또한 식 (9)~식 (12), 식 (14)는 식 (15)와 같이 표현될 수 있다. 그러므로 P2는 다음과 같은 수학모형으로 표현될 수 있다.

P3 : 목적식 : 식 (6)
s.t. 식 (7), 식 (8), 식 (13), 식 (15)

식 (15)의 제약식은 선형계획문제(linear programming)를 푸는데 널리 알려진 Simplex 방법에 진입제한규칙(restricted entry rule)을 적용하여 표현하면 P3의 해를 구할 수 있다. 진입제한규칙이란 Simplex 방법에서 진입변수를 선택할 때 비기저변수중 곱해서 0이 되어야 하는 h_{ijk}와 l_{ijk}, 혹은 L_{ij}와 H_{ij}의 각 쌍 중 하나가 이미 기저변수라면 진입변수선택에서 제외하는 방법이다. 즉, 곱해서 0이 되어야 하는 두 변수가 동시에 기저변수가 될 수 없으므로 동시에 값을 가지지 못하도록 하는 규칙이다[7].

그러므로 제 2.4절에서 제시한 적합도 함수에 대해 진입제한규칙을 이용한 선형계획법으로 국부최적해를 찾을 수 있다.

4. 예제 및 실험

개체들 간의 비유사성이 1-2는 1, 1-3은 3, 2-3은 2로 주어져 있다고 하자. 이때의 수학모형은 다음과 같다.

$$\begin{aligned}
Min\ & H_{12}+L_{12}+H_{13}+L_{13}+H_{23}+L_{23}\\
s.t.\ & -x_{11}+x_{21}+h_{121}-l_{121}=0\\
& -x_{12}+x_{22}+h_{122}-l_{122}=0\\
& -x_{11}+x_{31}+h_{131}-l_{131}=0\\
& -x_{12}+x_{32}+h_{132}-l_{132}=0\\
& -x_{21}+x_{31}+h_{231}-l_{231}=0\\
& -x_{22}+x_{32}+h_{232}-l_{232}=0\\
& 1-h_{121}-l_{121}-h_{122}-l_{122}-H_{12}+L_{12}=0\\
& 1-h_{131}-l_{131}-h_{132}-l_{132}-H_{13}+L_{13}=0\\
& 1-h_{231}-l_{231}-h_{232}-l_{232}-H_{23}+L_{23}=0\\
& h_{121}l_{121}+h_{122}l_{122}+h_{131}l_{131}+h_{132}l_{132}\\
& +h_{231}l_{231}+h_{232}l_{232}+H_{12}L_{12}+H_{13}L_{13}\\
& +H_{23}L_{23}=0\\
& \forall x_{ik}, l_{ijk}, h_{ijk}, L_{ij}, H_{ij}\ge 0
\end{aligned}$$

위 모형의 초기실현 가능해를 찾아야 하는데 이는 보조 목적식(auxiliary objective function)을 이용한 방법을 적용하였다[15]. 즉, 실현 가능해는 모든 제약식을 만족해야 하는데 이런 제약을 만족하도록 보조 목적식을 추가하고 이 보조 목적식의 값이 0이 되도록 하여 초기 실현 가능해를 찾는 방법이다. 이를 살펴보면 다음과 같다.

$$z_1 = -x_{11} + x_{21} + h_{121} - l_{121}$$
$$z_2 = -x_{12} + x_{22} + h_{122} - l_{122}$$
$$\vdots$$
$$z_9 = 1 - h_{231} - l_{231} - h_{232} - l_{232} - H_{23} + L_{23}$$

보조 목적식은

$$z' = -z_1 - z_2 - \cdots - z_9$$
$$= -6 + 2x_{11} + 2x_{21} + \cdots + H_{23} - L_{23}$$

이 때 진입제한 규칙을 적용하므로 마지막 제약식인 비선형제약식은 고려치 않았다.

<표 1>에 최초 심플렉스 모형이 있다. 2행인 z는 목적식이며 3행부터 11행까지는 제약식이며 마지막 12행은 보조 목적식이다. <표 2>는 최종테이블 직전의 테이블이다. L_{12}이 진입변수로 z8이 탈락변수로 결정되었다. 여기서 볼드체로 표시된 x_{12}, x_{21}, x_{22}, x_{32}, h_{121}, l_{131}, h_{132}, l_{231}은 기저변수이며 나머지 변수는 비기저변수이다. h_{131}, l_{132}는 보조 목적식의 계수가 2로 크지만 l_{131}, h_{132}가 각각 기저변수이므로 진입제한규칙에 의해 진입변수가 될 수 없다. L_{12}, H_{13}, L_{23}은 모두 보조 목적식의 계수가 1로 동일한데 임의로 L_{12}가 진입변수로 결정되었다.

<표 3>은 보조 목적식의 값(z')이 0이므로 초기실현 가능해를 구하였다. 여기서 목적식(z)를 최소화하도록 피봇을 하여야 하지만 계수가 양수인 h_{131}, l_{132}는 l_{131}, h_{132}가 각각 기저변수이므로 진입변수가 될 수 없고 이외의 변수들은 모두 0 혹은 음수이므로 진입변수가 될 수 없다. 그러므로 <표 3>이 최종테이블인 것을 알 수 있다. 이때의 해를 정리하면 다음과 같다.

$$x_{12} = x_{22} = 0,\ x_{21} = 1, x_{32} = 3, h_{132} = L_{12} = 0,$$
$$h_{121} = 1, l_{131} = 3, l_{231} = 2$$

<표 1> 초기해

	RHS	x_{11}	x_{12}	x_{21}	x_{22}	x_{31}	x_{32}	h_{121}	l_{121}	h_{122}	l_{122}	h_{131}	l_{131}	h_{132}	l_{132}	h_{231}	l_{231}	h_{232}	l_{232}	H_{12}	L_{12}	H_{13}	L_{13}	H_{23}	L_{23}
z	0	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1	-1	-1
z1	0	**-1**	0	1	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z2	0	**0**	-1	0	1	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
z3	**0**	-1	**0**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	1	-1	0	0	0	0	0	0	0	0	0	0	**0**	**0**
z4	0	**0**	-1	0	0	0	1	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0	0	0
z5	0	**0**	0	-1	0	1	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0	0	0
z6	0	**0**	0	0	-1	0	1	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0
z7	1	**0**	0	0	0	0	0	-1	-1	-1	-1	0	0	0	0	0	0	0	0	1	-1	0	0	0	0
z8	3	**0**	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1	0	0	0	0	0	0	1	-1	0	0
z9	2	**0**	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1	0	0	0	0	1	-1
z'	-6	**2**	2	0	0	-2	-2	0	2	0	2	0	2	0	2	0	2	0	2	1	-1	1	-1	1	-1

<표 2> 최종 테이블 직전

	RHS	x_{11}	x_{12}	x_{21}	x_{22}	x_{31}	x_{32}	h_{121}	l_{121}	h_{122}	l_{122}	h_{131}	l_{131}	h_{132}	l_{132}	h_{231}	l_{231}	h_{232}	l_{232}	H_{12}	L_{12}	H_{13}	L_{13}	H_{23}	L_{23}
z	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	**-1**	-1	-1	-1	-1
h121	1	0	0	0	0	0	0	-1	-1	-1	-1	0	0	0	0	0	0	1	0	-1	**1**	0	0	0	0
h132	0	0	0	0	0	0	0	0	0	1	-1	0	0	-1	1	0	0	-1	-1	0	**0**	0	0	0	0
l131	3	0	0	0	0	0	0	-2	0	-1	-1	1	-1	0	0	-2	0	1	-1	-1	**1**	0	0	-1	1
x12	0	0	-1	0	0	0	1	0	0	1	-1	0	0	0	0	0	0	0	-1	0	**0**	0	0	0	0
x21	1	1	0	-1	0	0	0	-2	0	-1	-1	0	0	0	0	0	0	1	0	-1	**1**	0	0	0	0
x22	0	0	0	0	-1	0	1	0	0	0	0	0	0	0	0	0	0	-1	-1	0	**0**	0	0	0	0
x32	3	1	0	0	0	-1	0	-2	0	-1	-1	0	0	0	0	-2	0	0	-1	-1	**1**	0	0	-1	1
z8	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**2**	**0**	**0**	**2**	**-2**	**0**	**0**	**-2**	**2**	**-1**	**-1**	**2**	**1**	-1	**-1**	**1**	**1**	**-1**
l231	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1	0	**0**	0	0	-1	1
z'	0	0	0	0	0	0	0	-2	0	0	-2	2	0	0	2	-2	0	0	-2	-1	**1**	1	-1	-1	1

<표 3> 최종 테이블

	RHS	x_{11}	x_{12}	x_{21}	x_{22}	x_{31}	x_{32}	h_{121}	l_{121}	h_{122}	l_{122}	h_{131}	l_{131}	h_{132}	l_{132}	h_{231}	l_{231}	h_{232}	l_{232}	H_{12}	L_{12}	H_{13}	L_{13}	H_{23}	L_{23}
z	0	0	0	0	0	0	0	-2	0	0	-2	2	0	0	2	-2	0	0	-2	-2	0	0	-2	-2	0
h_{121}	1	0	0	0	0	0	0	1	-1	-1	1	-2	0	0	-2	2	0	0	2	0	0	-1	1	1	-1
h_{132}	0	0	0	0	0	0	0	0	0	1	-1	0	0	-1	1	0	0	1	-1	0	0	0	0	0	0
l_{131}	3	0	0	0	0	0	0	0	0	-1	1	-1	-1	0	-2	0	0	-1	1	0	0	-1	1	0	0
x_{12}	0	0	-1	0	0	0	1	0	0	1	-1	0	0	0	0	0	0	1	-1	0	0	0	0	0	0
x_{21}	1	1	0	-1	0	0	0	0	0	-1	1	-2	0	0	-2	2	0	0	2	0	0	-1	1	1	-1
x_{22}	0	0	0	0	-1	0	1	0	0	0	0	0	0	0	0	0	0	1	-1	0	0	0	0	0	0
x_{32}	3	1	0	0	0	-1	0	0	0	-1	1	-2	0	0	-2	0	0	-1	1	0	0	-1	1	0	0
L_{12}	0	0	0	0	0	0	0	2	0	0	2	-2	0	0	-2	2	0	0	2	1	-1	-1	1	1	-1
l_{231}	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-1	-1	-1	-1	0	0	0	0	-1	1
z'	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

나머지 변수들의 값은 모두 0이다. 즉, 개체 1, 개체 2, 개체 3의 좌표는 (0, 0), (1, 0), (3, 0)이며 이때의 목적식의 값은 0이다.

본 연구에서 사용한 벤치마킹 데이터는 복잡 네트워크(complex networks)[18]라 불리는 네트워크를 통해 생성할 수 있는 생물학적 네트워크 중, E-coli와 neuron 그리고 임의 네트워크(random network)를 적용하여 생성한 비유사도 데이터이다. 복잡 네트워크란 연구의 대상이 되는 시스템을 그 시스템을 구성하는 개체와 개체 사이의 상호 연관성을 연결(link, 혹은 edge)로 표현하고, 개체와 연결의 집합으로 구성된 네트워크를 여러 분석법을 사용하여 그 특성을 연구하는 분야이다. 복잡 네트워크는 개체와 연결의 개수를 임의로 조절할 수 있다는 장점이 있기 때문에 개체와 연결의 개수에 따라 개체들 간의 다양한 비유사도 δ_{ij}를 생성할 수 있다. E-coli, neuron, random 등 3개의 데이터 셋을 이용하여 개체 수가 5, 10, 15, 20, 25, 30인 경우에 대해 목적식의 값과 계산 시간(CPU time, 단위 : 초)을 구하였다. 이러한 데이터를 사용하여 목적식을 계산한 결과는 <표 4>, <표 5>, <표 6>와 같다.

<표 4> E-coli의 개체수별 결과

개체수	목적식 값	시간(초)
5	108.38	0.17
10	116.04	0.17
15	118.07	2.18
20	103.10	14.32
25	90.78	51.82
30	81.56	192.53

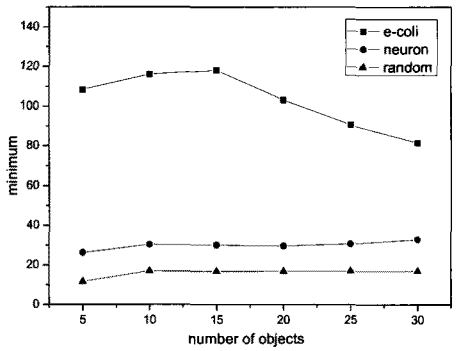
<표 5> Neuron의 개체수별 결과

개체수	목적식 값	시간(초)
5	26.32	0.57
10	30.32	0.21
15	30.14	2.11
20	29.64	13.70
25	30.75	59.41
30	32.95	193.94

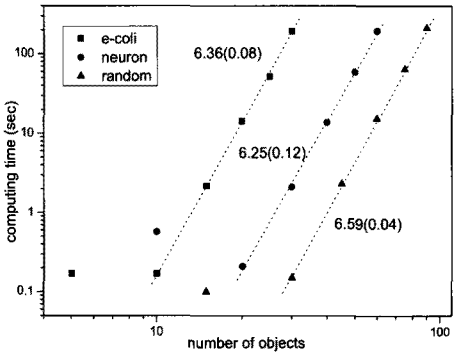
<표 6> Random의 개체수별 결과

개체수	목적식 값	시간(초)
5	11.62	0.10
10	16.88	0.15
15	16.69	2.31
20	16.83	15.35
25	16.86	64.33
30	16.92	211.41

E-coli의 경우를 제외하면 개체쌍 당 목적식 값은 일정한 것으로 보인다. 개체쌍 당 계산 시간은 개체의 수가 많아질수록 급격히 증가하는 것으로 보인다. 개체의 수와 계산 시간을 자세히 살펴보기 위해 개체 쌍의 수와 계산 시간을 각각 로그를 취한 후 회귀분석을 한 결과는 <그림 2>와 같다. 기울기를 보면, 개체 수에 대하여 6.36, 6.25, 6.59으로 약 6에 가까우므로 "개체쌍의 수"[3]에 비례하여 계산시간이 증가하는 것으로 보인다.



<그림 1> 목적식 값



<그림 2> 개체 수에 따른 계산 시간(초). 각 데이터를 구별하기 위하여 neuron과 random 데이터에 대하여 개체 수를 각각 2배와 3배로 수평 방향으로 이동시켰다.

목적식 값과 계산 시간은 <그림 1>, <그림 2>와 같다.

5. 결론 및 미래연구과제

비계량 다차원척도법(nMDS)은 개체(object)들의 비유사성(dissimilarity)에 대한 비계량 쌍별 자료만 존재할 때 개체들 간의 전체관계를 저차원 공간상에 나타내어 시각화하는 기법이다. 비유사성과 적합된 거리간의 단조 관계의 정도를 나타내는 척도(measure)를 적합도(fitness)함수라 부른다. 기존의 적합도 함수들은 이상치(outlier)의 영향을 많이 받으므로 본 연구에서는 이러한 적합 될 수 없는 자료의 영향을 줄이기 위해 절대값을 이용한 새로운 적합도 함수를 제안하였다.

제안된 적합도 함수는 기존의 최대경사법으로 해를 구할 수 없어 진입제한규칙을 적용한 선형계획법 기법을 개발하였다. 이해를 돕기 위해 간단한 예제를 이용하여 기법을 살펴보고 개체쌍의 수별 목적식 값의 변화와 계산시간의 변화를 실험을 통해 살펴보았다. 개체수가 늘어남에 따라 개체쌍 당 목적식 값의 변화는 대체로 일정하며 계산시간의 경우에는 비선형적으로 급격히 증가함을 알 수 있었다.

본 연구에서 제시한 기법은 국부최적해로 수렴하므로 전체최적해(global optimum) 혹은 근사 최적해로 수렴하는 기법에 대한 연구가 필요하다.

참고문헌

[1] 김미향, 문형태, 신상희, 손명백, 변주영, 최휴창, 손민호; "월성원자력발전소 주변 해역 동물플랑크톤의 군집 특성", 환경생물학회지, 29 : 40-48, 2010.

[2] 노의경, 유효선; "면직물의 구성특성이 시지각에 미치는 영향과 이미지 스케일에 관한 연구", 한국의류학회지, 28 : 1142-1152, 2004.

[3] 박기용, 안성식, 정기용; "다차원척도법을 이용한 외식기업 경쟁요인 비교분석에 관한 연구", 외식경영학회, 9 : 93-114, 2006.

[4] 이창용, 이동주; "담금질을 사용한 비계량 다차원 척도법", 정보과학회 논문지 : 컴퓨팅의 실제 및 레터, 16 : 648-653, 2010.

[5] Abbiw-Jackson, R., Golden, B., Raghavan, S., and Wasil, E.; "A divide-and-conquer local search heuristic for data visualization," *Computers and Operations Research*, 33 : 3070-3087, 2006.

[6] Groenen, P. J. F. and Heiser, W.; "The Tunneling Method For Global Optimization In Multidimensional Scaling," *Psychometrika*, 61 : 529-550, 1996.

[7] Hiller, F. S. and Lieberman, G. J.; *Introduction to Operations Research*, 8th Ed, McGraw-Hill, 2004.

[8] Kirkpatrick, S., Gelatt, C. D., and Vecchi. M. P.; "Optimization by simulated annealing," *Science*, 220 : 671-680, 1983.

[9] Klock, H. and Buhmann, J. M.; "Data Visualization by multidimensional scaling : a deterministic annealing approach," *Pattern Recognition*, 33 : 651-669, 2000.

[10] Kruskal, J. B.; "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, 29 : 115-129, 1964.

[11] Leung, P. L. and Lau, K.; "Estimating the city-block two-dimensional scaling model with simulated annealing," *European Journal of Operaional Research*, 158 : 518-524, 2004.

[12] Leeuw, J. D.; "Differentiability of Kruskal's Stress At A Local Minimum," *Psychometrika*, 49 : 111-113, 1984.

[13] Malone, S. W., Tarazaga, P., and Trosset, M. W.; "Better initial configurations for metric multidimensional scaling,"

Computational Statistics and Data Analysis, 41 : 143-156, 2002.

[14] Muller, U. A., Dacorogna, M. M., Dave., R. D., Olsen, R. B., Pictet, O. V., and Weizsacker, J. E.; "Volatilities of different time resolutions-Analyzing the dynamics of market components," *Journ. of Empirical Finance*, 4 : 213-239, 1997.

[15] Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling W. T.; Numerical Recipes in C(3rd Editiion), Cambridge Univ Pr. N.Y., 2003.

[16] Taguchi, Y. h. and Oono, Y.; "Relational patterns of gene expression via non-metric multidimensional scaling analysis," *Bioinformatics*, 21 : 730-740, 2005.

[17] Zilinskas A., Zilinskas, J.; "On Multidimensional Scaling with Euclidean and City Block Metrics," *Okio Technologinis Ir Ekonominis Vystymas*, 69-75, 2006.

[18] Newman, M.; "The structure and function of complex networks," *SIAM Rev*, 45 : 167-256, 2003.