담금질을 사용한 비계량 다차원 척도법
(Non-Metric Multidimensional Scaling using Simulated Annealing)

이 창 용 † 이 동 주 ††
(Chang-Yong Lee) (Dongju Lee)

요 약 비계량 다차원 척도법은 개체들 간의 비유사성이 비계량으로 주어져 개체들 간의 거리 개념을 설정하기 어려운 경우에 개체들을 유클리드 공간 상으로 사상하여 개체 간의 관련성을 연구하는 방법으로 지역 최적치가 많은 최적화 문제로 간주할 수 있다. 비계량 다차원 척도법을 위한 기존의 알고리즘은 최대 경사법을 사용함으로 일단 지역 최적치에 도달하면 더 이상 향상된 해를 찾기 어렵다는 단점이 있다. 이러한 단점을 해결하기 위하여 본 논문에서는 담금질 방법을 비계량 다차원 척도법에 접목하여 지역 최적치에 빠지지 않고 전역 최적치를 효율적으로 찾을 수 있는 새로운 비계량 다차원 척도법 알고리즘을 제안하였다. 제안한 알고리즘을 벤치마킹 문제에 적용하고 실험을 통하여 기존 알고리즘과 비교 분석한 결과, 제안한 알고리즘은 기존 알고리즘 대비 0.7%에서 3.2%의 향상률을 보였다. 또한 통계적 가설 검정을 통하여 제안한 알고리즘의 우수성을 입증하였다.

키워드 : 비계량 다차원 척도법, 담금질, 순수 nMDS, 전역 최적치, 임의 네트워크

Abstract The non-metric multidimensional scaling (nMDS) is a method for analyzing the relation among objects by mapping them onto the Euclidean space. The nMDS is useful when it is difficult to use the concept of distance between pairs of objects due to non-metric dissimilarities between objects. The nMDS can be regarded as an optimization problem in which there are many local optima. Since the conventional nMDS algorithm utilizes the steepest descent method, it has a drawback in that the method can hardly find a better solution once it falls into a local optimum. To remedy this problem, in this paper, we applied the simulated annealing to the nMDS and proposed a new optimization algorithm which could search for a global optimum more effectively. We examined the algorithm using benchmarking problems and found that improvement rate of the proposed algorithm against the conventional algorithm ranged from 0.7% to 3.2%. In addition, the statistical hypothesis test also showed that the proposed algorithm outperformed the conventional one.

Key words : Non-metric multidimensional scaling, simulated annealing, pure nMDS, global optimum, random network

· 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(과제번호: 2009-0073427)

† 정 회 원 : 공주대학교 산업시스템공학과 교수
clee@kongju.ac.kr
(Corresponding author임)
†† 비 회 원 : 공주대학교 산업시스템공학과 교수
djlee@kongju.ac.kr

논문접수 : 2010년 1월 18일
심사완료 : 2010년 3월 29일

정보과학회논문지: 컴퓨팅의 실제 및 레터 제16권 제6호(2010.6)

1. 서 론

비계량 다차원 척도법(nMDS, Non-metric Multi-Dimensional Scaling)[1]은 N개 개체(object)를 저차원(보통 2차원) 유클리드 공간(Euclidean Space)에 동상사상(同相寫像, embedding)하여 개체들 간의 상호 관련성을 시각적으로 표현함으로써 개체들 간의 관련성을 연구하는 방법으로 공학, 생물정보학(bioinformatics), 순수과학, 심리학 등에 널리 적용되는 방법론이다. 개체들 간의 비유사성(dissimilarity)이 계량적(metric)으로 주어진 경우에 사용하는 다차원 척도법(MDS, Multi-Dimensional Scaling)[1]과 달리, nMDS는 N개 개체들 간의 비유사성 δ_{ij} $(i,j=1,2,\cdots,N)$가 비계량(non-

metric)으로 주어져 유클리드 공간 관점에서 개체들 간의 거리(distance) 개념을 설정하기 어려운 경우에 사용하는 방법이다. 일반적으로 비유사성은 $\delta_{ij}=\delta_{ji}$ 를 만족하고 $\delta_{ii}=0$으로 정의됨으로 N개 개체들 간의 비유사성은 대각원소가 0인 $N\times N$ 대칭 행렬을 이룬다.

방법론적 측면에서 볼 때 nMDS는 개체들을 유클리드 공간에 적합하게 배치한 후 계산할 수 있는 개체간의 거리와 주어진 비유사성 사이에 최적의 단조 관계(monotone relationship)가 형성되도록 유클리드 공간상에 개체들의 배치(configuration)를 찾는 최적화 알고리즘으로 간주할 수 있다. 여기서 단조 관계란 비유사성과 거리 각각을 크기 순서로 정렬한 후 각 개체 쌍에 대한 비유사성의 정렬된 순서와 거리의 정렬된 순서 간의 차이를 말한다. 이를 위하여 nMDS 알고리즘에서는 스트레스(stress)[2,3] 혹은 포텐셜 함수(potential function)[4] 등의 적합도 함수(fitness function)를 사용한다. 이 적합도 함수는 개체들의 비유사성과 거리간의 단조 관계의 정도를 나타내는 측도(measure)이며, 따라서 이러한 적합도 함수의 최적치를 찾는 것이 nMDS 알고리즘의 목표이다. nMDS의 적합도 함수는 일반적으로 다변량(multivariate)이고 비선형(non-linear)으로 지역 최적치가 많은 울퉁불퉁한 모양(rugged landscape)을 띄고 있는데, 이러한 적합도 함수의 전역 최적치를 찾는 것은 시간 복잡도(time complexity)가 $O(N!)$인 NP-complete 혹은 NP-hard 문제[5]에 속한다.

nMDS는 PCA(Principal Component Analysis)[1]나 비유사성이 계량적인 양일 때 사용할 수 있는 MDS 등의 선형 방법론들을 적용할 수 없는 경우에 사용할 수 있는 비선형 다변량 데이터 해석 방법론으로 데이터 마이닝(data mining) 분야에서 그 잠재력을 인정받은 방법론이다. 특히 nMDS는 유클리드 공간 상에 개체를 위치시킴으로 시각화가 가능하다는 장점이 있기 때문에 수치적 해석에 상대적으로 익숙하지 않는 생명과학 혹은 의학 분야의 연구자들에게 중요한 분석 방법으로 자리 잡고 있다. 최근 들어 생물정보학 분야에서 실험적으로 생산된 다양한 유형의 다변량 데이터들을 nMDS로 분석하는 시도가 늘고 있고, 특히 종양(tumor) 형태의 분류 및 마이크로 어레이(microarray) 데이터 등을 nMDS에 적용하여 유전자 발현에 대한 상대적 패턴 등 상호 관련성을 밝히는 연구가 진행되고 있다[4,6].

2. 관련 연구 동향

초기 nMDS 알고리즘은 R. Shepard에 의하여 그 아이디어가 정립[7]되었고, J. Kruskal에 의해 1964년 최초의 nMDS 알고리즘이 개발되었다[2,3]. J. Kruskal은 개체들 사이의 비계량적인 비유사성 δ_{ij}가 주어진 경우, 최소자승 단조 회귀법(least square monotone regression)을 사용하여 개체들의 유클리드 공간 상의 위치와 그에 따른 거리 d_{ij}를 구하는 방법을 제안하였다. 이 방법은 비유사성 δ_{ij}에 단조 변환(monotone transformation)을 통하여 계량적 성질을 만족하는 $\widehat{d_{i,j}}$를 구하고, d_{ij}와 $\widehat{d_{i,j}}$의 차이가 최소화되도록 개체들의 유클리드 공간 상의 위치를 결정하는 방법이다. J. Kruskal은 d_{ij}와 $\widehat{d_{i,j}}$의 차이를 표현하기 위해 스트레스라는 개념을 도입하였다. 이 방법은 단조 변환을 통해 $\widehat{d_{i,j}}$를 구해야 하는 단점이 있으며, 또한 스트레스 함수는 비선형으로 다수의 지역최적치가 존재하므로 단조 회귀법과 같은 최대 경사법(steepest descent)은 전역 최적치를 보장하지 못할 뿐 아니라, 우수 해를 찾는 데에도 많은 시간을 소모한다.

J. Kruskal이 제안한 방법 외에 메타 휴리스틱(meta-heuristic)을 적용하여 최적치를 개선하는 연구들이 있다. P. Leung[8] 등은 유클리드 공간상의 거리 d_{ij} 대신 개체간의 거리를 절대값으로 표현한 절대값 거리(city block distance)를 사용하고 담금질(SA, Simulated Annealing)을 적용하여 최적치를 구하였으며, A. Zilinskas[9] 등은 유클리드 공간상의 거리 d_{ij}와 city block 거리 두 경우에 대하여 담금질을 적용하여 최적해를 구하였고, S. Malone[10] 등은 nMDS에서 초기해의 질이 좋으면 더 나은 해를 탐색해낸다는 것에 착안하여 더 나은 초기 해를 찾는 기법을 제안하였다. 국내의 nMDS에 대한 연구는 다양한 분야에 nMDS을 적용한 연구가 주를 이루고 있다. 박기용[11] 등은 외식기업에 대한 유사성 자료로 nMDS을 적용하여 외식기업간의 경쟁관계, 브랜드별 외식기업의 이미지 유사성 등을 살펴보았고, 김철수[12] 등은 공동주택에서 주거유형별 만족도를 nMDS으로 위치화하고 분석하였다.

그러나 이러한 방법들은 J. Kruskal 알고리즘의 변형으로 기본적으로 최적치를 찾기 위한 스트레스 함수에 대한 계산량이 많은 단점이 있다. 또한 위에서 언급한 nMDS 알고리즘은 소규모 개체에 적용하기 위하여 개발된 방법론이며 $\widehat{d_{i,j}}$를 구하는데 걸리는 시간 복잡도가 크기 때문에 많은 개체(보통 100개 이상)를 포함하는 nMDS 문제에는 적용하기 어려운 실정이다. 특히 최근 들어 생물정보학 등에서 분석에 요구되는 개체의 수가 100개가 넘는 경우가 많은 점을 고려하면 J. Kruskal 방법과 이를 보완한 방법들은 한계가 있다고 할 수 있다.

J. Kruskal이 제안한 nMDS 방법론의 단점을 극복하기 위한 방법으로 Y. Taguchi와 Y. Oono 등에 의해

제안된 "순수 nMDS"(purely nMDS)[4]가 있다. 이 방법은 단조변환을 통하여 구한 $\widehat{d_{i,j}}$ 를 사용하지 않는 방법으로, 비계량인 δ_{ij} 들을 크기 순서로 나열한 순서 척도(ordinal scale)를 통해 d_{ij} 를 결정하는 방법이다. 이 알고리즘의 핵심은 $\widehat{d_{i,j}}$ 를 사용하지 않고 순수하게 d_{ij} 와 δ_{ij} 를 크기순으로 정렬하여 정렬된 d_{ij} 와 δ_{ij} 의 순위를 각각 T_{ij} 과 τ_{ij} 으로 나타낸 다음, 모든 i,j 에 대하여 그 차이의 제곱의 합인 포텐셜 함수(potential function)

$$\Delta = \sum_{(i,j)} (T_{ij} - \tau_{ij})^2 \qquad (1)$$

가 최소화되도록 유클리드 공간상의 개체들의 배열(configuration)을 찾는 방법이다. 포텐셜 함수는 J. Kruskal 방법의 스트레스 함수와 유사한 역할을 하며 최적화 알고리즘 측면에서는 적합도 함수(fitness function)에 해당한다. 이 방법은 $\widehat{d_{i,j}}$ 를 구하지 않기 때문에 $\widehat{d_{i,j}}$ 를 구하는데 걸리는 계산 시간을 줄일 수 있고, 유일하게 결정할 수 없는 $\widehat{d_{i,j}}$ 를 배제함으로 보다 안정적이며, 또한 개체의 수가 100개 이상인 경우에도 현실적인 계산이 가능하다는 장점이 있다. 그러나 이 방법 역시 포텐셜 함수를 미분한 결과를 적용하는 최대 경사법(steepest decent)을 사용하여 최적치를 구하기 때문에 J. Kruskal 알고리즘과 마찬가지로 적합도 함수의 해가 일단 어느 지역 최적치(local optimum)에 도달하면 더 이상 향상된 해를 찾기 어렵다는 단점이 있다. 특히 Y. Taguchi와 Y. Oono는 전역 최적치를 찾는 알고리즘을 사용하여 최적의 배열을 구한다고 할지라도 그들이 제안한 방법론인 순수 nMDS와 별다른 차이가 없을 것이라고 주장하고 있다[4]. 그러나 이러한 주장은 검정되지 않은 상태이다. 물론 주어진 문제의 특성에 따라 지역 최적치와 전역 최적치의 차이가 크지 않아서 순수 nMDS가 효율적인 방법일 수 있으나, 일반적인 경우를 고려하면 위의 주장은 설득력이 떨어진다.

위의 주장을 검정하기 위하여 본 논문에서는 Y. Taguchi와 Y. Oono가 제안한 순수 nMDS에서 사용한 포텐셜 함수를 적합도 함수로 간주하고 최적화 관점에서 적합도 함수의 전역 최적치를 근사적으로 찾는 휴리스틱 알고리즘을 제안하고, 또한 실험을 통하여 순수 nMDS 알고리즘과 제안한 알고리즘을 비교 분석하고자 한다. 포텐셜 함수의 전역 최적치를 근사적으로 찾기 위하여 유전자 알고리즘 및 진화 프로그래밍 등의 진화 연산과 타부 탐색(taboo search) 등 여러 가지 휴리스틱 방법[13]을 적용할 수 있으나, 본 논문에서는 휴리스틱 방법론 중에 담금질(Simulated Annealing)[14]을 적용하여 포텐셜 함수를 최적화하는 방법을 제안한다. 따라서 제안된 담금질을 사용한 최적화 알고리즘은 포텐셜 함수의 전역 최적치를 찾기 위한 휴리스틱 알고리즘의 초기 연구 성격을 띠고 있다고 할 수 있다.

3. 담금질을 사용한 nMDS 알고리즘

Y. Taguchi와 Y. Oono가 제안한 순수 nMDS 알고리즘에서 사용한 포텐셜 함수는 비선형으로 다수의 지역 최적치가 존재함에도 불구하고 순수 nMDS 알고리즘은 최대 경사법으로 해를 구하기 때문에 어느 특정 지역 최적치에 일단 수렴하면 더 이상 향상된 최적치(혹은 전역 최적치)를 찾지 못하는 단점이 있다. 서론에서 언급된 Y. Taguchi와 Y. Oono의 주장처럼 만약 순수 nMDS 알고리즘을 사용하여 구한 지역 최적치가 전역 최적치와 별다른 차이가 없다면, 지역 최적치를 가지는 개체들의 공간 상 배열을 섭동(perturbation)시킨 후 다시 순수 nMDS 알고리즘을 실행하여 새로운 지역 최적치를 구해도 원래 찾은 지역 최적치와 크게 다르지 않을 것으로 예상할 수 있다. 따라서 지역 최적치를 가지는 배열의 섭동에 따른 최적치 향상 여부를 조사하는 것이 Y. Taguchi와 Y. Oono의 주장을 검정하는 방법이 될 수 있다.

이러한 논리는 지역 최적치를 가지는 배열에 대한 섭동을 담금질에 접목시켜 nMDS에 적용하는 새로운 알고리즘으로 구현될 수 있다. 담금질[14]은 글로벌 탐색 기법인 메타 휴리스틱중 하나로 Kirkpatrick 등에 의해 제안되어 현재 여러 분야에 폭 넓게 사용되고 있다. 담금질은 섭동을 사용하여 후보 해를 바꾸어 가면서 적합도 함수를 개선해가는 방법론으로 "비탈 오름"(uphill move)라는 개념을 도입하여 적합도 함수를 향상시키지 못하는 해에 대해서도 메트로폴리스 알고리즘(Metropolis Algorithm)[15]을 적용하여 그 해를 선택할 확률을 부여하여 점진적으로 전역 최적치에 접근하도록 고안된 방법론이다. 담금질을 적용한 nMDS 알고리즘은 지역 최적치에 수렴한 배열을 무작위로 섭동한 후, 섭동된 배열의 수락 여부를 메트로폴리스 알고리즘으로 결정하고, 다시 순수 nMDS 알고리즘을 적용하여 새로운 지역 최적치에 수렴하는 배열을 구하는 과정을 반복하는 것이다. 담금질을 적용한 nMDS 알고리즘의 순서는 아래와 같다.

1) N 개의 개체에 대하여 초기 배열(initial configuration)을 임의로 선정하고, 초기 "온도" T 를 설정한다.
2) 선정된 배열 $\overrightarrow{X} = [\overrightarrow{X_1}, \overrightarrow{X_2}, \cdots, \overrightarrow{X_N}]$ 에 대하여 순수 nMDS 알고리즘을 적용하여 포텐셜 함수 Δ 가 지역 최적치가 되는(즉, $\Delta = \Delta_{local}$) 배열 $\overrightarrow{X_{local}}$ 을 구한다.
3) 메트로폴리스 알고리즘을 적용한다.

3a) 무작위 섭동 $\overrightarrow{\delta X}$ 을 사용하여 섭동된 배열 $\overrightarrow{X_{new}} =$

$\overrightarrow{X_{local}}+\overrightarrow{\delta X}$를 구하고, 섭동된 배열 $\overrightarrow{X_{new}}$에 대한 포텐셜 함수 Δ_{new}를 구한다.

3b) $p_0=\exp(-(\Delta_{new}-\Delta_{local})/T)$와 일양 분포 $U[0,1]$을 따르는 임의의 값 p를 구한다.

3c) $p_0 \geq p$일 때까지 (3a)와 (3b)를 반복한다.

4) $\overrightarrow{X} \leftarrow \overrightarrow{X_{new}}$로 치환하고 온도를 $T \leftarrow \alpha T$로 낮춘다.

5) (2)-(4)를 정해진 횟수인 m 만큼 반복한다.

4. 실험 결과 및 분석

제안한 담금질 nMDS 알고리즘을 성능 측면에서 순수 nMDS 알고리즘과 비교 분석하기 위해서 벤치마킹 문제가 필요하다. 벤치마킹 문제는 복잡계 네트워크(complex networks)[16]라 불리는 네트워크 모델을 통해 생성할 수 있는 임의 네트워크(random network)[16]를 사용하였다. 복잡계 네트워크란 연구의 대상이 되는 시스템을 그 시스템을 구성하는 개체와 개체 사이의 상호 연관성을 개체들을 연결하는 링크(link, 혹은 edge)로 표현한 다음, 개체와 링크의 집합으로 구성된 네트워크를 여러 분석법을 사용하여 그 특성을 연구하는 분야이다. 임의 네트워크 모델은 네트워크를 구성하는 개체와 링크의 개수에 따라 개체들 간의 다양한 비유사성 δ_{ij}를 생성할 수 있으며, 또한 개체와 링크의 개수를 임의로 조절할 수 있다는 장점이 있기 때문에 본 연구를 통해 제안한 nMDS 알고리즘의 성능을 분석하는데 적당한 모델이다. 실험을 위하여 임의 네트워크 모델에서 링크 확률을 0.1로 고정하고 개체 수를 변화시켜 ($N=100, 150, 200, 250, 300$) 비유사성 데이터를 생성하였다.

담금질 nMDS 알고리즘에서 사용된 매개변수는 초기 온도 T, 담금질 스케줄 α, 반복 횟수 m, 그리고 섭동 $\overrightarrow{\delta X}$이다. 일반적으로 담금질에서 초기 온도 T는 가장 큰 $\Delta_{new}-\Delta_{local}$보다 매우 큰 값을 취함으로 임의의 배열 여러 개에 대하여 포텐셜 함수를 구하여 $T=100(\Delta_{new}-\Delta_{local})$ 정도가 되도록 선정하였고, 담금질 스케줄은 $\alpha=0.9$로 설정하였으며, 반복 횟수는 $m=30$으로 하였다. 또한 섭동 $\overrightarrow{\delta X}$는 각 개체($i=1,2,\cdots,N$)와 배열의 차원(2차원 $j=1,2$)에 대하여 $\delta X_{ij}=\beta U[-1,1]$로 설정하였다. 섭동의 크기를 나타내는 β는 개체 수 N에 의존하며 실험적으로 $\beta \approx 0.02(N-100)^2+5$를 얻었다. 또한 $U[-1,1]$은 [-1, 1]사이의 일양분포를 따르는 확률 변수이다.

순수 nMDS와 담금질 nMDS 알고리즘 간의 공평한 비교를 위하여 담금질에서 적용한 반복 횟수와 동일하게 순수 nMDS를 m번 반복 실행하여 그 중에서 가장 최적의 포텐셜 함수를 선택하였다. 따라서 $m=30$인 경우, 담금질 nMDS 알고리즘의 결과 1개와 순수 nMDS 알고리즘을 30번 적용하여 그 중에서 가장 우수한 결과 1개를 취하여 비교에 사용하였다. 이러한 실험을 개체 수 $N=100, 150, 200, 250, 300$ 각각에 대하여 10번 반복하여 최적 포텐셜 함수를 구하였으며, 그 결과에 대한 통계치는 아래 표 1과 같다. 개체 수와 매개변수에 따른 실험 결과는 섭동의 크기를 제외하면 매개변수 값에 비교적 무관하였다.

표 1 각 개체에 대해 10번 독립적인 실행을 한 후 구한 최적 포텐셜 함수의 표본 평균과 표본 표준편차 (괄호 안) 및 향상률

N	순수 nMDS	담금질 nMDS	향상률 η
100	1.047E+10 (9.051E+07)	1.015E+10 (8.081E+07)	3.2%
150	1.594E+11 (1.724E+09)	1.550E+11 (1.394E+09)	2.8%
200	9.843E+11 (4.825E+09)	9.779E+11 (5.355E+09)	0.7%
250	3.775E+12 (2.070E+10)	3.721E+12 (1.756E+10)	1.5%
300	1.111E+13 (4.850E+10)	1.096E+13 (4.804E+10)	1.4%

담금질을 사용한 nMDS 알고리즘의 성능을 순수 nMDS 경우와 정량적으로 비교하기 위하여 향상률을 $\eta=\frac{\text{순수}nMDS-\text{담금질}nMDS}{\text{담금질}nMDS}\times 100\%$로 정의하여 계산하였으며, 그 결과를 표 1에 포함시켰다. 표 1에서 볼 수 있듯이 향상률은 개체의 수가 증가함에 따라 작아지나, 개체의 수가 많아짐에 따라 최적화가 지수적으로 힘듦을 고려할 때 자연스러운 결과라 하겠다.

순수 nMDS와 담금질 nDMS의 성능을 비교 분석하기 위하여 위의 결과에 대한 통계적 검정을 실시하였다. 통계적 검정은 각 개체 수에 대하여 독립적으로 실시하였고, 동일한 개체 수에 대하여 두 가지 알고리즘을 사용하여 구한 최적 포텐셜 함수 값의 평균에 대한 검정을 실시하였다. 검정을 위하여 설정한 귀무가설(null hypothesis) H_0와 대립가설(alternative hypothesis) H_1은 다음과 같다.

$$H_0: \mu_{sa}=\mu_{pure} \quad H_1: \mu_{sa}<\mu_{pure} \tag{2}$$

여기서 μ_{sa}와 μ_{pure}는 각각 담금질과 순수 nDMS를 사용하여 구한 최적 포텐셜 함수의 평균을 나타낸다. 즉, 귀무가설은 두 알고리즘의 결과가 통계적으로 차이가 없다는 것이고, 대립가설은 포텔셜 함수를 최소화하는데 담금질 nMDS가 순수 nMDS 보다 우수하다는 것이다.

표본의 개수가 약 30개 이상인 경우에는 중심극한정리(Central Limit Theorem)에 의하여 정규 분포를 사용하여 위의 가설을 검정할 수 있으나, 본 실험의 경우

에는 표본의 개수가 10임으로 t-검정[17]을 실시하여야 한다. t-검정은 일반적으로 두 모집단의 분산이 동일한 경우와 그렇지 않는 경우 등 두 가지로 구분하는데, 표 1을 통해 볼 때 두 방법을 통한 결과에 대한 표본 표준 편차가 심각하게 다르지 않기 때문에 등분산 가정 하에 t-검정을 실시할 수 있다. 이 경우 검정통계량은

$$t_0 = \frac{\overline{X_1} - \overline{X_2}}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \quad (3)$$

로 주어진다. 여기서 $\overline{X_1}$, $\overline{X_2}$는 각각 순수 nMDS와 담금질 nMDS 알고리즘의 결과에 대한 표본 평균이며, n_1, n_2는 표본 개수 ($n_1 = n_2 = 10$), 그리고 S_p는 등분산 가정에 대한 합동추정량으로 $S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}$ 이고, S_1, S_2는 각각의 표본 표준 편차를 나타낸다. 아래 표 2는 개체의 수에 따라 t-검정을 사용해 구한 검정통계치이다. 유의 수준 $\alpha = 0.05$에서 검정을 실시한 결과, 표 2에서 볼 수 있듯이 모든 개체 수에 대하여 $t_0 > t(18, 0.05) = 1.734$ 임으로 식 (2)의 귀무가설 H_0를 기각하고 대립가설 H_1을 채택할 수 있다. 따라서 실험을 통해 볼 때 유의 수준 $\alpha = 0.05$에서 담금질 nMDS 알고리즘이 순수 nMDS 알고리즘보다 성능 면에서 우수함을 통계적 검정을 통해 입증할 수 있다.

표 2 각 개체 수에 대한 t-검정을 사용한 검정통계치

N	100	150	200	250	300
t_0	8.453	6.384	2.808	6.263	7.017

5. 요약 및 결론

본 논문에서는 비계량 다차원 척도법 문제를 위하여 제안된 순수 nMDS 알고리즘의 단점인 지역 최적치 수렴 현상을 해결하기 위하여 담금질에 기반한 nMDS 알고리즘을 제안하였다. 담금질을 적용한 nMDS 알고리즘은 지역 최적치로 수렴하는 배열을 섭동시킨 후 메트로폴리스 알고리즘을 적용하여 선택된 배열에 대하여 순수 nMDS를 적용하여 보다 향상된 최적치를 찾아가는 특징을 가지고 있다. 제안한 담금질 nMDS 알고리즘의 향상된 성능을 입증하기 위하여 임의 네트워크를 통해 생성된 비유사성 데이터에 담금질 nMDS 알고리즘을 적용하여 순수 nMDS 알고리즘의 결과와 비교 분석하였다. 실험 결과를 통계적 검정을 사용하여 분석한 결과, 고려한 모든 개체 수에 대하여 제안한 담금질 nMDS 알고리즘이 순수 nMDS 알고리즘 보다 최적화 측면에서 향상된 결과를 나타내었다.

비계량 다차원 척도법은 공학 분야뿐만 아니라 생물정보학, 경제학 등에 널리 적용될 수 있는 방법론임으로 새로운 nMDS 알고리즘에 대한 연구는 최적화 이론에 관한 기초 분야뿐만 아니라 응용 분야에도 적용할 수 있으므로 그 파급효과가 클 것으로 생각된다. 특히 최근 마이크로 어레이 관련 연구에 nMDS가 많이 적용되고 있음을 고려할 때 관련 바이오산업에 큰 효과를 미칠 것으로 보인다. 따라서 향후 다양한 휴리스틱 방법을 적용하여 보다 효율적인 최적화 알고리즘을 개발하는 연구가 필요하다고 하겠다.

참 고 문 헌

[1] J. Lattin et al, *Analyzing Multivariate Data*, Thomson, Nelson, 2003.

[2] J. Kruskal "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, vol.29, pp.1-27, 1964.

[3] J. Kruskal "Nonmetric Multidimensional Scaling," *Psychometrika*, vol.29, pp.115-129, 1964.

[4] Y. H. Taguchi, and Y. Oono, "Relational patterns of gene expression via non-metric multidimensional scaling analysis," *Bioinformatics*, vol.21, pp. 730-740, 2005.

[5] M. Garey and D. Johnson, *Computers and Intractability: A Guide to the theory of NP-completeness*, New York, W. H. Freeman, 1979.

[6] I. Shmulevich and W. Zhang, "Binary analysis and optimization-based normalization of gene expression data," *Bioinformatics*, vol.18, pp.555-565, 2002.

[7] R. Shepard "The analysis of Proximities: Multidimensional Scaling with an unknown Distance Function," *Psychometrika*, vol.27, pp.125-140, 1962.

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

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

[10] S. W. Malone, P. Tarazaga, and M. W. Trosset, "Better initial configurations for metric multidimensional scaling," *Computational Statistics & Data Analysis*, vol.41, pp.143-156, 2002.

[11] 박기용, 안성식, 정기용, "다차원척도법을 이용한 외식기업 경쟁요인 비교분석에 관한 연구", *외식경영학회*, vol.9, pp.93-114, 2006.

[12] 김철수, 정병두, 이원수, "공동주택 외부공간의 주거만족도 분석", *국토연구*, vol.53, pp.277-294, 2007.

[13] D. B. Fogel, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*, New York, IEEE Press, 1995.

[14] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi,

"Optimization by simulated annealing," *Science*, vol.220, pp.671-680, 1983.
[15] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, "Equation of state calculations for fast computing landscape," *J. Chem. Phys.*, vol.2, pp.1087-1093, 1953.
[16] M. Newman, "The structure and function of complex networks," *SIAM Rev.*, vol.45, pp.167-256, 2003.
[17] L. Chao, *Statistics: Methods and Analyses*, McGraw-Hill, N.Y., 1966.

이 창 용

1983년 서울대학교 계산통계학과 졸업(이학사). 1995년 미국 텍사스 주립대학교(Univ. of Texas at Austin) 물리학과 졸업(이학박사). 1996년~1998년 한국전자통신연구원 선임연구원. 1998년~2007년 공주대학교 산업정보학과 교수. 2007년~현재 공주대학교 산업시스템공학과 교수. 관심분야는 진화 연산 알고리즘 및 최적화 문제, 복잡계 네트워크(complex networks), 생물정보학(bioinformatics) 등

이 동 주

1996년 동아대학교 산업공학과 졸업(공학사). 1998년 Texas A&M University 석사, 2002년 Texas A&M University 산업공학과 졸업(공학박사). 2002년~2003년 미국 Texas Transportation Institute 연구원, 2003년~현재 공주대학교 산업시스템공학과 부교수. 관심분야로는 최적화, SCM, Simulation 등