저차원 영역에서 고차원 데이터 집합의 표현 방법

서동훈*, 아나스타씨야*, 이원돈*
*충남대학교 컴퓨터공학과
e-mail : sm1835dh@hanmail.net

Visualizing a Multi-Dimensional Data Set in a Lower Dimensional Space

Dong-Hun Seo*, Kolesnikova Anastasiya*, Won Don Lee*
*Dept. of Computer Science & Engineering, Chungnam National University

요 약

본 논문에서는 고차원 영역의 데이터 집합을 저차원 영역으로 표현하는 방법에 대해서 제안한다. 특별히 고차원 영역을 2 차원 영역으로 실험하였다. 제안한 방법은 사람이 데이터 객체 사이의 거리나 관계를 직관적으로 인지할 수 있도록 하는 방법이다. 데이터 객체 사이의 거리나 관계를 계산하기 위하여 Kullback-Leibler divergence 를 사용하였다. 이 방법은 확률 분포를 갖는 벡터들 사이의 거리를 계산하여 사용한다. Kullback-Leibler divergence 를 사용하여 계산된 거리 값들은 저차원 영역에서 객체들의 좌표를 계산하기 위하여 사용된다. 좌표계산을 위해서 Simulated Annealing 란 최적화 기법을 사용하였다. 실험 결과를 통해 다차원 데이터를 2 차원 영역으로 표현한 것이 충분히 직관적임을 보였다.

1. 서론

일반적으로 정보는 문서의 형태보다 그림의 형태를 가질 때 더 쉽게 뜻을 전달할 수 있다. 그래서 포스터와 그래프처럼 많은 시각화 방법들은 기존의 정보를 보다 더 잘 전달하기 위해 발명되었다.

본 논문에서 제안된 방법은 거리를 측정하는 프로세스와 포지셔닝 프로세스로 구성되었다. 거리를 측정하는 프로세스는 데이터들 사이의 거리를 계산하기 위한 것이다. 계산하는 방법으로는 Kullback-Leibler divergence 를 사용하였다. 유클리안 거리와는 달리 이 방법은 비대칭적이기 때문에 일반적인 거리의 개념으로 직접 사용할 수 없다. 그래서 우리는 거리 행렬을 사용하여 수학적인 의미를 부여하여 사용하였다. 포지셔닝 프로세스는 고차원 데이터를 저차원 영역의 좌표로 옮기기 위하여 거리 행렬과 Simulated Annealing 을 사용하였다.

2. Kullback-Leibler divergence and Simulated Annealing

2.1 Kullback-Leibler divergence

Kullback-Leibler divergence 는 이미 정보이론에서 잘 알려진 개념이다[1]. 이산분포를 위하여 Kullback-Leibler divergence p 와 q 사이를 아래와 같이 정의 하였다. $KL(p,q) = \sum p(x) \log \frac{p(x)}{q(x)}$

이 식을 2 개의 확률 분포들 사이의 거리로 사용하기 위해선 Kullback-Leibler divergence 의 비대칭적인 특성을 해결해야 한다. Kullback-Leibler divergence 에 대칭성을 주기 위한 많은 시도가 있었다.[2] 그리고 우리는 거리를 사용하기 위해서 수학적 평균을 사용하였다. 수식은 아래와 같다.

$$Dist(p,q) = (KL(p,q) + KL(q,p)) / 2$$

2.2 Mean Field Annealing

솔루션의 품질을 보장하면서 솔루션을 빠르게 구해야 하는 조합된 NP-complete 최적화 문제에 Mean Field Annealing 신경 회로망은 널리 사용되고 있다[5].

n 상태의 이산 변수를 갖는 회로망 모델에서 평형 상태 를 갖는 $< s_i >$의 스핀 평균 값은 모든 스핀들이 평형 상태의 온도를 갖는 볼츠만 분포를 가정함으로써 계산할 수 있다. 수식은 아래와 같다.

$$< s_i > = \Pr\{s_i = 0\} \times 0 + \cdots + \Pr\{s_i = n-1\} \times (n-1)$$
$$= \sum_{i=0,n-1} s_i \exp(-H_i / T) / \sum_{i=0,n-1} \exp(-H_i / T)$$

$H_i = < H(s) > | s_i$ =i, $H(s)$ 이 Hamiltonian 이고 $s_i = \{0,...,n-1\}$ 인 경우.

연속 변수를 갖는 mean field 이론 신경 회로망 안의상황은 각 스핀들이 실수 값을 갖는 것을 제외하곤 이산 변수를 갖는 회로망과 동일하다. 그래서, 평형 상태를 갖는 스핀의 평균을 이산 회로망으로부터 계산할 수 있다. 이산 회로망은 실수 값과 함께 스핀의 확률로 구성한다. 이것을 수식으로 표현하면 다음과 같다.

$$< s_i > = \frac{\int s_i \exp(-H(s)/T) ds_i}{\int \exp(-H(s)/T) ds_i}$$

2.3 저차원 영역에서 고차원 데이터 집합의 표현 방법

본 논문에서 제안된 방법은 다음의 2 부분으로 구성된다.

1) 거리를 측정하는 프로세스 : 그룹들 간의 거리를 측정하는 프로세스. 여기서, 그룹이란 특성들로 표현된다. 각각의 특성들은 모아진 많은 객체들로부터 추출한 값들을 의미한다. n 개의 속성들을 갖는 데이터 집합을 위해서 그룹의 확률 분포를 계산한다. 그리고 그룹들 사이의 거리를 Kullback-Leibler divergence 를 사용하여 계산한다. 순서는 아래와 같다.

① 속성들의 개수 'n'을 결정한다.
② 각각의 속성들의 빈도 수를 확인한다.
③ 각각의 속성들의 본도 수로부터 그룹의 벡터를 만든다.

$$V = (v_1, v_2, ..., v_n)$$ (n: 속성들의 수)

④ 벡터 V 를 표준화를 하여 확률 분포로 변형시킨다.

$$P = (v_1/K, v_2/K, ..., v_n/K)$$

$$(K = \sum_{i=0}^{n} v_i)$$

⑤ 그룹 'p'와 'q' 사이의 거리를 Kullback-Leibler(KL) divergence 를 사용하여 계산한다.

$$(KL(p,q) + KL(q,p))/2 \quad (1)$$

⑥ 그룹들의 거리 행렬을 만든다.

<표 1> 거리 행렬의 예제

0	1	$\sqrt{2}$	1
1	0	1	$\sqrt{2}$
$\sqrt{2}$	1	0	1
1	$\sqrt{2}$	1	0

2) 포지셔닝 프로세스 : 저차원 영역에서 그룹들간의 상대적 거리를 고려하여 표현하는 프로세스. 예제 데이터 집합인 <표 1>을 2 차원 영역으로 표현 가능한 예 중 하나는 (그림 1)과 같다. 그룹 a 는 (1,1)의 좌표에 위치하고, 그룹 b 는 (2,1), 그룹 c 는 (2,2) 그리고 그룹 d 는 (1,2)에 위치한다. <표 1>에서 보는 것과 같이 각각의 그룹들 간의 거리들이 같다는 것을 알 수 있다. (그림 1)에서 각각의 그룹들 간의 거리도 <표 1>과 같이 표현되었음 알 수 있다. 포지셔닝 프로세스의 순서는 아래와 같다.

① 그룹들을 표현할 차원의 수를 결정한다.
② 비용 함수를 정의 한다.

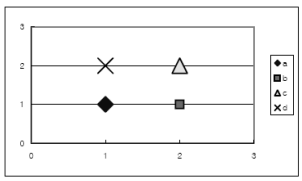
$$Cost = \sum |Dist_{real} - Dist_{cur}| \quad (2)$$

③ 비용을 최소화하면서 각각의 그룹들의 좌표 값들을 Simulated Annealing 을 통하여 정한다. 여기서, $Dist_{real}$ 는 거리 행렬 안에 쓰여지는 두 그룹들간의 거리이다. 그리고 $Dist_{cur}$ 는 두 그룹 간의 현재 유클리안 거리이다. 그리고 식 (2)를 통한 비용과 실제 거리의 합의 비율로 에러 식 (3)를 정의 한다.

$$Error = Cost / \sum Dist_{real} \quad (3)$$

$Dist_{real}$ 은 거리 행렬을 2 로 나눈 것의 대각 성분들의 합이다.

에러를 최소화하기 위하여 Simulated Annealing 을 사용하여 그룹들 간의 좌표를 계산하였다.



(그림 1) 거리 행렬을 사용하여 2 차원으로 표현한 예제

3. Experiments and Results

실험은 UCI depository 에 있는 multi-feature digit 데이터 집합을 사용하였다. 이 집합은 필기체 숫자들 ('0'부터 '9')의 특성들로 만들어진 10 개의 패턴들로 구성되었다. 한 숫자 데이터 집합은 200 개의 패턴들을 가지고 있고, 전체 데이터 집합은 총 2000 개의 패턴들로 구성되었다. 이 데이터 집합 안에는 6 개의 데이터 집합이 있고 각각은 다른 특성 집합들로 구성되었다.

이 중 우리는 거리 행렬을 계산하고 표현하기 위해 숫자의 형태에 대한 퓨리에 계수들로 이루어진 데이터 집합을 사용하였다.

퓨리에 계수들은 Kullback-Leibler divergence 로 계산하기에 부적절 하므로 이 값들을 확률 분포 형태로 바꿀 필요가 있다. 또한 한 숫자에 대해 200 개의 패턴들 모두를 2 차원 영역에 표현하는 것은 계산 량의 문제로 불가능하다. 그래서 200 개의 패턴에 대한 중심 값을 먼저 계산한 후, 숫자 패턴들의 중심 값들의 거리를 이용하여 거리 행렬을 구하였다. 이렇게 얻은 10 개의 그룹들은 각각 76 개의 표준화된 퓨리에 계수들로 구성된다. 표준화 하는 것으로부터 우리는 Kullback-Leibler divergence 계산시 이 값을 확률 분포처럼 사용할 수 있게 되었다.

우리는 그룹들 사이의 거리를 얻기 위하여 10 * 10 Kullback-Leibler 를 만들었다. 여기서 그룹들의 총 개수는 10 개이다.

하지만 Kullback-Leibler divergence 의 특성상 대칭이 아니므로 (i,j)의 값과 (j,i)의 값이 다르다.

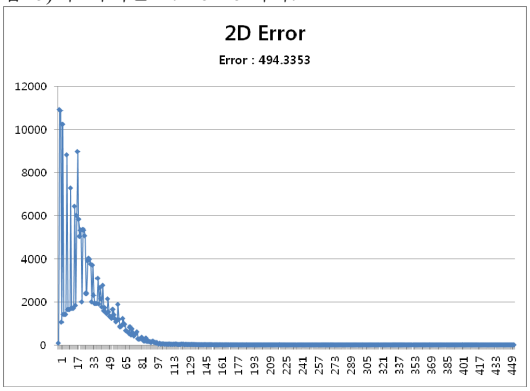
<표 2> 그룹들간의 거리 행렬

0	2.704554	2.093457	2.470168	1.926781	1.448448	2.931705	1.652479	2.550256	1.816372
2.704554	0	3.878108	3.559997	3.717953	3.377634	4.264472	4.180896	3.684899	4.113597
2.093457	3.878108	0	3.645257	3.323072	3.972165	3.731020	4.243260	3.814103	4.335361
2.470168	3.559997	3.645257	0	3.912784	3.560531	3.982981	3.690097	3.938813	3.972205
1.926781	3.717953	3.323072	3.912784	0	4.439165	3.807345	3.635116	3.495409	3.438541
1.448448	3.377634	3.972165	3.560531	4.439165	0	4.529304	3.885749	3.785129	4.071760
2.931705	4.264472	3.731020	3.982981	3.807345	4.529304	0	4.067664	3.968370	3.198556
1.652479	4.180896	4.243260	3.690097	3.635116	3.885749	4.067664	0	3.825300	3.535964
2.550256	3.684899	3.814103	3.938813	3.495409	3.785129	3.968370	3.825300	0	3.442366
1.816372	4.113597	4.335361	3.972205	3.438541	4.071760	3.198556	3.535964	3.442366	0

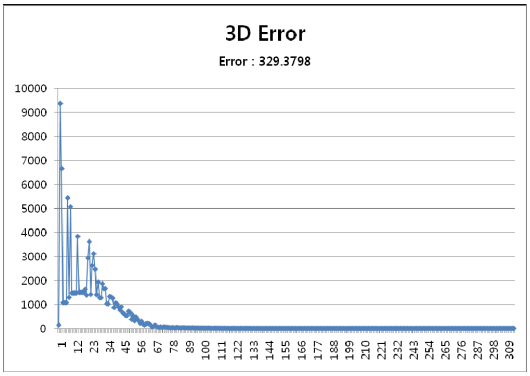
비록 Kullback-Leibler divergence 는 두 그룹간의 차이를 반영하지만 비대칭 특성은 두 그룹간의 값을 거리로 사용하기에 어렵게 만든다.

그래서 우리는 (i,j)와 (j,i)의 값을 더한 후 둘로 나누어 거리 값을 얻었다. 결과적으로 이 값을 가지고 이 그룹들의 거리 행렬을 만들었다.

(표 2)는 그룹들간의 거리 행렬을 나타낸다. 예를 들어 숫자 '4' (그룹 4) 그리고 숫자 '6' (그룹 6)은 거리가 3.807345 이다. 숫자 '0' (그룹 0) 와 숫자 '5' (그룹 5)의 거리는 1.448448 이다.

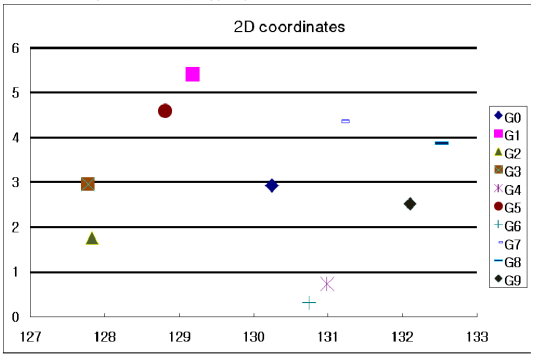


(그림 2) 2 차원 영역에서 에러값

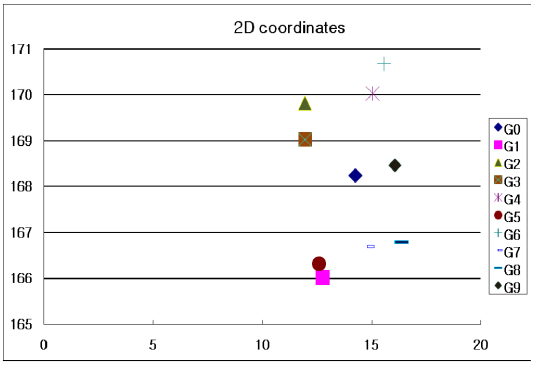


(그림 3) 3 차원 영역에서 에러값

(그림 2)와 (그림 3)는 2 차원과 3 차원에서 에러값을 보여준다. 이 값은 Simulated Annealing 에서 온도값으로 사용하였다. 그래프 제목 밑의 에러는 마지막 에러값을 의미한다. 제안된 방법으로 1 차원 영역에 표현한 경우 결과가 해석하기 힘들게 나타났다. 이것은 1 차원 좌표는 그룹들 간의 관계를 표현하기에 별로 좋지 않다는 것을 보여준다. 본 논문에서 제안된 방법은 2 차원과 3 차원의 영역에서 좋은 결과를 얻는 것을 보여준다. 논문에서 Simulated Annealing 을 사용함으로써 빠르게 에러가 적은 쪽으로 결과를 수렴하는 것을 볼 수 있다. 2 차원 영역에서 마지막 에러는 494.3353 이고 3 차원 영역에서 마지막 에러는 329.3798 임을 볼 수 있다.



(그림 4) 첫번째 2 차원 좌표 영역



(그림 5) 두번째 2 차원 좌표 영역

2 차원 영역에서 많은 실험을 하였지만 본 논문에서는 2 차원 영역의 (그림 4)와 (그림 5)를 실험 결과로 보여주고 있다. 이것들은 제안된 방법들의 특징을 잘 나타내 주고 있다.

여기서 제안된 방법은 그룹들간의 관계를 보여주는

방법으로써 우리는 그룹들간의 거리 관계를 표현 하는 방법에 관심이 있다. 이것을 하기 위하여 그룹들간의 관계를 저차원 영역에 보여주기 위해서 그룹들간의 좌표를 결정 하는데 최적화 알고리즘을 사용한 이유이다.우리는 그룹들간의 관계를 분석하기 위하여 저차원 영역에서 좌표값을 발견하는 것에 집중을 하였다.

(그림 4)와 (그림 5)는 실험 결과를 보여준다. (G1, G5), (G4, G6), (G2, G3), (G7, G8)은 다른 것들보다 더 가깝게 있음을 알 수 있다. 여기서 그룹 Gi 는 숫자 i 의 중심 좌표를 나타낸다. 예를 들어 G5 은 숫자 '5' 를 의미한다. 다르게 말하자면, 숫자 (1,5), (4,6), (2,3), (7,8)은 다른 경우보다 더 밀접한 관계를 가진 것을 의미한다. 그리고 이것은 사람이 직관적으로 판단하기에 부합하다.

4. Conclusion

본 논문에서는 UCI repository 에 있는 숫자 데이터 집합을 가지고 실험을 하였다. 이 데이터 집합은 확률 분포를 갖는 벡터들로 변환 후 Kullback-Leibler divergence 를 사용하여 각각의 그룹들간의 거리를 측정하였다. 만약 거리를 측정하기에 잘 정의된 데이터 집합이라면 Kullback-Leibler divergence 를 사용할 필요는 없다. 여기선 두 확률 분포의 비율을 사용하였다.

본 논문에서 제안된 방법은 고차원의 데이터 집합을 저차원 영역으로 재배치하여 표현함으로써 데이터들 간의 관계를 적절히 표현하였다. 실험을 통하여 저차원 영역에서의 좌표 값들을 계산하였다. 실험을 통하여 제안된 방법이 사람이 직관적으로 이해하기 쉬운 형태로 표현되는 것을 보였다.

참고문헌

[1] Robert M. Gray, "Entropy and Information Theory (Book style)" revised November 2000. Available at: http://www-ee.stanford.edu/~gray/.

[2] D. H. Johnson and S. Sinanovic, "Symmetrizing the Kullback-Leibler distance", Technical report, Rice University, 2001. Available at: http://cmc.rice.edu/docs/.

[3] Dezhong Peng, Zhang Yi "Dynamics of Generalized PCA and MCA Learning Algorithms", Neural Networks, IEEE Transactions on, Dec. 2007. Vol. 18, Issue: 6, pp. 1777 – 1784.

[4] Nasser A, Hamad D, Nasr C, "Visualization Methods for Exploratory Data Analysis", Information and Communication Technologies, 2006. ICTTA '06. 2nd. Vol. 1, pp. 1379 – 1384.

[5] Jain A.K. and J. Mao, "Artificial Neural Network for Nonlinear Projection of Multivariate Data", Proc. IEEE Int. Joint. Conf. on Neural Networks, 1992. Vol. 3, pp. 335-340.

[6] Chi-Hwa Song, Jin-Ku Jeong, Dong-Hun Seo, Won Don Lee, "A Mean Field Annealing Algorithm for Fuzzy Clustering", Fourth International Conference on Fuzzy Systems and Knowledge Discovery 2007, 2007. Vol. 2, pp. 193-197.

[7] Tae-Hyoung Kim, Chi-Hwa Song, Won Don Lee, Jae-Cheol Ryou, "Building a Packages Delivery Schedule Using Extended Simulated Annealing", 2006 International Joint Conference on Neural Networks, 2006, pp. 2691-2695.

[8] Oh-Jun Kwon, Won-Don Lee and Sung-Yang Bang, "Modified mean field annealing algorithm for combinatorial optimization problems with continuous state space", ELECTRONICS LETTERS, 22nd May 1997 Vol. 33, No. 11, pp. 968-969.