연속 변수 함수 최적화를 위한 Variational 혼합 인자 분석 베이지안 진화 연산

조동연° 장병탁
서울대학교 컴퓨터공학부 바이오지능연구실
{dycho°, btzhang}@bi.snu.ac.kr

Bayesian Evolutionary Computation by Variational Mixtures of Factor Analyzers for Continuous Function Optimization

Dong-Yeon Cho° Byoung-Tak Zhang
Biointelligence Laboratory
School of Computer Science and Engineering, Seoul National University

요 약

연속 변수 함수 최적화를 위한 진화 연산에서는 전통적으로 확률 분포를 도입하여 새로운 세대를 생성하는 기법을 사용하고 있다. 최근 들어 이러한 확률 분포를 개체군으로부터 추정하여 보다 효율적으로 최적화를 해결하려는 연구가 진행되고 있다. 본 논문에서는 variational 베이지안 혼합 인자 분석 기법(Bayesian mixtures of factor analyzers)을 사용한 개체군의 분포 추정을 통해 연속 변수 함수의 최적화 문제를 해결하는 방법을 제안한다. 이 기법은 혼합 분포의 개수 추정을 자동화하여 개체군의 다양성을 유지할 수 있기 때문에 지역 최적점으로 일찍 수렴하는 현상을 방지할 수 있으며, 세부 개체군 내의 분포 추정을 통해 탐색을 효율적으로 수행할 수 있다. 잘 알려진 평가 함수들에 대하여 다른 분포 추정 진화 연산과 비교하여 제안하는 방법의 우수성을 검증하였다.

1. 서 론

진화 전략(evolution strategy)과 같이 연속 변수를 다루는 진화 연산에서는 전통적으로 확률 분포를 사용하여 새로운 세대를 생성하는 기법을 사용해 오고 있다. 최근 들어 이 확률 분포를 개체군으로부터 추정하여 보다 효율적으로 탐색을 수행하려는 시도가 이루어지고 있다. 분포 추정 알고리즘(Estimation of Distribution Algorithm, EDA)으로 불리워지는 이러한 기법들은 현재 개체군 중 적합도가 좋은 개체들을 데이터로 사용하여 그 분포를 표현할 수 있는 확률 모델을 학습하고, 이 모델로부터 새로운 개체들을 생성하게 된다. 즉, 기존의 진화 연산과는 달리 교차나 돌연변이 연산을 이용하지 않고 학습된 확률 분포로부터의 표본 추출을 통해 새로운 탐색점을 만들어 낸다.

여기서 우수한 개체들의 확률 분포를 추정한다는 것은 주어진 문제에 존재하는 변수들의 간의 관계를 파악하는 것을 의미하여 확률 모델의 관계 표현 능력에 따라 알고리즘의 분류가 가능하다[1,2]. 특별히 연속 변수를 다루는 문제에 있어서는 주로 정규 분포가 사용되어 왔으며, 단순한 문제에서는 각 변수들의 독립성을 가정하기도 하지만 보통 각 변수들 간의 관계는 공분산 행렬을 통해 표현된다. 문제가 복잡해지면 일반적인 정규 분포를 이용하더라도 변수들 사이의 관계를 정확하게 파악하기 어렵기 때문에, 단 하나의 정규 분포를 사용하여 최적점을 탐색할 때 지역 최적점으로 일찍 수렴하는 현상이 발생하게 된다.

이러한 단점을 극복하기 위해 정규 혼합(normal-mixture) 모델을 도입하는 연구가 이루어졌다[3]. 이는 여러 개의 개체군을 사용하는 진화 연산 기법과도 그 맥을 같이 하며, 전체 개체군의 분포를 다수의 세부 개체군에 대한 혼합 분포로 표현함으로써 개체군의 다양성을 유지하는데 기여하게 된다. 그러나 이러한 혼합 분포를 추정하는 경우 몇 개의 분포로 개체군을 표현하는 것이 적합한 가를 파악하기 어려운 단점이 있다.

본 논문에서는 variational 베이지안 혼합 인자 분석 기법(Bayesian mixtures of factor analyzers)을 사용한 개체군의 분포 추정을 통해 연속 변수 함수의 최적화 문제를 해결하는 방법을 제안한다. 이 기법은 혼합 분포의 개수 추정을 자동화하여 개체군의 다양성을 유지할 수 있기 때문에 지역 최적점으로 일찍 수렴하는 현상을 방지할 수 있으며, 세부 개체군 내의 분포 추정을 통해 탐색을 효율적으로 수행할 수 있다.

2. Variational 베이지안 혼합 인자 분석을 이용한 진화 연산

일반적인 인자 분석(factor analysis) 기법은 가장 단순한 잠재 변수(latent variable) 모델 중의 하나로 다음과 같은 선형 관계를 가정한다.

$$x = Az + u + e$$

여기서 z는 표준 정규 분포를 따르는 잠재 변수를 나타내고, u은 주어진 데이터 x의 평균값이며, e는 평균이 0이고 공분산 행렬 C가 대각 행렬이 되는 정규 분포로부

터의 오차를 나타낸다. 따라서 $p(\mathbf{x}|\mathbf{z})$는 평균이 $\mathbf{Az} + \mathbf{u}$ 이고, 공분산 행렬이 C인 정규 분포가 된다. 결국 인자 분석은 데이터를 가장 잘 설명할 수 있는 인자적재행렬 (factor loading matrix) A와 오차의 공분산 행렬 C를 찾는 것이 된다. 이것은 보통 잠재 변수 값의 추정과 함께 EM 알고리즘의 적용을 통하여 이루어진다[4].

본 연구에서는 S개의 인자 분석 모델을 기반으로 다음과 같은 혼합 분포를 고려한다.

$$p(\mathbf{x}) = \sum_{s=1}^{S} h_s p(\mathbf{x}|s) = \sum_{s=1}^{S} \int p(\mathbf{x}|\mathbf{z},s)p(\mathbf{z}|s)p(s)d\mathbf{z}$$

여기서 $p(\mathbf{x}|\mathbf{z},s)$는 하나의 인자 분석 모델이 나타내는 분포이다. 따라서 이 분포는 잠재 변수 $\mathbf{z}$가 주어졌을 때 다루고자 하는 데이터, 즉 선택된 개체군의 서로 다른 부분을 각각의 인자 분석 모델로 나타내게 된다. 또한 $h_s = p(s)$는 혼합 비율이며, $p(\mathbf{z}|s)$는 하나의 인자 분석 모델에서와 마찬가지로 s에 관계없이 모두 표준 정규 분포를 따른다. 혼합 인자 분석 모델에서는 각 성분의 인자 분석 모델을 결정짓는 인자적재행렬과 잠재 변수 및 공분산 행렬에 추가적으로 혼합 비율을 고려하여 EM 알고리즘으로 학습이 가능하다[5].

그러나 일반적인 EM 알고리즘은 혼합 분포의 개수가 늘어나면 계속해서 가능도(likelihood)가 커지게 된다. 즉 몇 개의 분포를 사용하여 주어진 데이터를 설명하는 것이 가장 좋은지를 결정할 수가 없게 된다. 이러한 단점을 해결하기 위해 그림 1과 같은 혼합 인자 분석에 대한 베이지안 확률 모델을 도입하게 된다. 여기서는 각 파라메터에 대한 확률 분포와 이 분호를 위한 사전 확률 분포 및 각 사전 확률 분포를 위한 초파라메터(hyper-parameter)들을 정의한다.

이와 같은 확률 모델 위에서 로그 주변 가능도 (log marginal likelihood)의 하한을 Jensen의 부등식을 이용하여 다음과 같이 정의할 수 있다.

$$\log p(\mathbf{x}) = \log \int p(\mathbf{x},\mathbf{w})d\mathbf{w} \geq \int p'(\mathbf{w})\log\{p(\mathbf{x},\mathbf{w})/p'(\mathbf{w})\}d\mathbf{w}$$

여기서 $\mathbf{w}$는 잠재 변수를 포함하여 우리가 모르는 모든 은닉 변수를 의미하며, $p'(\mathbf{w})$는 사후 확률 $p(\mathbf{w}|\mathbf{x})$에 대한 다루기 쉬운 근사화를 나타낸다. Variational 베이지안 EM 알고리즘은 반복적인 과정을 통해 위의 하한 값을 최대화하여 결과적으로는 로그 주변 가능도를 최대화한다. 이 과정은 $p'(\mathbf{w})$과 $p(\mathbf{w}|\mathbf{x})$의 KL 발산량(Kullback-Leibler divergence)을 최소화하는 것과 같다. 이와 더불어 잠재 변수의 분포를 나타내기 위한 평균과 공분산, 그리고 혼합 인자 분석을 위한 인자 적재행렬과 오차의 공분산 등도 함께 구하게 된다. (잠재 변수 및 다른 은닉 변수들에 대한 자세한 사전 확률 분포와 초파라메터의 설정 및 반복적인 갱신 과정의 유도는 [6]에 제시되어 있다.)

정해진 탐색 범위 내에서 임의로 생성된 초기 개체군에 대하여 주어진 함수로 적합도를 계산한 후에 토너먼트 선택 기법에 의하여 부모들을 고른다. 이것을 데이터로 하여 위에서 설명한 variational 베이지안 EM 알고리즘으로 혼합 인자 분석 모델을 학습한다. 이렇게 학습된 확률 모델로부터 새로운 개체를 생성하기 위하여, 데이터로 사용된 부모들 중 임의로 한 개를 고른 후에 그 데이터에 대한 혼합 비율에 따라 인자 분석 모델을 선택하다. 이로부터 잠재 변수의 평균값을 사용하여 $p(\mathbf{x}|\mathbf{z})$로부터 다시 새로운 개체 $\mathbf{x}$를 생성할 수 있다. 이러한 생성 과정을 정해진 수만큼 반복하여 자손들을 만들고 그들의 적합도도 평가한다. 각각의 자손 개체에 대하여 이전 세대의 전체 개체군 중 일부를 임의로 선택하고 그 중에서 가장 비슷한, 즉 실수 공간상에서는 거리가 가장 가까운 개체와 비교하여 새로운 개체가 더 좋다면 그 개체를 대체하게 된다. 이러한 제한된 토너먼트 대체 기법 (restricted tournament replacement, RTR)은 혼합 분포로 학습된 개체군의 다양성을 유지하는데 더욱 도움을 준다[7].

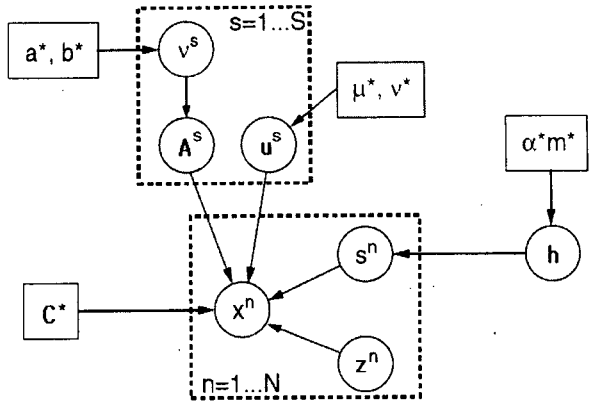

그림 1. 혼합 인자 분석에 대한 베이지안 확률 모델 ([6]에서 변형, N은 데이터의 개수)

3. 실험 및 결과

제안된 방법의 성능을 평가하기 위해 널리 사용되고 있는 연속 변수 함수 중에서 다음의 두 함수를 선택하였다.

$$f_{Ro}(\mathbf{x}) = \sum_{i=1}^{9}(100(x_{i+1}-x_i^2)^2 + (1-x_i^2)),\ |x_i| \leq 5.12$$

$$f_{FP}(\mathbf{x}) = \sum_{i=1}^{10}(\sum_{j=1}^{10}(a_{ij}\sin a_j + b_{ij}\cos a_j) - \sum_{j=1}^{10}(a_{ij}\sin x_j + b_{ij}\cos x_j))^2,$$
$$|x_i| \leq \pi \quad |a_j| \leq \pi \quad a_{ij},\ b_{ij} \in \mathrm{Int}[-100, 100]$$

두 함수 모두 각 변수들이 서로 연관되어 있어서 각 변수들을 분리하여 최적화하기는 불가능하다. f_{Ro}는 x_i가 모두 1일 때, 그리고 f_{FP}는 $x_j = a_j$ 일 때 최소값 0을 갖는다. (FP 함수의 정의를 위해 [8]에 제시된 값을 사용하였다.)

크기가 2인 토너먼트 선택 기법에 의하여 전체 개체군 중 절반을 부모로 골라 variational 베이지안 EM 알고리즘으로 혼합 인자 분석 모델을 학습한다. 이 학습은 로그 주변 가능도의 변화가 5% 미만이 될 때까지 반복된다. 앞 절에서 설명한 방법으로 새로운 개체를 생성하고 적합도를 평가한 후에 전체 개체군에서 5%의 개체들을 선택하여 RTR 기법으로 개체군을 갱신한다. 이 과정을

적합도 평가 회수가 10^6이 될 때까지 진행한다.

기존의 확률 분포 추정 진화 연산 기법 중 그 성능이 좋다고 알려진 MIDEA[3]와 MBOA[7]를 앞에서 설명한 함수들에 적용하여 제안된 알고리즘과 비교하였다. MIDEA는 개체군에서 성능이 좋은 30%를 선택하여 BEND leader 알고리즘으로 군집화를 한 후 (분계점은 0.3으로 설정), 각 군집에 대하여 Gaussian 네트워크로 확률 분포를 추정한다. MBOA에서는 결정 트리와 유사하게 축에 평행한 값으로 데이터 공간을 나눈 후에 단말 노드에서는 Gaussian 커널을 사용하여 확률 분포를 추정한다. 이 기법에서도 역시 RTR을 채택하였다.

개체군의 크기를 200, 400, 800, 1600, 3200, 6400으로 변화 시키며 각 알고리즘을 20번씩 수행하여 가장 좋은 성능을 얻은 결과가 그림 1에 제시되어 있다 (20번 수행의 평균). 전체적으로 MIDEA보다는 다른 두 알고리즘이 더 좋은 성능을 보이며, 특히 지역 최적점이 많은 Fletcher-Powell 함수에 대하여 제안된 방법이 월등히 우수한 성능을 나타내고 있다. 또한 가장 좋은 성능을 얻기 위해 필요한 개체군의 크기도 제안된 방법이 가장 적다.

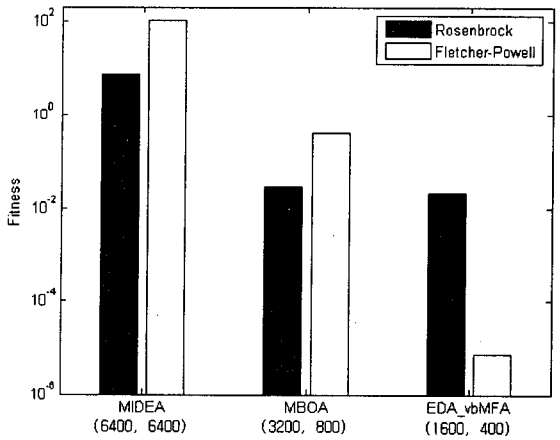

그림 2. 테스트 함수에 대한 알고리즘의 성능 비교 및 최적 개체군의 크기

각 알고리즘에 대하여 최적 성능을 얻기 위해 소요된 적합도 평가 회수와 상대적인 계산 시간(RT)이 표 1에 제시되어 있다. MIDEA는 비록 그 효율은 좋지만 최적점을 구하는 데 모두 실패하였다. 제안된 방법은 Rosenbrock 함수에 대해서 MBOA와 성능은 비슷하나 더 효율적이고, Fletcher-Powell 함수에 대해서는 비록 더 많은 계산량을 요구하지만 월등히 좋은 성능을 나타낸다.

표 1. 각 알고리즘의 계산 비용 비교

Algorithm	Rosenbrock		Fletcher-Powell	
	#Eval.	RT	#Eval.	RT
MIDEA	937328	6.888	986843	0.255
MBOA	969520	5135.15	992280	7.537
EDA_vbMFA	981920	346.871	878520	270.892

4. 결론

본 논문에서는 variational 베이지안 혼합 인자 분석 기법을 사용한 개체군의 분포 추정을 통해 연속 변수를 다루는 진화 연산의 성능을 향상 시키는 방법을 제안하였다. 여기서는 개체군의 군집화를 통한 다양성 유지가 혼합 분포의 개수를 자동적으로 추정함으로써 이루어지고 각 개별 분포는 인자 분석 기법에 의하여 효율적으로 파악된다. 제안된 기법을 잘 알려진 테스트 함수에 적용하여 기존의 분포 추정 진화 연산과 비교한 결과, 최적화 성능 면에서 혹은 알고리즘의 계산 효율 면에서 그 우수성을 입증하였다.

감사의 글

본 연구는 교육인적자원부 BK21-IT, 산업자원부 차세대 신기술 개발 사업의 분자 진화 컴퓨팅 과제 및 과학기술부 국가지정연구실 과제에 의하여 일부 지원되었다. 또한 이 연구를 위해 장비를 지원하고 공간을 제공한 서울대학교 컴퓨터연구소에도 감사드린다.

참고문헌

[1] P. Larranaga, A review on estimation of distribution algorithms, *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*, pp. 57-100, 2001.

[2] M. Pelikan, D. E. Goldberg, F. G. Lobo, A survey of optimization by building and using probabilistic models, *Computational Optimization and Applications*, vol. 21, no. 1, pp. 5-20, 2002.

[3] P. Bosman, D. Thierens, Advancing continuous IDEAs with mixture distributions and factorization selection metrics, *Proceedings of 2001 GECCO Workshop Program*, pp. 208-212, 2001.

[4] D. B. Rubin, D. T. Thayer, EM algorithms for ML factor analysis, *Psychometrika*, vol. 47, no. 1, pp. 69-76, 1982.

[5] Z. Ghahramani, G. E. Hinton, The EM algorithm for mixtures of factor analyzers, Department of Computer Science, University of Toronto, Technical Report CGG-TR-96-1, 1997.

[6] M. J. Beal, Variational Algorithms for Approximate Bayesian Inference, Ph.D. Thesis, The Gatsby Computational Neuroscience Unit, University College London, 2003.

[7] J. Ocenasek, J. Schwarz, Estimation of distribution algorithm for mixed continuous-discrete optimization, *Proceedings of the 2nd Euro-International Symposium on Computational Intelligence*, pp. 227-232, 2002.

[8] T. Bäck, *Evolutionary Algorithms in Theory and Practice*, Oxford University Press, 1996.