잠재 고객 예측을 위한 능동 학습 기법

박상욱, 장병탁
서울대학교 컴퓨터공학부
{swpark, btzhang}@scai.snu.ac.kr

Active Learning for Prediction of Potential Customers

Sang-Wook Park Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

본 논문에서는 상거래 환경에서 구매자와 비구매자들에 대한 데이터를 학습한 후, 잠재 고객들 중에서 구매 확률이 높은 사람을 예측하는 문제에 효율적으로 접근하기 위해 능동적인 데이터 선택 기법을 이용한다. 실험 데이터는 CoIL Challenge 2000에서 얻은 데이타로서, 구매자들의 정보보다 비구매자들의 정보가 더 많기 때문에 상당히 균형이 맞지 않는다. 따라서 모든 데이터를 한꺼번에 학습하는 경우에 성능이 좋지 않다. 본 논문에서는 이러한 불균형 분포를 갖는 실제적인 문제에 있어서 RBF 기반의 신경망을 가지고 능동 학습을 함으로써 기존의 뱃치학습 보다 예측의 정확도를 향상시킬 수 있음을 보인다.

1. 서론

전통적인 신경망 알고리즘에서는 학습 데이터가 외부 환경이나 외부 실험자에 의해 모두 주어진다고 가정한다. 따라서, 신경망의 학습은 신경망의 자유 변수들의 조정에 초점을 맞추게 된다. 반면에 능동 학습에서는 학습자가 자신의 학습 데이터를 스스로 선택하거나 혹은 자신이 학습 데이터에 어떤 영향력을 행사할 수가 있다고 가정한다[1]. 일반적으로 학습의 문제는 훈련 데이터에 기반하여 그 입력과 출력 사이의 대응 관계를 찾아내는 것으로 생각할 수 있다. 이 때, 능동 학습자는 제한된 데이터 집합으로부터 새로운 입력을 반복적으로 선택하고 그 결과 값을 관찰한 후 새로운 데이터 집합을 자신의 훈련 데이터에 포함시키는 행위를 반복한다.

능동 학습에 있어 가장 중요한 문제는 새로운 데이터를 선택하는 방법이다. 그러한 방법으로는 현재 훈련 데이터가 존재하지 않는 곳에서 새로운 데이터를 선택하는 방법, 성능이 나쁘거나 신뢰성이 떨어지는 곳에서 데이터를 선택하는 방법, 현재의 모델을 변경시킬 수 있을 만한 데이터를 선택하는 방법, 총괄적인 분산이 최소화될 수 있는 데이터를 선택하는 방법 등이 있다. 한편, 결정적인 알고리즘이 지역 최적화에 빠지는 문제가 있기 때문에 확률적인 능동 학습 방법도 제안되었다[2,3,4,5].

본 논문에서는 현재까지 학습된 신경망에서 학습이 가장 어려운 데이터를 새로운 학습 데이터로 선택하는 방법을 사용한다. 이는 은닉 뉴런을 스스로 증가시키는 RBF 신경망인 ARAN(Active RAN)으로 구현되었다[6]. 실험 데이터는 CoIL Challenge 2000에서 사용된 것으로 이동식 주택을 위한 보험의 잠재 고객에 관한 정보로 이루어져 있고, 구매자들의 정보보다 비구매자들의 정보가 더 많은 불균형 데이터 집합이다. 따라서 이 데이터는 뱃치학습 방법으로는 성능이 좋지 않기 때문에 능동 학습 기법을 이용하여 예측의 정확도를 향상시키고자 한다.

2장에서는 ARAN 알고리즘에 대해서 기술하고, 3장에서는 ARAN을 이용해 COIL Challenge 2000 데이터 집합에 대한 능동 학습의 실험 결과를 보인다. 끝으로 4장에서는 이 논문의 결론 및 앞으로의 과제에 대해 기술한다.

2. ARAN

ARAN 알고리즘은 Platt이 제안한 RAN 알고리즘에 능동 학습 기법을 추가한 것이다[7]. 따라서 그림 1과 같이 은닉 뉴런층이 하나인 RBF 신경망의 모습을 가진다[9]. 여기서 x는 입력 데이터이고, c는 은닉 뉴런의 중심값, w는 은닉 뉴런의 너비가 되며, h는 은닉 뉴런과 출력 뉴런 사이의 가중치가 된다. 그림에서는 입력의 차원이 m이고, 은닉 뉴런의 개수가 k인 신경망을 보여준다. k

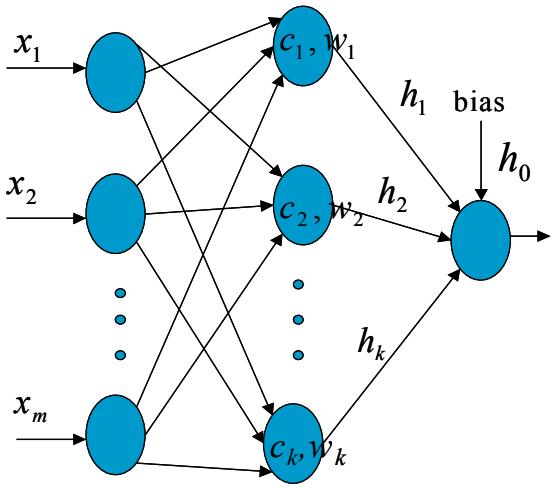

그림 1 : RBF 신경망의 일반적인 구조

의 값은 학습이 진행되는 동안 일정 수준까지 계속 늘어나게 되어, 적당한 은닉 뉴런의 개수를 찾아줄 수 있도록 한다. 전체적인 알고리즘의 구성은 그림 2에 순서도로 표현되어 있다.

알고리즘을 세부적으로 살펴보면, 먼저 은닉 뉴런이 없는 임의의 초기 신경망을 구성한다. 학습에 사용할 데이터는 학습에 사용되지 않은 데이터로 이루어진 후보 집합과 학습에 사용된 훈련 집합으로 나누어 지는데, 후보 집합의 모든 데이터에 대한 에러를 계산한다. 그 후에 가장 큰 에러를 가지는 데이터를 일정 개수 선택하고, 그 에러의 크기가 크고, 기존 은닉 뉴런과 선택된 데이터 사이의 거리가 큰 경우에 은닉 뉴런을 생성한다. 그렇지 않은 경우에는 에러를 최소화시키는 학습만 수행한다. 선택된 데이터는 훈련 집합에 추가한다. 다시 훈련 집합의 데이터로 학습된 신경망에 후보 집합의 데이터들에 대해 에러를 다시 계산한 후 가장 큰 에러를 갖는 데이터를 선택한다. 이 과정을 반복하는 것이 전체적인 학습 알고리즘이다. 자세한 알고리즘 및 그에 따른 수식은 [6]에 나와 있다.

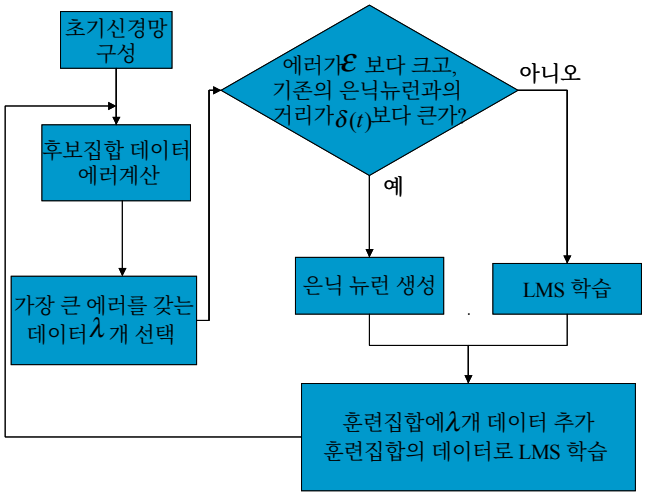

그림 2 : ARAN의 알고리즘

3. 실험 및 결과

실험은 Coil challenge 2000 데이터를 이용한다[8]. 이 데이터는 실제 비즈니스 데이터로서 이동식 주택을 위한 보험의 잠재 고객에 관한 정보로 이루어져 있다. 하나의 잠재 고객에 대한 전체 86개의 속성이 있고, 그중 43개가 고객의 우편 번호를 통해 알아낼 수 있는 사회 인구 통계 정보이고, 나머지는 이 보험 상품 구매와 관련된 상품의 구매 정보이다. 학습 데이터는 5822명의 구매자와 비구매자로 이루어져 있고, 테스트 데이터는 4000명의 잠재 고객으로 이루어져 있다.

한편 이 데이터는 전체 데이터의 약 6%정도만 실제 고객으로 그 분포가 매우 불균형적으로 되어있다. 테스트 데이터 4000명중에서 오직 238명만 고객이므로 고객인지 아닌지를 정확히 판별하는 것은 매우 어려운 문제이다. 이에 실제 CoIL Challenge 2000에서는 4000명의 잠재 고객 중 고객이 될 확률이 높은 800명을 선택해서 이들 중 몇 명이 실제 고객인지를 측정함으로써 우승자를 결정했다. 이는 어떤 회사가 상품 광고 메일 보내는 경우에, 원하지 않는 고객에게 메일을 보내는 것을 피하고, 구매할 가능성이 높은 고객에게만 메일을 보내려고 하는 경우에 유용할 수 있다. 한편, 가장 좋은 선택은 800명 중에서 실제 고객 238명을 전부 예측하는 것이지만, 실제로는 121명을 예측한 참가자가 우승을 하였다. 임의로 800명을 선택한다면 평균 42명이 포함되어 있을 것이다. 아래 표 1에 대회 결과가 나와있다.

표 1 : CoIL Challenge 2000 결과

방법	결과(전체 800명중 실제 고객수)
최적의 방법	238
임의선택 방법	42
1등	121
2등	115
3등	112

실험은 다음과 같이 두가지로 하였다. 첫째, 전체 85개의 입력 속성을 모두 이용하여 능동 학습을 이용하는 경우와 그렇지 않은 경우에 대해서 실험을 하였다. 이 실험을 통하여 능동 학습을 이용하여 성능이 좋아짐을 보이고자 했다. 둘째, 주어진 모든 입력 속성을 이용하는 대신에, 가장 중요한 속성으로 알려진 4개의 속성만을 이용해서 앞에서와 같은 방법으로 실험을 하였다. 이 실험을 통하여 CoIL Challenge 2000 데이터처럼 복잡한 문제에서 데이터 특징 추출이 중요함을 보이고자 했다. 표 2에서는 이러한 실험의 결과를 보여준다.

표 2를 보면 예상한 바와 같이, 85개의 속성을 사용한 경우와 4개의 속성을 사용한 경우 모두 능동 학습 방법을 이용했을 때, 더 좋은 성능을 보인다. 이는 실제 구매자에 대한 정보가 훨씬 적은 불균형한 데이터에서 실제 구매자에 대해 더 가중치를 줌으로써 더 좋은 성능을 보이는 것으로 보인다.

표 2 : 실험 결과

사용된 속성 수	결과(전체 800명중 실제 고객수)	
	능동 학습 방법	뱃치 학습 방법
85	110	104
4	118	109

한편, 85개의 속성을 전부 사용한 경우보다, 오히려 더 적은 4개의 속성을 사용한 경우가 능동 학습 방법과 뱃치 학습 방법에서 모두 더 좋은 성능을 보인다. 이는 잠재 고객 중에서 실제 고객을 예측하는 문제에서 85개의 속성 전부가 문제 해결에 관련이 있는 것이 아님을 보여준다. 즉, 오히려 더 많은 정보를 가지고 학습하는 것이 그 안에 노이즈가 들어갈 수 있으므로 예측 성능에 좋지 않은 결과를 나타낸다. 따라서 CoIL Challenge 2000 데이터와 같이 많은 속성을 가지는 데이터를 학습하는 경우에는 학습 모델을 최적화시키기 이전에 데이터의 중요한 특징을 파악하는 것이 중요할 수 있다. 표 1과 표 2에서의 성능 차이는 데이터의 중요한 특징 파악 능력의 차이라고 볼 수 있다.

그러나 본 논문에서 실험한 결과는 데이터의 특징 파악없이도 CoIL Challenge 2000 입선자들의 결과와 많은 차이를 보이지 않는데, 이는 본 논문의 실험이 이미 테스트 데이터에 대한 답을 알고 이루어졌기 때문이다. 즉, 이미 답을 알고 있었기 때문에, 좋은 성능을 내는 쪽으로 학습을 진행시킬 수 있었다. 실제 대회에 참가했을 경우에는 본 논문의 결과보다 성능이 더 좋지 않았을 것이다.

4. 결론

본 논문에서는 상거래상에서 얻을 수 있는 잠재적 고객 데이터에 대해서 실제 구매자를 예측하는 실험을 하였다. 이 데이터는 실제 구매자보다는 비구매자들이 훨씬 많이 포함되어 있는 불균형 데이타이기 때문에 단순히 데이터를 한꺼번에 학습하는 경우에 어려움이 있다. 여기서는 능동학습 방법을 적용함으로써 예측의 정확도를 향상시킬 수 있음을 보였다.

그러나 실제 CoIL Challenge 2000의 데이터처럼 그 속성이 많으면서도, 상당히 불균형한 데이터에서는 단순한 학습 알고리즘을 사용하는 것보다 그 데이터의 속성에 관한 특징을 파악한 후에 학습을 하는 것이 더 중요하다. 실제로 CoIL Challenge 2000 결과 보고서의 대부분이 데이터의 속성중에서 중요한 속성을 골라내는 것에 치중을 하였으며, 속성 파악을 잘한 참가자가 좋은 성적을 낼 수 있었다. 우리의 실험 결과에서도 주어진 입력 속성 85개를 다 사용하는 것보다 중요한 4개의 속성을 사용한 것이 결과가 더 좋아지는 것을 보면, 데이터의 특징 파악이 중요함을 알 수 있었다. 따라서 더 좋은 예측 성능을 얻기 위해서는 이러한 데이터들에 대해 그 특징을 잘 파악할 수 있는 연구가 이루어져야 할 것이다.

한편 본 논문에서 사용한 알고리즘에서는 데이터를 능동적으로 학습을 하는 경우에 합리적인 종료 알고리즘이 없다. 따라서 모든 데이터를 끝까지 사용한 다음에 가장 좋은 성능을 나타내는 것을 결과로 사용했기 때문에 문제가 생긴다. 능동 학습 알고리즘의 안정성을 위해 종료 조건에 대한 연구를 하는 것이 앞으로의 연구 과제이다.

감사의 글

본 연구는 학술진흥재단 자유공모과제(1999-001-E01025)와 첨단정보기술연구센터(AITRC)에 의하여 일부 지원되었음.

참고 문헌

[1] M. Plutowski and H. White, Selecting concise training sets from clean data, *IEEE Transactions on Neural Networks*, vol. 4, pp. 305-318, 1993.

[2] David A. Cohn et al, Active Learning with Statistical Models, *Journal of Artificial Intelligence Research 4*, pp. 129-145, 1996.

[3] Kenji Fukumizu, Statistical Active Learning in Multilayer Perceptrons, IEEE Transactions on Neural Networks, Vol. 11, No.1 , pp. 17-26, 2000

[4] B.T. Zhang, Learning by Incremental Selection of Critical Examples, *Arbeitspapiere der GMD, No 735, German National Research Center for Computer Science (GMD), St. Augustin/Bonn,* 1993.

[5] B.T. Zhang, Accelerated Learning by Active Example Selection, *International Journal of Neural Systems 5(1)*, pp. 67-75, 1994.

[6] 박상욱, 장병탁, 능동적인 데이터 선택에 의한 RBF 신경망의 학습, *한국정보과학회 봄 학술발표 논문집 (B)*, 제27권 1호, pp. 478-480, 2000.

[7] John Platt, A Resource-Allocating Network for Function Interpolation, *Neural Computation Vol 3 No. 2,* pp. 213-225, 1991.

[8] http://www.dcs.napier.ac.uk/coil/challenge

[9] Lu Yingwei et al, Performance Evaluation of a Sequential Minimal Radial Basis Function (RBF) Neural Network Learning Algorithm, *IEEE Transactions on Neural Networks Vol 9 No.2,* pp. 308-318, 1998.