제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

이동형 단말기를 이용한 길 안내 SOM의 학습 알고리즘을 이용한 데이터마이닝 응용에 관한 연구

이대영, 배상현, 정명진, 송병호
조선대학교 전산통계학과
e-mail : cssna01@hanmail.net

A Study On Application Of Data Mining Using SOM Studying Algorithm

Dae-young Lee, Sang-hyun Bae, Myong-jin Jung, Byoung -ho Song
Dept of Computer Science and Statistics, Chosun University

요약

본 논문에서는 실제 경영의 의사결정 등을 위한 활용가치가 있는 정보를 추출해 내는 방법론으로 SOM을 적용하였다. SOM은 자율(upsupervised)과 경쟁(competitive) 학습을 한다. 데이터를 입력하였을 때, SOM의 출력 노드중에서 다른 출력 노드과 비교해서 가장 강하게 반응하는 노드가 있을 것이며, 그러한 출력 노드를 더욱 더 강하게 반응하게끔 반복적으로 학습시키는 것이다. 입력에 대해 자연스럽게 반응하는 출력 노드를 선택하여 반복 학습을 시켜면, 후에는 결과적으로 어떤 출력 노드가 반응되는지를 조사하면 거꾸로 입력을 알 수 있게 되는 것이다. 대량의 데이터, 잠재적으로 활용가치가 있는 데이터를 SOM을 통해 유용한 정보들을 추출할 수 있으며 이는 실제 경영의 의사결정을 위한 수단으로 충분히 활용될 수 있을것이다.

1. 서론

인간의 대뇌 피질은 각각의 기능을 맡는 구역(region)이 정해져 있으며, 이는 생각하고, 말하고, 듣고 그리고 판단하는 인간의 중요한 기능을 처리한다. 뇌를 다쳐서 다른 건 멀정한데 과거의 기억을 잃어 버린다던가, 혹은 말만 못한다던가 하는 것, 이런 이유는 바로 대뇌피질의 각 구역마다 맡은 기능이 존재하기 때문이다. 이러한 특징을 Ordered Feature Maps이라 하며, SOM은 바로 이러한 대뇌피질을 닮았다. SOM은 신경망(neural network)의 일종이며, 자율(upsupervised)학습과 경쟁 학습 방법을 이용한다. 또 BPN(Back Propagation Network)과는 다르게, 마치 대뇌피질의 구역이 의미가 있는 것처럼 출력 노드의 위상(topology)자체가 의미가 있다. 간단하게 BPN과 비교를 해보자. 숫자를 인식할 때 예를 들어보자. BPN은 이미지 데이터를 입력으로, 그리고 그 이미지 데이터가 나타내는 숫자를 출력으로 하여 입력, 출력의 쌍으로 하여 지도 학습(supervised leaning)을 시킨다. 하지만 SOM은 그렇지 않다. 자율(upsupervised)과 경쟁(competitive) 학습을 한다. 이미지 데이터를 입력하였을 때, SOM의 출력 노드중에서 다른 출력 노드과 비교해서 가장 강하게 반응하는 노드가 있을 것이며, 그러한 출력 노드를 더욱 더 강하게 반응하게끔 반복적으로 학습시키는 것이다. 다시 말해, SOM에서는 BPN처럼 출력노드를 미리 기능적으로 정하지 않는다는 말이다. 입력에 대해 자연스럽게 반응하는 출력 노드를 선택하여 반복 학습을 시켜면, 후에는 결과적으로 어떤

출력 노드가 반응되는지를 조사하면 거꾸로 입력을 알 수 있게 되는 것이다.

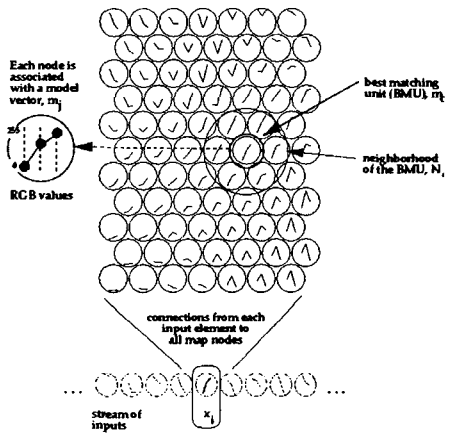
2. SOM 알고리즘

입력을 $x(t)$라 하고, 출력 노드의 가중치를 $m_i(t)$, 모델 벡터(model vector)라고 하였다. 그리고 SOM 알고리즘을 아래와 같은 두 단계로 표현하였다. 물론 초기에 $m_i(t)$는 랜덤값이다.

1. 입력 $x(t)$에 대해 가장 잘 반응하는 $m_i(t)$를 선택한다. 다시 말해, $x(t)$와 가장 비슷한 벡터를 선택한다. 그리고 가장 반응하는 출력노드를 winner라고 하였으며, $m_i(t)$는 winner
의 가중치 벡터인 셈이다.

2. winner와 그 주위 노드(neighboring nodes)들을 학습시킨다. 【그림 1】을 자세히 살펴보자. 그림 아래 부분에 입력 스트림이 있고, 그리고 원 안에 BMU(Best Matching Unit)라 하여 가장 잘 반응하는 출력 노드가 있다.

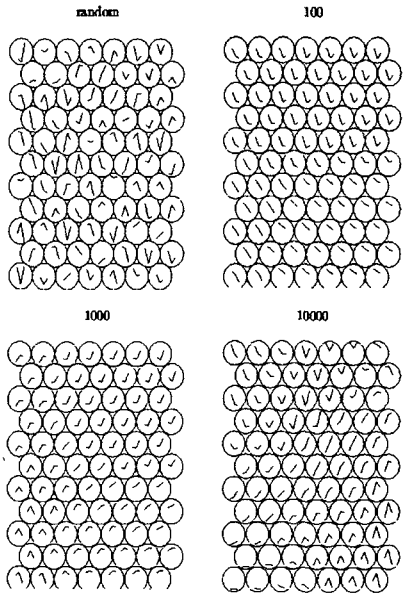
이는 SOM에 대해 Timo Honkela가 설명한 부분이다.



【그림 1】 SOM의 과정, 가장 잘 반응하는 BMU와 그 주위 노드

단계 1, 2를 반복한 것이 【그림 2】이다. random은 초기 상태이고 각각 100번, 1000번, 10000번씩 반복 학습시켰을 때 나타나는 출력 노드들의 가중치 값들이다. 즉 $m_i(t)$들의 값이다.

SOM의 알고리즘을 간단히 요약하면, 특정한 입력에 대해 가장 잘 반응하는 출력 노드와 그 주위의 노드를 학습시킴으로써 자연스럽게 그 입력에 대한 대표자(representative)로써의 역할이 가능하게끔 하는 것이다. 대뇌 피질을 예를 들어 설명해보자. 인간은 수많은 정보를 받아들이고 처리하게 되는데. 수많은 정보들중에 특정 정보를 잘 받아들이는 뇌 세포가 있을 것이며, 나중에는 신경세포의 연결 강도가 변하게 되며, 또한 주위의 신경세포도 변함으로써 대뇌피질의 어느 한 구간으로 자리잡게 된다. 그럼으로써 인간은 입력되어지는 수많은 정보들을 구별한다. 수많은 입력값으로부터 특정 정보를 구별하는 것과 비슷하다.



【그림 2】 단계마다의 출력 노드 가중치값 변화

SOM의 학습 알고리즘

$m_i(t+1) = m_i(t) + \alpha(t)[x(t) - m_i(t)]$ for each $i \in N_c(t)$,

$m_i(t+1) = m_i(t)$ otherwise

단, $x(i)$는 입력,

$m_i(t)$는 출력 노드의 가중치(모델 벡터),

$\alpha(t) \in [0,1]$는 학습률이다. $N_c(t)$는 winner의 주위 노드들이다.

BPN의 가중치 학습과는 상이한 모습이다. BPN처럼 출력과의 오류 정보를 이용하지 않고 입력 정보만이 이용되기 때문이다. 입력에 따라 가장 잘 반응하는 출력 노드와 그 출력 노드의 주위 노드($N_c(t)$)

들을 선택하고, 학습률에 비례하여 가중치(모델 벡터)를 변화시킴으로써 학습하는 것이다.

3. Data mining

데이터마이닝이란 대량의 데이타로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 데이터마이닝의 개념은 정보기술의 발달과 비즈니스적 요구에 의해 시장에 등장하게 되었다. 이를 살펴보기 전에 먼저 정보시스템의 발전과정을 살펴보는 것이 좋겠다. 정보화의 초창기에 EDPS라는 개념이 한때 유행을 하다가 MIS의 개념으로 옮겨지게 되었다. 경영층의 의사결정에 도움을 주는 고급정보를 가공하고 축적하는데 관심을 가졌던 의사결정지원시스템(DSS : Decision Support System)은, 정보화의 개념을 조직의 하부계층의 반복업무를 지원하는 자동화 업무에서 전사적인 개념으로 확장시키는 역할을 하였다. 그런데 이를 구축하기 위하여 선결과제가 발견되었는데 바로 전사적인 시스템을 통합관리하는 통합데이타베이스의 구축이었다. 각 부서별로 독립적으로 운영되는 시스템으로는 경영층이 의사결정을 내리는데 별로 도움이 되지 않았던 것이다. 통합 데이터베이스 구축이 어느 정도 이루어졌을 때 발생한 또다른 문제점은, 방대한 데이터와 정보들 가운데서 찾고자 하는 정보를 정확하고 빠르게 찾는다는 것이 아주 힘들다는 점이다. 결국 이를 해결하고자 하는 노력을 등장한 개념이 바로 데이터마이닝, 데이터웨어하우징 등의 개념이다. (데이터웨어하우징이란 대용량의 데이터베이스를 실제 업무에 있어서 활용도를 높이기 위해 데이타를 좀더 정제되고 일관성있게 통합된 형태로 쌓아두고자 하는 시도이다.) 이러한 문제는 실제 비즈니스에서 더욱 개선의 필요성으로 등장하였다. 즉, 고객, 상품, 경쟁사 관련 데이터 등 기업이 얻고자 하는 정보를 보다 손쉽게 접근할 수 있고 효과적으로 활용할 수 있도록 하는 도구가 필요하게 된 것이다. 점점 심해지고 있는 기업 경쟁의 상황에서 더욱 다양해지고 개성화되고 있는 고객들의 필요를 만족시키기 위하여는 그런 요구에 대한 빠른 대응이 요구되었고 이것이 기업간의 경쟁력 척도가 되었다. 또한 지속적으로 경쟁우위를 확보하기 위하여는 효과적이고 합리적인 신속한 의사결정이 더욱 중요하게 되었다. 따라서 기업들의 관심은 데이터를 잘 쌓아놓는 단계에서 벗어나 방대한 데이터의 창고에서 보다 가치있는 정보를 효과적으로 신속하게 찾아내고자 하는 방법으로 모아지고 있는 것이다.

4. SOM을 이용한 데이터마이닝

데이터마이닝이란 묵시적이고 잘 알려져 있지 않지만 잠재적으로 활용가치가 있는 정보, 즉 대량의 데이터로부터 쉽게 드러나지 않는 유용한 정보들을 추출하는 과정을 말한다. 다시말해 기업이 보유하고 있는 일일 거래자료, 고객자료, 상품자료, 마케팅 활동의 피드백 자료와 기타 외부자료를 포함하여 사용가능한 데이타를 기반으로 마케팅을 기획·수립하게 되고 본격적으로 상품을 시중에 선보이게된다. 하지만 숨겨진 정보, 기대하지 못했던 패턴, 새로운 법칙과 관계들로 인해 기대 이상 혹은 기대 이하의 예상치 못했던 결과물들을 얻게된다. 본 논문에서는 실제 경영의 의사결정 등을 위한 활용가치가 있는 정보를 추출해 내는 방법론으로 SOM을 적용하였다. SOM은 자율(upsupervised)과 경쟁(competitive) 학습을 한다. 데이터를 입력하였을 때, SOM의 출력 노드중에서 다른 출력 노드과 비교해서 가장 강하게 반응하는 노드가 있을 것이며, 그러한 출력 노드를 더욱 더 강하게 반응하게끔 반복적으로 학습시키는 것이다. 입력에 대해 자연스럽게 반응하는 출력 노드를 선택하여 반복 학습을 시켜면, 후에는 결과적으로 어떤 출력 노드가 반응되는지를 조사하면 거꾸로 입력을 알 수 있게 되는 것이다. 대량의 데이터, 잠재적으로 활용가치가 있는 데이터를 SOM을 통해 유용한 정보들을 추출할 수 있으며 이는 실제 경영의 의사결정을 위한 수단으로 충분히 활용될수 있을 것이다.

5. 참고문헌

[1] Usama. M Fayyad et al., "Advances in knowledge discovery and data mining", MIT Press, 1996

[2] Tom M. Mitchell, "Machine Learning and Data Mining", Communications of the ACM, Vol. 42, No. 11, 1999

[3] William J. E. Potts, "Generalized Additive Neural Networks", SAS Institute Inc., ACM KDD-99, 1999

[4] Kohonen, T., "Self-Organizing Maps", Berlin: Springer-Verlag. Second edition, 1997

[5] J.P.Bigus, Data Mining with Neural Networks, McGraw-Hill, 1996