인공신경망분석의 이해와 활용

Constructs and Application of Artificial Neural Networks(ANN)

유제민 (강남대학교, pavlov2@kangnam.ac.kr)

초록 데이터마이닝(Data Mining)은 방대한 양의 자료 속에서 의미 있는 패턴과 규칙을 찾아내기 위해 자동적이거나 반자동적인 방식으로 자료를 탐색하고 분석하는 것이다. 즉, Data간의 관계, 패턴, 규칙 등을 찾아내고 모형화 한 후 최적의 의사결정을 목표로 스스로 정보를 변환하는 수학적 기법이다. 마이닝에서는 각 개인이 시시각각 발생시키는 중요한 정보들을 컴퓨터로 분석한 후 패턴을 실시간으로 분석한다. 데이터 마이닝의 프로그램들이 개발되고 다양한 학문 분야에 적용되기 시작하면서 21세기에 세상을 바꿀 10대 발명품 중에서도 가장 주목받는 상품으로 언급되기도 하였다. 본 워크샵에서는 신경망 데이터 마이닝 프로그램인 Neural Connection을 중심으로 통계적인 논리 사용법을 숙지한다.

본론: 전통적인 통계 분석에서는 통계적 이론 모형이 설정되고 그에 따라 분석이 이루어지는 *하향적* 방식인데 비해 데이터 마이닝은 통계적 가정이나 전제 없이도 자료 속에서 자료구조의 패턴을 바탕으로 새로운 예측이나 설명모형을 만들어 내는 *상향적*(BOTTOM-UP)방식이라는 기본적 차이가 있다. 즉, 일반적인 통계와 데이터 마이닝은 접근하는 방식이 상반된다. 무엇보다도 데이터 마이닝은 정상성이나 선형성 그리고 등분산성 같은 통계적 가정에 좌우되지 않는다. 일반적인 통계방법에서는 척도(scale)가 혼재된 변인들을 처리하는데 한계가 있다. 가(dummy) 변수를 만들거나 척도를 미리 통일하는 과정에서 중요하지만 변환시키기 어려운 자료들을 버리는 경우도 있을 수 있다. 이렇게 가공한 변인들이 정상성을 크게 충족시키지 못하면 일반화기도 어렵다. 따라서 다양한 척도의 다변인 자료를 동시에 처리할 때 모수적 가정 고집하는 것은 비현실적이다. 이런 점에서 본다면 데이터 마이닝은 비 모수적(non-paramatic) 기법에 가깝다.

연구자들은 연구 설계에 맞는 통계를 적용시키려고 노력한다. 가설설정-설계-통계로 이어지기 때문에 하나라도 수정되면 나머지도 바꾸어야 한다. 연구자는 이 과정을 능숙하고 빠르게 시행하는 전문가이다. 그러나 신경망 분석(Neural Networks)에서는 이 과정을 스스로 수행한다. 신경망 모형은 인간두뇌의 학습과정을 시냅스 작용을 통한 연산과정으로 재현한 데이터 마이닝 기법의 일종이다. 이 방법은 실제 출력 값과 기대 출력 값을 비교하고 그 차이를 시냅스 역할을 하는 노드(node)에 가중치를 주어 조정한 후 그 과정을 신경망 구조가 안정화될 때까지 반복함으로써 예측 및 분류 모형을 자가 창출한다. 스스로 피드백을 주어가면서 수정하는 역-전파망 방식(Back-Propagation Network)으로 경험을 통한 학습 과정을 기계적으로 단축시킨 것이다. 이에 따르면 연구자들의 가장 중요한 능력은 변인들의 선정과 관련 데이터의 확보 그리고 산출된 결과의 해석과 그 응용력이다.

기본적인 통계로도 충분히 분석이 가능한 주제들에 관해 데이터 마이닝을 할 필요는 없다.

데이터 마이닝을 적용할 때 고려해야 하는 사항들이 있다. 첫째, **데이터의 사용이 용이한가(available)** 확인해야 한다. 많은 자료들을 모으는 과정에서 각기 다른 장소에 다른 형식(format)으로 보관되어 있거나 컴퓨터 파일의 형태가 아닌 방식으로 입력(coding)되어 있는 경우가 자주 있다. 데이터를 통일하기 전에 각 자료의 장/단점을 파악하고 이를 토대로 분석 방법을 미리 생각해야 한다. 둘째, **데이터가 풍부한 설명요소들(factors)을 가지고 있는지** 확인해야 한다. 마이닝의 결과가 효과적이기 위해서는 보유한 데이터가 다양하고 중요한 설명요소들로 이루어져야 한다. 물론 유용한 설명요소들을 확인하는 것이 마이닝의 목적이 될 수 있지만 데이터 자체가 어떤 문제에 대한 실마리가 있으면 보다 다양한 기법을 동원할 수 있다. 셋째, **데이터에 심각한 noises가 있는지** 확인해야 한다. 노이즈는 오차의 집합체라고 할 수 있다. 클레멘타인(Clementine)의 기계학습 시스템에서는 최고 50%까지 노이즈를 조정할 수 있고 Neural Connection에서도 오차 조정 옵션이 있다. 그러나 결측치나 요소의 속성 변화가 많을 때 모형의 타당성 자체가 저하되는 것은 당연하다. 넷째, **데이터의 양이 충분한가(enough)** 점검해야 한다. 마이닝에서는 데이터의 양이 적을 때 이를 고려하는 방안이 있다. 따라서 데이터의 양이 적다는 것은 마이닝을 할 수 없다는 것이 아니라 결과의 해석이나 신뢰도에 문제가 발생한다는 것이다. 특히 데이터의 패턴이나 상관관계가 복잡하다고 예측될 경우 표집은 많을수록 바람직하다.

또 하나 주의해야 할 사항은 고객 데이터와 프라이버시(privacy)에 관한 문제이다. 프라이버시 문제는 항상 존재해왔으나 CRM시스템이 점차 강력해지고 데이터 마이닝과의 결합이 이루어짐에 따라 점점 더 중요한 문제로 부각되고 있다. 미국의 경우 데이터 마이닝을 비즈니스에 사용하면서 법적인 문제들이 드러나기 시작했다. 만약 기업들이 인종이나 성별 그리고 출신지 등을 차별대우적 의사결정에 활용하는 것이 일상적이고 당연한 것이라면 어떻게 될 것인가? 인터넷을 통한 개인 정보의 유출은 우리나라에서도 심각한 수준에 이르고 있다. 미국에서는 공정한 정보관행(fair information practice)이 법률에 의거해 다각적으로 시행되고 있지만 Web Mining과 Data Farming 등의 기법이 확산되면서 프라이버시 문제와 끊임없이 충돌하고 있다. 현재 신원에 대한 정보보다는 **익명성**의 정보를 사용하도록 하고 상세 데이터보다는 **통합 데이터**를 활용하며 목표 선정보다는 **측정을 위한 공공 정보**로 데이터 마이닝을 권장하고 있다. 데이터 마이닝의 한계는 그 자체의 기법적인 제약이 아니라 사생활과 인권 그리고 법률적인 규제에 달려 있다.