범주형 자료에 대한 데이터 마이닝 분류기법 성능 비교

손소영* · 신형원*

* 연세대학교 산업시스템공학과

Abstract

본 연구에서는 범주형 자료의 특성에 따라 세가지 데이터 마이닝 기법(신경망, Decision Tree, 로지스틱 회귀분석)의 분류성능을 비교하였다. 이를 위하여 범주형 자료의 특성을 나타내는 네가지 인자((1)입력변수의 범주별 비율, (2)입출력변수간의 함수, (3)error의 크기, (4)출력변수의 범주별 비율)를 바탕으로 모의자료를 만들고 이를 세가지 데이터 마이닝 기법(신경망, Decision Tree, 로지스틱회귀분석)에 적용하여 분류정확성을 분석하였다. 분석결과 네가지 인자중 함수의 종류와 에러의 크기, 출력변수의 범주별 비율, 데이터 마이닝 기법이 α=0.05에서 유의하게 분류정확성에 영향을 미치는 것으로 나타났다. 범주형 자료의 특성에 따라 적합한 데이터 마이닝 기법을 찾기 위하여 교호작용을 통하여 Duncan 검정한 결과, 함수의 종류가 간단할 때는 Decision Tree, 함수의 종류가 복잡할 때 로지스틱 회귀분석이 적합한 것으로 나타났으며 에러의 크기가 작을 때는 로지스틱 회귀분석, 에러의 크기가 클 때는 신경망이 적합한 것으로 나타났다. 한편, 4가지 인자중 입출력변수간의 함수관계는 주어진 자료에서 현실적으로 알 수 없는 성격이므로 다구찌 디자인을 이용하여 비제어 인자로 간주하고 실험한 Duncan 검정결과, 출력변수의 범주별 비율이 어느 한 쪽으로 치우치지 않았을 경우 신경망과 Decision Tree의 분류정확성이 로지스틱 회귀분석에 비하여 높은 것으로 나타났다. 이상의 실험에 사용된 일부요인 실험계획(Half Fractional Factorial Design)에서 검정하려는 효과는 다른 효과와 교락(Confounding)되지 않는 것으로 나타났다. 향후 연구방향으로, 본 연구와 같이 모의 자료를 이용하여 데이터 마이닝 기법의 분류성능을 측정한 연구를 취합하여 메타분석을 함으로써 좀 더 다양한 인자와 수준을 고려한 데이터 마이닝 기법의 성능 비교를 제시 하고자 한다.