데이터 마이닝을 이용한 고혈압 요인분석

이제영[1], 이용원[2], 사공준[3], 이윤수[4]

요약

데이터 마이닝을 이용하여 20대 이상의 사람들을 대상으로 남녀간의 고혈압 발병여부에 따른 고혈압 위험요인에 대하여 분석하였다. 분석된 위험요인들의 개별적인 영향력을 알아보고, 이를 바탕으로 남녀간의 고혈압 발병여부에 대하여 적절한 모형을 만들었다.

keyword : 데이터 마이닝, 고혈압

1. 서론

컴퓨터와 네트워크의 엄청난 발전으로 데이터베이스를 만들어 많은 양의 자료를 저장, 보관하게 되었다. 여러 분야에서 저장, 보관되는 대량의 데이터를 효과적으로 분석하여 유용한 정보를 획득하기 위해 새로이 등장한 것이 데이터 마이닝이다. 이러한 데이터 마이닝 기법을 이용하여 다양한 종류의 데이터의 기초적인 특성을 바탕으로 모델링하여 새로운 데이터가 들어왔을 때 최적의 분류를 가능하게 했으며(Hussein A. Abbass 등, 2004), 웹 페이지 분석을 통한 성별과 관심분야에 대한 웹 페이지 성향 파악이나(Baglion 등, 2004), 두 개의 서로 다른 분포에서 나온 데이터가 섞여있을 때 데이터 마이닝의 Neural Network 기법을 이용하여 올바른 판별을 할 수 있게 되었다(이, 2003). 또한, 주식의 지수예측 모형을 만들어 주식현물시장에서 개별주식을 예측하거나(이, 2003), 특정 병을 가진 환자를 대상으로 한 약물의 유효성 평가에 영향을 주는 변수 선택(전, 2003) 등 기술, 경제적인 분야뿐만 아니라 의학에서도 이용될 만큼 데이터 마이닝의 분야는 매우 광범위하다. 본 논문은 서울 소재 건강검진센터에 건강검진을 위하여 온 사람들 중 20세 이상의 사람들을 대상으로 데이터 마이닝을 이용하여 고혈압 발병여부에 대한 고혈압 위험요인을 찾아내어 향후 고혈압의 조기 발견 및 고혈압의 예방과 관리를 위하여 도움이 될 수 있도록 하였다.

1) Professor, Department of Statistics, Yeungnam University, 214-1 Daedong, Kyungsan, Kyungbuk, 712-746, South Korea
E-mail : jlee@yu.ac.kr

2) Graduate, Department of Statistics, Yeungnam University, 214-1 Daedong, Kyungsan, Kyungbuk, 712-746, South Korea

3) Associate Professor, College of Medicine, Yeungnam University, 317-1 DaeMyoung-dong, Namgu Daegu, 705-717, South Korea

4) Graduate, Department of Statistics, Yeungnam University, 214-1 Daedong, Kyungsan, Kyungbuk, 712-746, South Korea

2. 연구 대상 및 내용

2003년 1월 1일부터 12월 31일까지 1년간 서울 소재 모 종합건강검진센터에 건강검진을 위하여 온 사람들 중 20세 이상의 39,900명을 대상으로 분석을 실시하였다.

면담 조사시 고혈압 과거력, 고혈압 가족력, 과거 고혈압 약 복용여부, 과거 당뇨약 복용 여부, 흡연, 음주, 운동습관, 성별, 연령, 신장, 체중, BMI지수, 총 콜레스테롤, HDL 콜레스테롤, 중성지방, 혈당, 혈압 측정치에 관한 내용이 포함되어 있다.

고혈압 환자를 1차 고혈압군과 2차 고혈압군으로 분류하였다. 1차 고혈압 환자군의 정의는 고혈압 약을 복용한 적이 있으며, 이완기 혈압 90mmHg이상 또는 수축기 혈압 140mmHg이상인 경우로 하였다. 그리고, 정상군은 고혈압 약을 복용한 적이 없으며, 이완기 혈압 90mmHg미만 또는 수축기 혈압 140mmHg미만인 사람들로 하였다. 2차 고혈압 환자군의 정의는 고혈압 약을 복용한 적이 있으며, 이완기 혈압 100mmHg이상 또는 수축기 혈압 160mmHg이상인 경우로 하였다. 그리고, 정상군은 고혈압 약을 복용한 적이 없으며, 이완기 혈압 100mmHg미만 또는 수축기 혈압 160mmHg미만인 사람들로 하였다.

3. 데이터 마이닝을 이용한 고혈압의 위험요인 분석

고혈압의 위험요인을 분석하기 위하여 데이터 마이닝을 이용하며, 신경망과 의사결정나무분석방법을 사용하였다. 전체 데이터는 39,900명이며, 남녀간에 고혈압 발생 위험요인의 차이가 있는 것으로 생각되어 전체 데이터를 성별로 나누었다. 종속변수를 고혈압으로 하고, 독립변수로 고혈압 과거력, 고혈압 가족력, 흡연력, 음주력, 운동습관, 나이, BMI 지수, 총 콜레스테롤, HDL 콜레스테롤, 중성지방, 당뇨 이분화의 총 11개의 변수를 이용하였다. 데이터는 Train Data와 Test Data로 나누어 데이터 마이닝 기법 중 신경망기법과 의사결정 나무분석을 실시하였다. Train Data를 이용하여 각각의 분석방법으로 만들어진 모형을 Test Data에 적용하여 평가함으로써 공정한 모형 평가가 가능하게 하였다.

남성인 경우 1, 2차 고혈압에서 Neural Network기법이 다른 기법들 보다 예측력이 더 좋게 나왔다. Neural Network기법을 이용하여 선별된 1차 고혈압의 위험요인으로는 나이, BMI, 고혈압 가족력이 가장 큰 위험요인으로 선별 되었으며, 2차 고혈압의 위험요인에도 동일하게 나이, BMI, 고혈압 가족력이 가장 큰 위험요인으로 선별 되었다. BMI 지수는 1, 2차 고혈압에 동일하게 영향을 주지만, 나이와 고혈압 가족력은 2차 고혈압에 더욱 영향을 주는 것으로 볼 수 있다. 또한, 당뇨나 중성지방, HDL 콜레스테롤은 그 위험도가 다른 요인들보다 큰 경향이 있으므로 잠재 위험요인으로 볼 수 있다. 특히, 당뇨인 경우 2차 고혈압에서, 중성지방인 경우 1차 고혈압에서 조금 더 영향은 준다고 할 수 있다.

여성인 경우 1, 2차 고혈압에서 CART기법이 다른 기법들 보다 예측력이 더 좋게 나왔다. CART기법을 이용하여 선별된 1차 고혈압의 위험요인으로는 나이, BMI 지수, 고혈압 가족력이 가장 큰 위험요인으로 선별 되었으며, 2차 고혈압의 위험요인에도 동일하게 나이, BMI 지수, 고혈압 가족력이 가장 큰 위험요인으로 선별 되었다. BMI 지수나 고혈압의 가족력은 1, 2차 고혈압에 동일하게 영향을 주지만, 나이는 2차 고혈압에서 더욱 큰 영향을 주는 것을 알 수 있다. 또한, 당뇨나 중성지방, HDL 콜레스테롤은 그 위험도가 다른 요인들보다 큰 경향이 있으므로 잠재 위험요인으로 볼 수 있다. 특히 당뇨인 경우 2차 고혈압에서, HDL 콜레스테롤인 경우 1차 고혈압에서 조금 더 영향은 준다고 할 수 있다.

4. 결론

본 논문은 고혈압의 위험요인에 대하여 데이터 마이닝을 이용하여 총 예측도가 높은 분류기법을 찾고, 고혈압의 위험요인의 중요도에 관하여 분석을 실시하였다. 남성인 경우 신경망기법이 다른 기법들 보다 총 예측도가 높았으며, 여성인 경우 CART기법이 다른 기법들보다 총 예측도가 높게 나타났다. 남성과 여성 모두 중요 위험요인으로 나이, BMI 지수, 고혈압의 가족력이 나타났다. 하지만, 여성인 경우 남성보다 나이가 많이 질수록 고혈압에 대한 위험이 더 커지며, 남성인 경우 여성보다 고혈압의 가족력이 있는 경우 고혈압에 대한 위험이 더 커지는 것을 알 수 있다.

참고문헌

1. 경북대학교 의과대학 (2003). 대구광역시 고혈압·당뇨병 역학조사
2. Hussein A. Abbass, Jaume Bacardit, Martin V. Butz and Xavier Llorà (2004). Online Adaption in Learning Classifier System : Stream Data Mining.
3. M.Baglion, U. Ferrar, A.Romei, S.Ruggier, F. Turini (2004). Preprocessing and Mining Web log Data for the Personalization
4. 이성원 (2003). Logistic modelling for receiver operation characteristic curves with neural network.
5. 이수연 (2003). 통계적 기법과 인공지능기법을 이용한 데이터 마이닝의 예측성과 비교 연구 -KOSPI 200지수와 개별주식을 중심으로-
6. 전용진 (2004). 약물의 유효성 평가에서의 변수선택법 연구 -통계적 기법과 데이터 마이닝 기법의 비교-
7. 이홍주 (2001). 일부 농촌 지역 주민을 대상으로 한 고혈압 위험요인에 대한 연구