1. 데이터 분석과 빅데이터 분석의 차이점

첫번째, 데이터의 확장입니다.  전통적 데이터 분석은 조직 내부의 데이터 분석을 중심으로 이루어졌다면, 빅데이터 분석은 웹 상, SNS상의 외부 데이터까지 활용합니다. 예를 들어, 과거 기업들은 자사의 매출 분석만 시행했다면 요즘은 웹과 SNS상의 외부 데이터까지 분석해 자사 제품에 대한 소비자의 인식을 확인한다고 볼 수 있습니다.

두번째, 데이터의 다양화입니다. 전통적 데이터 분석은 정형 데이터 분석 중심으로 이루어졌지만, 빅데이터 분석은 사진, 동영상, 텍스트 모두 포함하여 비정형 데이터까지 활용한다는 점에서 차이ㄴ가 있습니다.

 세번째, 데이터의 대규모화입니다. 전통적 데이터 분석에 비해서 빅데이터 분석은 분석 대상 데이터의 규모에 큰 차이가 있습니다.

1. 데이터 전처리 기술 3가지

1. 데이터 정제 기술

1) 세분화

계층적 방법 : 각 객체를 하나의 소집단으로 간주하고 단계적으로 유사한 소집단들을 합치는 응집분석법과, 전체 집단에서부터 유사성이 떨어지는 객체들을 분리하는 분할분석법이 있다.

비 계층적 방법: 인공신경망 모델과 중심좌표 간의 거리를 산출해 가장 근접한 소집단에 배정 후 해당 소집단의 중심좌표를 업데이트 하는 k평균 군집화 방법이 있다.

2) 결측값 처리

단순 대치법 : 결측값을 그럴듯한 값으로 대체하는 통계적 기법이다. 불완전 자료는 모두 무시하는 완전 분석법, 평균값으로 대치하는 평균 대치법, 어떤 적절한 확률값을 부여한 후 대치하는 단순확률 대치법이 있다.

다중 대치법 : m번의 대치를 통해 m개의 가상적 완전한 자료를 만들어서 분석하는 방법. 대치-분석-결합 3단계로 구성되어 있다.

3) 데이터 이상값 처리

마할라노비스 거리: 데이터의 분포를 고려한 거리 측도로, 관측치가 평균으로부터 벗어난 정도를 측정하는 통계량 기법.

LOF : 관측치 주변의 밀도와 근접한 관측치 주변의 상대적인 비교를 통해 이상값을 탐색하는 기법

ESD : 평균으로부터 3시그마 이상 떨어진 값을 이상값으로 판단

카이제곱 검정 : 데이터가 정규 분포를 만족하나, 자료의 수가 적은 경우에 이상값을 검정하는 방법

그럽스 T-검정 : 정규 분포를 만족하는 단변량 자료에서 이상값을 검정하는 방법

딕슨의 Q검정 : 오름차순으로 정렬된 데이터에서 범위에 대한 관측치 간의 차이와 비율을 활용. 데이터 수가 30개 미만인 경우에 적절하다.

이상값 처리 방법으로는 삭제, 대체, 변환, 박스플롯 해석을 통한 제거 등이 있다.

2. 분석 변수 처리

래퍼 기법 : 가장 좋은 성능을 보이는 하위 집합을 선택하는 기법. 그리디 알고리즘에 속하며, 과적합의 위험이 발생할 수 있다. 전진 선택법, 후진 제거법, 단계적 방법 등이 있다.

임베디드 기법: 모델의 정확도에 기여하는 변수를 학습하며, 패널티를 통해 좀 더 적은 계수를 가지는 회귀식을 찾는 방향으로 학습한다. 라쏘(L1), 릿지(L2), 엘라스틱넷 등이 있다.

차원축소 : 분석 대상이 되는 여러 변수의 정보를 최대한 유지하면서 데이터 세트 변수의 개수를 줄이는 탐색적 분석기법. 주성분분석(PCA)와 특이값 분해(SVD), 독립성분분석(ICA), 다차원 척도법(MDS) 등이 있다.​

3. 불균형 데이터 처리

언더 샘플링 : 다수 클래스의 데이터를 일부만 선택하여 데이터의 비율을 맞추는 방법. 데이터의 소실이 매우 크고, 중요상 정상 데이터를 잃을 수도 있다. ENN, 토멕링크, CNN, OSS 등이 있다.

오버 샘플링: 소수 클래스의 데이터를 복제 또는 생성하여 데이터의 비율을 맞추는 방법. 과적합을 초래할 수 있다. SMOTE, ADASYN 등이 있다.

그 외 임곗값 이동, 앙상블 기법(위 방법들을 섞는 방법) 등이 있다