Word2vec을 통한 머신러닝 문제의 예측변수 자동생성기법

저자
(Authors)

우덕채, 조윤호

출처
(Source)

한국지능정보시스템학회 학술대회논문집 , 2017.11, 37-38 (2 pages)

발행처
(Publisher)

한국지능정보시스템학회
Korea Intelligent Information Systems Society

URL

http://www.dbpia.co.kr/Article/NODE07284534

APA Style

우덕채, 조윤호 (2017). Word2vec을 통한 머신러닝 문제의 예측변수 자동생성기법. 한국지능정보시스템학회 학술대회논문집, 37-38.

Word2vec 을 통한 머신러닝 문제의 예측변수 자동생성기법

우덕채
국민대학교 데이터사이언스학과
woodc@kookmin.ac.kr

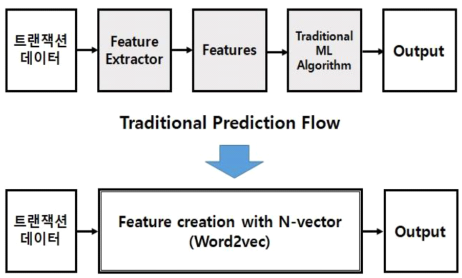
조윤호
국민대학교 경영학부 (교신저자)
www4u@kookmin.ac.kr

Abstract *– 트랜잭션 데이터를 이용하여 예측모델을 만들고자 할 때 원시 데이터로부터 모델링에 사용할 예측변수를 생성하고 선택하는 작업인* ***Feature Engineering*** *이 전체 과정의 80%이상을 차지하고 있다고 알려져 있다. 하지만 이러한 인위적인 노력에도 불구하고* ***Feature Engineering*** *의 방법과 결과에 따라서 예측성능에 현저한 차이를 보이고 있다. 본 연구에서는 Word2Vec 을 이용하여 트랜잭션 데이터로부터 의미 있는 예측변수를 자동으로 생성함으로써* ***Feature Engineering*** *이 필요 없는 새로운 분류예측 머신러닝 프로세스를 제안한다.*

Key Terms *– 분류예측, 머신러닝, Word2Vec, 트랜잭션 데이터,* ***Feature Engineering***

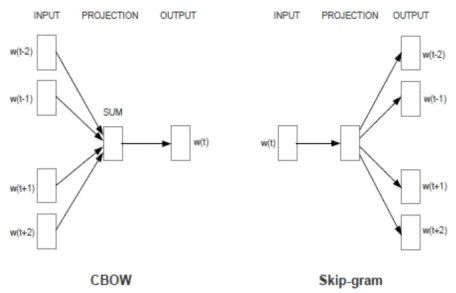
I. 서론

트랜잭션 데이터로부터 분류예측을 위한 의미 있는 예측변수(feature)를 추출하는 과정은 시행착오를 거듭할 수 밖에 없으므로 많은 노력과 시간이 필요하게 된다. 이러한 과정을 자동화하여 시간과 노력을 단축하자는 것이 본 연구의 주된 아이디어이다. <그림 1>에서 같이 기존의 머신러닝 프로세스에서는 Feature 추출작업과 Feature selection 작업이 중요한 절차이며 이후 거기에 맞는 머신러닝 알고리즘을 찾아내고 훈련시켜서 예측하는 프로세스를 거치지만, 본 연구에서는 Feature 를 생성하는 부분을 자동화하고 자동으로 생성된 Feature 를 이용해서 곧 바로 예측 할 수 있는 방법을 제안한다.



<그림 1> **Concept of Proposed Flow**

Feature 를 자동으로 생성하기 위해서 최근 텍스트 분석에서 많이 사용되고 있는 Word2vec 을 활용한다. Word2vec 은 raw text 로부터 워드 임베딩을 학습하는 model 인데, <그림 2>와 같이 CBOW, Skip-gram 두 가지 방식이 있다(Mikolov, et. al., 2013). Word2vec 은 단어간 근접성을 기반으로 수백 차원의 벡터 공간에서 단어를 표현함으로써 벡터 값을 이용하여 단어 간의 의미관계를 파악할 수 있게 한다(Item2vec: Ozsoy, 2016).

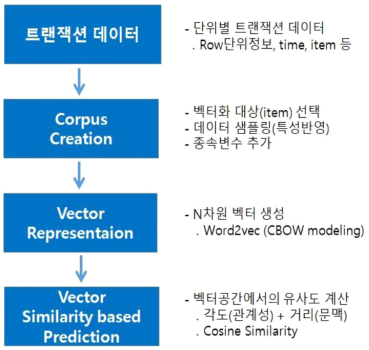


<그림 2> Word2vec Architecture (Mikolov, et. al., 2013)

본 연구에서는 이 중 CBOW 모델을 사용하여 트랜잭션 데이터의 특성을 표현할 수 있는 Feature 를 자동 생성 하고, N 차원 벡터공간에서 유사도를 계산하여 종속변수를 예측하고자 한다.

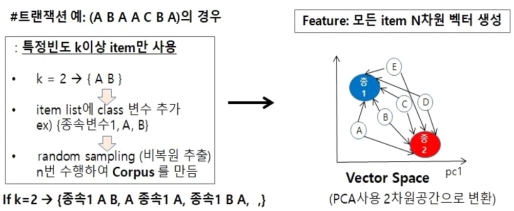
II. 방법론

본 연구에서 제시하는 방법은 <그림 3>과 같은 프로세스를 가진다. 첫 번째 단계에서 트랜잭션 데이터에서 item을 특성 구분단위 별로 집계한다. 그리고 텍스트 분석에서와 같이 희귀성 배제를 위하여 빈도하한을 설정한다. 두 번째 단계에서는 집계된 item으로 Corpus를 생성한다. 구분단위 별 종속변수를 추가하여 random sampling 을 한다. 이 때 비복원 방식으로 추출하고 빈도반영은 하지 않는다. sampling 횟수는 전체 데이터 양에 따라 차이가 있을 수 있지만 20회 정도가 가장 적절하다.



<그림 3> 제안 프로세스

세 번째 단계로 Word2vec을 이용하여 Corpus로부터 모든 item을 N차원 벡터로 매핑한다. CBOW 모델을 사용하고 벡터공간은 300 차원으로 설정한다. <그림 4>는 이 과정을 도식한 것이다.



<그림 4> item 벡터 생성 과정

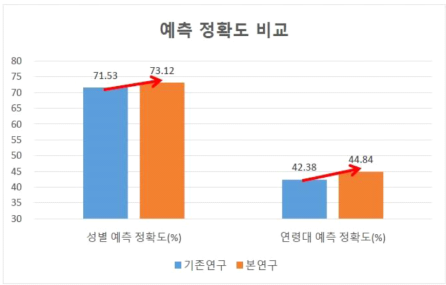
마지막 단계로 N차원 벡터공간에서 items 평균 벡터(feature)와 종속변수와의 Cosine 유사도를 계산하여 종속변수를 예측한다.

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

A : items mean vector B: 종속변수 vector

III. 성능 평가

본 연구의 실험에 사용된 데이터는 국내의 한 인터넷 사이트 순위 분석 전문업체로부터 패널 5,000명에 대해 2012/7/1부터 2013/6/30까지 1년 동안의 온라인 활동기록을 제공받은 패널 데이터 형태의 클릭스트림 데이터이다. 클릭스트림 트랜잭션 데이터로부터 성별과 연령대를 예측하는데, 훈련용 데이터에 50%, 검증용 데이터에 50%를 사용하였으며, 성능평가를 위해 기존논문(Jiae Park, Yoonho Cho, 2016)의 예측결과와 비교하였다. <그림 5>와 같이 binary class(성별)를 예측하는 경우 뿐만 아니라 multi-class(연령대)를 예측하는 경우에서도 기존 연구보다 향상된 결과를 보였다. 따라서 본 연구에서 제안하는 방법론이 Feature Engineering에 소요되는 시간과 노력을 대폭 축약할 수 있으며, 예측정확도에 대한 준거 모델로 활용할 수 있다고 판단한다.



<그림 5 > 성능평가 비교

IV. 참고문헌

Tomas Mikolov, and Ilya Sutskever, and Kai Chen, and Greg Corrado, and Jeffrey Dean, "Distributed Representations of Words and Phrases and their Compositionality", 2013

Makbule Gulcin Ozsoy, "From Word Embeddings to item Recommendation", Middle East Technical University Ankara, Turkey, 2016

Jiae Park,Yoonho Cho, "Clickstream Big Data Mining for Demographics based Digital Marketing" J Intell Inform System 2016 Sep.:22(3)