딥러닝과 통계 모델을 이용한 T-커머스 매출 예측

T-Commerce Sale Prediction Using Deep Learning and Statistical Model

저자
(Authors)

김인중, 나기현, 양소희, 장재민, 김윤종, 신원영, 김덕중
Injung Kim, Kihyun Na, Sohee Yang, Jaemin Jang, Yunjong Kim, Wonyoung Shin, Deokjung Kim

출처
(Source)

정보과학회논문지 44(8), 2017.8, 803-812(10 pages)
Journal of KIISE 44(8), 2017.8, 803-812(10 pages)

발행처
(Publisher)

한국정보과학회
KOREA INFORMATION SCIENCE SOCIETY

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE07226478

APA Style

김인중, 나기현, 양소희, 장재민, 김윤종, 신원영, 김덕중 (2017). 딥러닝과 통계 모델을 이용한 T-커머스 매출 예측. 정보과학회논문지, 44(8), 803-812

ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 44, No. 8, pp. 803-812, 2017. 8
https://doi.org/10.5626/JOK.2017.44.8.803

딥러닝과 통계 모델을 이용한 T-커머스 매출 예측

(T-Commerce Sale Prediction Using Deep Learning and Statistical Model)

김 인 중 † (Injung Kim)　나 기 현 †† (Kihyun Na)　양 소 희 †† (Sohee Yang)　장 재 민 †† (Jaemin Jang)
김 윤 종 †† (Yunjong Kim)　신 원 영 †† (Wonyoung Shin)　김 덕 중 ††† (Deokjung Kim)

요 약 T-커머스는 양방향 디지털 TV를 기반으로 양방향 데이터방송 기술을 활용하여 상거래를 하는 기술융합형 서비스이다. 채널 번호와 판매상품이 제한된 환경에서 T-커머스의 매출을 극대화 하기 위해서는 각 제품의 시간대별 경쟁력을 고려하여 매출이 최대화 되도록 프로그램을 편성해야 한다. 이를 위해, 본 논문에서는 딥러닝을 이용해 T-커머스에서 각 상품을 각 시간대에 편성하였을 때의 매출을 예측하는 방법을 제안한다. 제안하는 방법은 심층신경망을 이용해 판매 상품과 시간대, 주차, 휴일 여부, 그리고 날씨를 입력 받아 실제 방송으로 편성했을 때 기대되는 매출을 예측한다. 그리고, 통계적 모델과 SVD (Singular Value Decomposition)를 적용하여 판매 데이터의 편중 및 희박성 문제를 완화한다. 실제 T-커머스 운영자인 (주)더블유쇼핑의 판매 기록 데이터에 대하여 실험하였을 때 실제 매출과 예측치의 차이가 0.12의 NMAE(Normalized Mean Absolute Error)를 보여 제안하는 알고리즘이 효과적으로 동작함을 확인하였다. 제안된 시스템은 (주)더블유쇼핑의 T-커머스 시스템 적용되어 방송 편성에 활용되었다.

키워드: T-커머스, 매출 예측, 딥러닝, SVD, 지능형 방송 편성

Abstract T-commerce is technology-fusion service on which the user can purchase using data broadcasting technology based on bi-directional digital TVs. To achieve the best revenue under a limited environment in regard to the channel number and the variety of sales goods, organizing broadcast programs to maximize the expected sales considering the selling power of each product at each time slot. For this, this paper proposes a method to predict the sales of goods when it is assigned to each time slot. The proposed method predicts the sales of product at a time slot given the week-in-year and weather of the target day. Additionally, it combines a statistical predict model applying SVD (Singular Value Decomposition) to mitigate the sparsity problem caused by the bias in sales record. In experiments on the sales data of W-shopping, a T-commerce company, the proposed

† 종신회원 : 한동대학교 전산전자공학부 교수(Handong Global Univ.)
ijkim@handong.edu
(Corresponding author임)
†† 비 회 원 : 한동대학교 전산전자공학부
kevinna95@gmail.com
coffee.into.code@gmail.com
jkugan@naver.com
kyjk3@naver.com
rebeccawyshin@gmail.com
††† 비 회 원 : (주)더블유쇼핑 연구소 연구소장
deokjung@mediawill.com

논문접수 : 2017년 1월 25일
(Received 25 January 2017)
논문수정 : 2017년 5월 22일
(Revised 22 May 2017)
심사완료 : 2017년 6월 12일
(Accepted 12 June 2017)

정보과학회논문지 제44권 제8호(2017. 8)

method showed NMAE (Normalized Mean Absolute Error) of 0.12 between the prediction and the actual sales, which confirms the effectiveness of the proposed method. The proposed method is practically applied to the T-commerce system of W-shopping and used for broadcasting organization.

Keywords: T-commerce, sales prediction, deep learning, SVD, broadcasting organization

1. 서 론

T-커머스는 양방향 디지털 TV를 기반으로 양방향 데이터방송 기술을 활용하여 상거래를 하는 기술융합형 서비스이다. 사용자는 TV 방송을 보고 상품을 결정한 후 TV 리모컨, 전화, 모바일 앱을 통해 구매한다. 인터넷에 연결된 디지털 TV를 이용하기 때문에 사용자와 판매자 간 양방향 통신이 가능하고, TV App을 통해 다양한 디지털 서비스와 결합될 수 있다는 점에서 기존 아날로그 TV홈쇼핑과 차별화된다. 또한, 방송을 통해 상품을 판매하고 디지털 TV를 통해 사용자와 소통한다는 점에서 E-커머스와 구분된다. T-커머스를 운영하기 위해서는 미래창조과학부의 사업자 승인이 요구되며, 현재까지 10개의 사업자가 T-커머스 운영을 승인받았다.

T-커머스는 수익성은 상품 자체의 경쟁력과 TV 방송 채널의 시청률에 큰 영향을 받는다. 인기 채널의 경우 경쟁이 심하고 수수료가 높아 T-커머스 사업자들이 활용하기에 부담이 크다. 또한, 중소규모의 T-커머스 사업자들의 경우 치열한 T-커머스/홈쇼핑 시장에서 경쟁력이 높은 상품을 확보하는 것 역시 한계가 있다. 따라서, 현실적으로 활용 가능한 채널을 이용해야 한다는 제약 하에서 수익을 극대화 하기 위한 기술 개발이 필요하다.

T-커머스의 매출은 방송 채널 및 상품 자체의 경쟁력 외에도 하루 24시간의 방송 시간 중 어떤 시간에 어떤 상품을 편성하였는가에 큰 영향을 받는다. 예를 들어, 음식은 식사시간 근처에 편성했을 때, 주부용품은 가족들이 출근, 등교한 시간에 배정했을 때, 그리고, 가족 용품의 경우 가족들이 집에 있는 주말에 배정했을 때 높은 매출을 얻을 수 있다. 그런데, T-커머스 방송 편성은 현재 담당자의 경험과 판단에 의존하고 있으며, 데이터에 기반한 체계적인 분석은 미흡한 실정이다. 그러나, T-커머스의 경우 ICT 기술을 이용한 데이터 수집이 가능하기 때문에 데이터를 이용해 각 상품의 시간대별 매출을 분석/예측하고, 그 결과를 편성에 반영한다면 방송 채널과 판매상품의 제약 하에서 T-커머스의 매출을 극대화하는 데 큰 도움이 된다.

본 논문에서는 T-커머스 방송 편성을 위해 각 상품의 시간 별 매출을 예측하는 방법을 제안한다. 제안하는 방법은 딥러닝 모델인 심층신경망을 이용해 각 상품정보, 판매 시간대, 주차, 휴일 여부 및 날씨 정보를 입력으로 받아, T-커머스에 편성했을 때의 매출을 학습하고, 예측한다. 또한, 각 상품들이 실제 편성/판매된 시간대가 편중되기 때문에 발생하는 데이터의 부재 및 희박성(sparsity) 문제를 개선하기 위해 통계적 모델과 특이값 분해(SVD, singular value decomposition)을 적용하였다. 심층신경망은 실제 T-커머스를 운영하는 (주)더블유쇼핑의 판매 기록 데이터를 이용해 학습하였으며, 그 예측 결과 역시 (주)더블유쇼핑의 방송 편성을 위해 적용되었다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구들을 소개한다. 3장에서는 T-커머스 방송편성 시스템의 구조를 설명하고 4장과 5장에서는 각각 심층신경망과 통계적 모델을 이용한 각 상품의 시간대별 예측 방법을 설명한다. 6장에서는 실험결과를 제시하고, 7장에서 결론을 맺는다.

2. 관련 연구

여러 기업에서는 상황이나 소비 트랜드의 대체 기술의 부상 등 경영환경의 변화에 적응하기 위해 수요예측 기법을 적용한다. 참고문헌 [1]에서는 다양한 수요 예측 기법을 정성적 기법, 정량적 기법, 시스템적 기법으로 구분하였다. 정성적 기법은 해당 제품, 또는 유사 제품 시장에 대한 경험과 지식을 보유한 전문가의 의견 활용, 제품의 기능이나 속성별 수요를 파악해 신제품/기능의 시장 반응을 예측하는 컨조인트 분석, 희소제품의 선택 가능성 예측에 적합한 인덱스 분석 등이 있다. 정량적 분석으로는 독립변수(입력변수)와 종속변수(출력변수) 간의 상관관계를 파악하는 회기분석, 다양한 변수, 시차 간의 인과관계를 모델링하는 시계열 분석, 신제품이나 신기술에 대한 수요를 예측하는 확산 모형 등이 있다. 시스템적 기법으로는 선물시장과 같은 배팅 게임 시스템을 구축하여 참여자들의 행동을 토대로 정보를 수집하고 전망을 예측하는 정보 예측 시장 기법, 변수들 간의 연쇄적 인과관계를 모형화하고 시뮬레이션을 통해 변화 과정을 분석하는 시스템 다이나믹스 기법, 그리고 인공신경망 등이 있다.

본 연구에서는 과거의 판매 데이터로부터 자동으로 각 제품의 시간별 매출을 예측하기 때문에 정량적 기법을 사용하여 상품, 판매시간대 등의 입력 변수로부터 매출을 추정하였다. 날짜(계절), 날씨, 휴일 여부 등 외부 요건을 함께 반영한 정교한 상관관계 추정을 위해 최근 다양한 분야에서 우수한 성능을 보이고 있는 딥러닝을 사용하였다. 딥러닝은 그림 1과 같이 다수의 계층으로 구성된 심

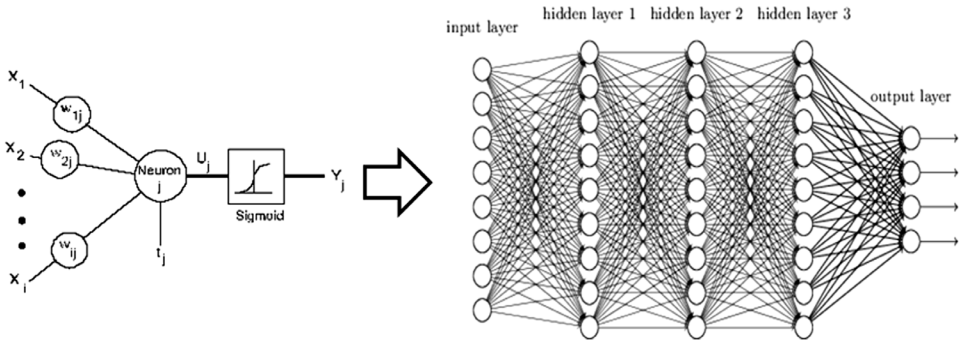


그림 1 심층신경망
Fig. 1 Deep neural networks

층신경망(DNN, deep neural networks)을 기반으로 데이터로부터 높은 수준으로 추상화된 정보를 학습한다[2].

인공신경망은 다수의 뉴런 계층으로 구성되어 입력 정보와 출력 정보 간의 상관관계를 학습/근사한다. 인공신경망의 각 뉴런은 뇌세포의 동작을 모방하여 입력 정보를 가중합의 형태로 조합하여 좀 더 높은 수준의 정보로 변환한 후 다음 뉴런으로 전달한다. 이를 수식으로 나타내면 다음과 같다. o_j 는 출력 뉴런의 값, x_j 는 입력 뉴런의 값, w_{ij} 은 o_j 와 x_j 간의 연결 가중치, 그리고, θ_i 는 편향(bias)을 나타낸다.

$$o_j = f(\sum_{i=1}^{n} w_{ij} x_j$$

인공신경망에서 지식은 연결 가중치 w_{ij} 의 집합에 의해 표현되며 연결 가중치는 학습 알고리즘에 의해 데이터로부터 학습된다. 인공신경망의 학습은 기울기 강하(gradient descent) 알고리즘에 의해 오차를 최소화하는 방향으로 진행된다. 먼저, 기대 출력(desired output)과 인공신경망의 실제 출력(output) 간의 차를 오차함수 E로 정의한다. 그리고, 랜덤하게 초기화된 가중치 w_{ij} 를 기울기(gradient)의 반대 방향으로 이동하면서 오차를 감소시킨다.

$$w_{ij}^{t+1} = w_{ij}^{t} - \eta \frac{\partial E}{\partial w_{ij}^{t}}$$

인공신경망의 각 계층은 이와 같은 뉴런들을 다수 포함하고 있으며, 각 계층은 각각 하위 계층으로부터 전달받은 정보를 좀 더 높은 수준의 정보로 추상화한다. 딥러닝에 사용하는 심층신경망은 많은 수의 계층으로 구성되어 일반적인 신경망보다 더 높은 수준의 추상화를 수행할 수 있다[3]. 특히, 심층신경망의 학습을 위해서는 비지도 학습 알고리즘을 이용한 사전학습(pre-training)이 많이 사용되는데, 이를 위해서는 RBM(restricted Boltzmann machine), SAE(stacked auto-encoder) 등이 사용된다[4-6].

3. 매출 예측 시스템의 구성

T-커머스의 방송편성을 위해서는 각 상품의 시간별 매출 예측이 요구되는데, 이는 분기별, 또는 월별 매출 예측에 비해 더 세밀한 예측이 요구된다. 먼저 T-커머스의 매출에 영향을 미치는 요인들에 대하여 실제 T-커머스를 운영하는 (주)더블유쇼핑의 방송 편성 담당자와의 인터뷰를 통해 조사하였다. 그 결과, 표 1과 같은 요인들이 T-커머스의 매출에 많은 영향을 끼치는 것으로 나타났다.

표 1 T-커머스 매출에 영향을 주는 요소
Table 1 Factors that affects T-Commerce sales

Factors affecting T Commerce sales	Utilization
Availability of bulk stock	Filter out unavailable items
Price and price adjustment	Included in input information
Day, date, holiday	Included in input information
Weather (more sales on rainy or hot days)	Included in input information
Sales items of other T Commerce companies at the same time	Not utilized
Broadcast TV programs at the same time	Not utilized
Frequency of sales broadcast	For each good, maximum # of sales broadcast per day is limited
Characteristics of product	Indirectly included in sales records

본 연구에서는 표 1의 요인 중 정량화가 용이한 상품의 가격, 날짜, 요일, 휴일 여부, 날씨를 이용하였다. 재고확보 가능여부 및 편성 빈도는 매출 예측에는 반영하지 않고, 상품 편성 단계에 반영한다. 동일 시간대 타 업체의 판매상품이나 일반 방송 프로그램도 매출에 영향을 끼치지만, 안정적인 데이터의 획득 및 정량화가 용이하지 않아 활용에 어려움이 있다. 따라서, 본 연구에서는 상품의 가격, 날짜, 요일, 휴일 여부, 날씨로부터 각 상품의 시간대별 매출을 예측하였다.

매출 예측에는 다양한 변수로 구성된 비선형 함수를 학습할 수 있으며, 높은 수준의 정보의 추상화에 뛰어나 복잡한 함수를 효과적으로 학습할 수 있는 심층신경망을 이용하였다. 그런데, T-커머스의 각 상품은 모든 시간에 고르게 편성되지 않는다. 따라서, 각 상품의 판매 기록을 시간별로 집계할 경우 데이터의 편중이 심하고, 그 결과, 특정 상품-시간 조합에 판매 기록이 없어 데이터 희박성 문제가 발생한다. 특히, 특정 시간대에 편성된 기록이 없는 상품의 매출은 학습 데이터의 부재로 인해 심층신경망을 학습하는 데 심각한 문제가 된다. 본 연구에서는 이와 같은 문제를 극복하기 위해 통계적 모델을 심층신경망과 결합하여 사용하였으며, 통계적 모델에 가우시안 평활화와 SVD(특이값 분해)를 통해 데이터 희박도 문제를 완화하였다.

본 연구에서 개발한 상품별, 시간별 T-커머스 매출 예측 시스템은 그림 2와 같이 심층신경망 예측기와 통계적 예측기의 조합으로 구성된다. 심층신경망은 판매기록이 있는 상품-시간대 조합에 대해 좋은 성능을 보이며, 날씨, 휴일 등 다양한 입력 변수들을 반영할 수 있다. 반면, 통계적 예측기는 심층신경망보다 간단하여 학습데이터가 부족으로 인한 오버피팅 문제가 적게 발생하며, 평활화, SVD 등의 기법을 통해 데이터에 희박성을 완화할 수 있다.

두 모델이 출력한 예상 매출은 다음 식과 같이 가중평균에 의해 결합된다. 여기에서 $S_{DNN}(x, y)$과 $S_{stat}(x, y)$은 각각 심층신경망과 통계적 예측기가 출력한 상품 x를 시간 y에 편성했을 때 예상되는 매출이며, $\alpha_{(x,y)}$는 심층신경망의 가중치를 나타낸다.

$$S(x, y) = \alpha_{(x,y)} S_{DNN}(x, y) + (1 - \alpha_{(x,y)}) S_{Stat}(x, y)$$

심층신경망은 예측 대상이 되는 상품-시간대 조합에 대한 판매 데이터가 존재할 경우 효과적으로 예측하지만, 해당 조합의 판매 데이터가 존재하지 않거나 너무 오래된 경우는 그렇지 못하다. 따라서, 심층신경망의 가중치는 예측 대상일로부터 해당 상품의 최근 판매일까지의 시간거리가 멀수록 작아지도록 다음과 같이 설정하였다. 여기에서 u는 1보다 작은 상수이며, 본 연구에서는 0.99를 사용하였다.

- 상품 x가 시간 y에 판매된 기록이 있을 경우:
  $\alpha_{(x,y)} = u^{(\langle \text{예측대상일} \rangle - \langle \text{최종판매일} \rangle)}$
- 상품 x가 시간 y에 판매된 기록이 있을 경우:
  $\alpha_{(x,y)} = 0$

4. 심층신경망을 이용한 상품별/시간별 매출 예측

심층신경망을 이용한 매출 예측기는 그림 3과 같이 구성된다. 심층신경망의 입력 변수로는 상품정보 x, 판매시간 y, 외부정보(날씨, 휴일 여부) z가 포함된다. 출력 노드는 1개이며 상품 x를 시간대 y에 편성했을 때의 예상 매출(분당판매량) $S_{DNN}(x, y)$이다.

T-커머스에서 각 상품은 상품코드로 구분된다. 본 연구에서는 각 상품코드를 1-of-C 코딩을 통해 표현하였다. 즉, 각 상품 x에 대하여 x번째 노드는 1, 다른 노드는 0을 입력으로 한다. 판매 시간은 주 단위로 표현하였다. T-커머스의 방송 편성은 보통 주 단위로 이루어지므로 판매 시간대를 24시간 * 7일 = 168개로 구분하였다.

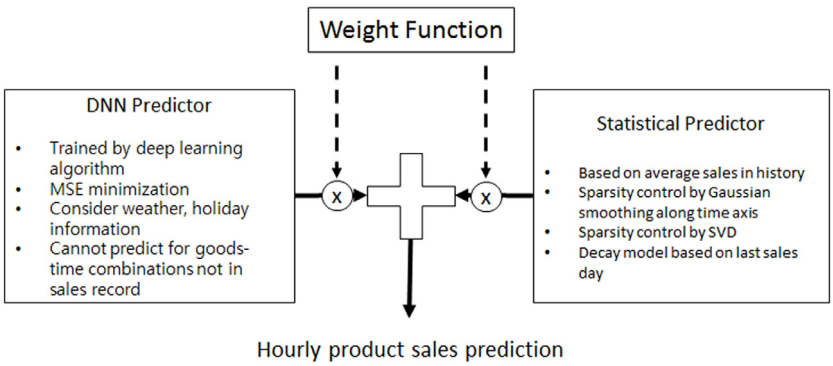


그림 2 T-커머스 매출 예측 시스템 구조

Fig. 2 Structure of T-Commerce sales prediction system

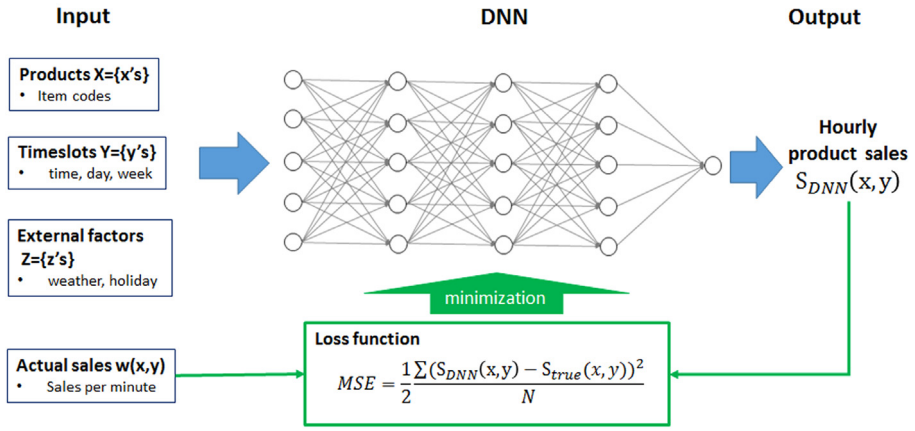


그림 3 심층신경망 매출예측기
Fig. 3 DNN sales predictor

표 1 심층신경망 매출예측기의 입력
Table 1 Input of DNN sales predictor

Categories	Input Information	Coding	Feature dim.
Product	Goods code	1-of-C	396
Timeslot	Day of week (0~6) and hour (0~23)	1-of-C	168
	Week of year (1~52)	1-of-C	52
External factors	Holiday	0(non-holiday) or 1(holiday)	1
	Weather (temperature, rainfall, wind speed, snowfall, cloud, effective temperature)	value (min-max normalized)	6

예를 들어, 일요일 오전 7시는 7번째 노드 (24 * 0(일) + 7 = 7)에 대응하고, 토요일 15시는 159번째 노드 (24 * 6 (토) + 15 = 159)에 대응한다. 그런데, 이와 같이 일주일 내 168개 시간대만으로 구분할 경우 연중 날짜나 계절 정보가 반영되지 않는다. 이를 보완하기 위해 판매일이 1년 중 몇 주차에 해당하는지를 입력 정보에 포함하였다. 1년은 52주로 구성되므로 52개의 입력 노드로 표현한다. 휴일 여부는 0, 또는 1로 표현하였으며, 날씨는 기온, 강수량, 풍속, 강설량, 구름의 양, 체감온도 등 6가지 수치를 min-max 정규화하여 사용하였다. 심층신경망의 입력 정보는 모두 623차원 벡터로 구성되며 요약하면 표 1과 같다.

심층신경망의 학습은 예상 매출과 학습데이터의 실제 매출간 오차를 최소화함으로써 이루어진다. 목적 함수에 해당하는 매출 오차는 다음과 같이 정의한다. $o(x, y)$는 심층신경망이 출력한 예측치를 의미하고, $d(x, y)$는 학습 데이터에 포함된 실제 매출을 의미한다. 모두 5개의 계층을 사용하였으며, 각 계층에는 각각 623, 400, 300, 200, 1개의 은닉 노드가 포함되었다. 신경망 가중치는 -0.1에서 +0.1사이의 난수로 초기화한 후 SGD(stochastic gradient descent) 알고리즘으로 학습하였다. 학습율은 0.01을 적용하였고, 배치의 크기는 64로 설정했다. 사전 실험에서 계층 및 은닉 노드를 추가하거나 RBM (Restricted Boltzmann Machine)을 이용하여 사전 학습을 적용해 보았으나 특별한 성능 개선은 보이지 않았다. 심층신경망과 학습 알고리즘은 모두 직접 구현하였다.

$$MSE = \frac{1}{2}\frac{\sum(o(x,y) - d(x,y))^2}{N}$$

5. 통계기반 상품별/시간별 매출 예측

통계적 예측기는 각 상품의 판매 시간의 편중에 의한 데이터 희박성 문제에 대하여 심층신경망을 보완하는 것을 목적으로 한다. 통계적 예측기의 구조는 그림 4와 같다.

먼저, 과거의 판매기록으로부터 상품들이 168가지 시간대에 편성되었을 때 각 상품-시간대 조합별 분당판매량의 평균을 기반으로 매출을 예측한다. 그런데, 예측대상일로부터의 거리가 가까울수록 예측에 중요하기 때문에 예측대상일과 과거의 판매일 간 시간거리에 반비례하는 가중치를 적용한 가중평균을 사용한다. 가중평균의 계산식은 다음과 같다. $SalesPerMin(d_{sales}, x, y)$는 판매기록 중 날짜 d_{sales}에 상품 x가 시간대 y에 판매된 분

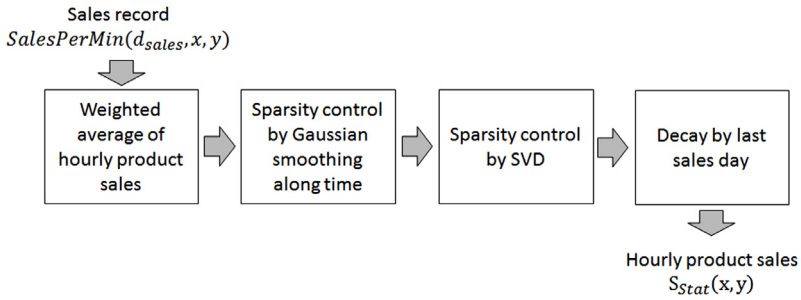


그림 4 통계적 매출 예측기
Fig. 4 Statistical sales predictor

당판매량을 나타낸다. d_{target}는 예측대상일을 의미하며 $w_{(dsales,\ x,\ y)}$는 해당 판매기록의 가중치이다.

$$\mu_{(x,y)} = \frac{\sum_{d_{sales}} SalesPerMin(d_{sales},x,y) * w_{(d_{sales},x,y)}}{\sum_{d_{sales}} w_{(d_{sales},x,y)}}$$

$$w_{(d_{sales},x,y)} = 0.9^{(d_{target}-d_{sales})}$$

이와 같은, 분당판매량의 가중 평균은 심층신경망보다 단순해서 오버피팅이나 데이터 부족에 의한 문제가 다소 적지만, 판매기록이 없는 상품-시간대 조합에 대한 예측이 어렵다는 문제는 여전히 존재한다. 이를 극복하기 위해 본 연구에서는 시간 축에 대한 가우시안 평활화(Gaussian smoothing)와 SVD 등 두 가지 기법을 적용하였다. 동일상품의 유사시간대 예상 매출은 서로 유사하다고 가정할 경우 매출기록이 없는 상품-시간대 조합의 예상 매출을 주변 시간대의 예상 매출로부터 다음과 같이 추정한다. $\hat{\mu}_{(x,y)}$는 가우시안 평활화를 통해 추정한 예상 매출을 의미하며, $G_{(0,\sigma)}(\cdot)$은 평균이 0, 표준편차가 σ인 가우시안 분포를 나타낸다.

$$\hat{\mu}_{(x,y)} = Max_{y'} \frac{G_{(0,\sigma)}(|y'-y|)}{G_{(0,\sigma)}(0)} \mu_{(x,y)}$$

추가적으로 추천시스템에 많이 사용되는 특이값 분해(SVD, singular value decomposition)를 적용하였다[8][9]. SVD는 특징의 차원을 축소하는 알고리즘이다. 데이터의 부족으로 인해 특징이 희박한 경우, 특징 벡터 간 비교나 정확한 파라미터 추정이 어렵다. 이 때 특징의 차원을 축소할 경우 데이터의 밀도가 높아지고, 데이터 부족으로 인한 노이즈가 감소한다[10,11].

SVD를 이용한 희박성 완화 알고리즘은 다음과 같다. 그림 5와 같이 희박한 행렬을 A가 주어졌을 때 SVD를 적용하면 A를 U, D, V^T 등 세 개의 행렬로 분해할 수 있다. 여기에서 D는 특이값(singular values)으로 구성된 대각 행렬이고, U와 V^T는 각각 좌측, 우측 특이벡터(singular vectors)로 구성되는 행렬이다. 여기에서 각

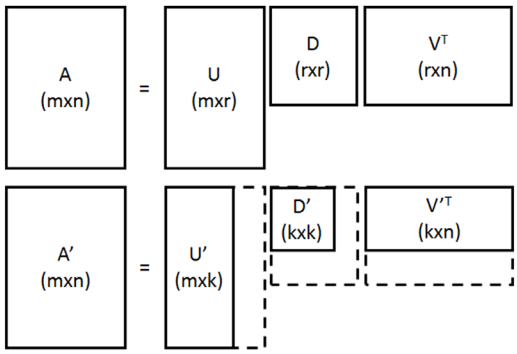


그림 5 SVD에 의한 희박성 완화
Fig. 5 Sparsity control by SVD

행렬의 특이값 및 특이벡터를 특이값이 큰 순서로 k개만 사용하여 다시 원래 행렬을 구성할 경우 데이터의 분별에 중요도가 낮은 차원은 무시된다. 그 결과, 노이즈가 제거되고 비슷한 행과 열은 좀 더 유사해지는 특성을 갖게 되는데, 이 과정에 행렬의 희박성 문제가 완화된다. SVD를 이용해 희박성을 완화한 예는 그림 6과 같다.

그런데, 특정 상품이 오랫동안 편성되지 않았다면 편성담당자가 해당 시간에는 그 제품의 경쟁력이 낮다고 판단한 것으로 추정할 수 있다. 따라서, 최종판매일이 오래된 경우 예상 매출을 감소시키는 감소 모델을 적용하였다. 최종적으로 통계적 모델의 예상 매출은 다음 식에 의해 계산된다.

$$S_{Stat}(x,y) = \hat{\mu}_{(x,y)} 0.9^{(<예측대상일>-<최종판매일>)/7}$$

6. 실 험

6.1 실험환경 및 데이터

실험은 인텔 제온 E5-2630 v4 2.2GHz 2개와 32GB RAM이 장착된 Linux 서버에서 수행하였다. 실험 데이터로는 (주)더블유쇼핑의 2015년 7월부터 2016년 9월까지의 15개월간의 판매 데이터를 사용했다. 판매기록에는

상품그룹코드	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00
70000022	-	-	-	-	-	-	-	-	-	-
70000023	2,452	4,870	4,107	3,871	4,065	5,714	7,069	14,156	9,315	-
70000024	51,534	48,600	32,911	20,173	29,573	34,177	68,770	121,287	326,467	270,708
70000025	45,276	72,009	48,631	36,701	30,859	30,821	46,214	202,516	188,729	200,307
70000027	4,102	8,147	4,234	4,178	6,800	5,290	6,689	17,453	12,263	8,979
70000029	-	-	-	-	-	9,388	20,243	70,656	81,035	69,885
70000030	-	-	578	681	893	1,567	2,421	6,055	4,976	3,075
70000031	29,737	41,506	27,275	25,713	53,197	54,941	144,274	242,500	219,470	122,820
70000034	31,996	56,869	42,919	36,197	34,012	35,014	47,589	146,911	195,252	202,820

(a) Hourly product sales prediction before applying SVD

상품그룹코드	0:00	1:00	2:00	3:00	4:00	5:00	6:00	7:00	8:00	9:00
70000022	32,620	43,653	27,582	22,449	24,419	31,528	47,369	114,897	113,989	91,614
70000023	2,452	4,870	4,107	3,871	4,065	5,714	7,069	14,156	9,315	71,017
70000024	51,534	48,600	32,911	20,173	29,573	34,177	68,770	121,287	326,467	270,708
70000025	45,276	72,009	48,631	36,701	30,859	30,821	46,214	202,516	188,729	200,307
70000027	4,102	8,147	4,234	4,178	6,800	5,290	6,689	17,453	12,263	8,979
70000029	27,815	36,363	22,679	19,021	21,157	9,388	20,243	70,656	81,035	69,885
70000030	14,110	15,570	578	681	893	1,567	2,421	6,055	4,976	3,075
70000031	29,737	41,506	27,275	25,713	53,197	54,941	144,274	242,500	219,470	122,820
70000034	31,996	56,869	42,919	36,197	34,012	35,014	47,589	146,911	195,252	202,820

(b) Hourly product sales prediction after applying SVD

그림 6 SVD에 의한 희박성 완화 예

Fig. 6 Example of sparsity control by SVD

396가지 상품에 대하여 23만 여 건의 판매에 대한 상품명, 상품코드, 가격, 판매 날짜 및 시간, 판매된 방송 사업자 등이 포함되어 있다. 각 상품의 판매 기록을 판매 시간 별로 분류, 누적하여 시간 별 매출을 계산해 실험에 사용하였다.

그 중 2016년 8월까지의 14개월간의 데이터를 학습에 사용하였고, 2016년 9월 1개월의 데이터를 테스트에 사용했다. 매출예측기의 출력과 실제 매출 데이터를 비교하기 위해서는 NMAE(normalized mean absolute error)를 사용하였다. MNAE의 계산식은 다음과 같다.

$$\text{NMAE} = \frac{1}{(Max - Min)}\frac{1}{N}\sum_{i}|S_i^{predict} - S_i^{true}|$$

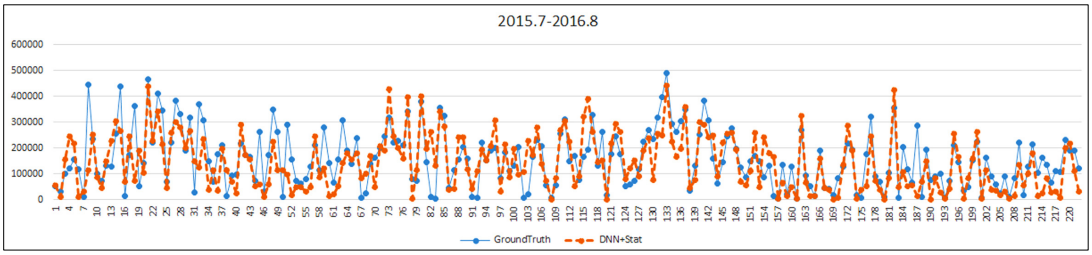
날씨가 매출에 미치는 영향을 파악하기 위해 위의 기간 중 비가 온 날과 그렇지 않은 날의 매출을 집계해 비교하였다. 그 결과 비가 오지 않은 날의 평균 분당 매출은 47,746원이었으나, 비가 온 날의 평균 분당 매출은 53,770원으로 나타났다. 평균적으로 비가 온 날의 매출이 비가 오지 않은 날보다 12.6% 상승하는 것으로 나타났다.

6.2 매출 예측 결과

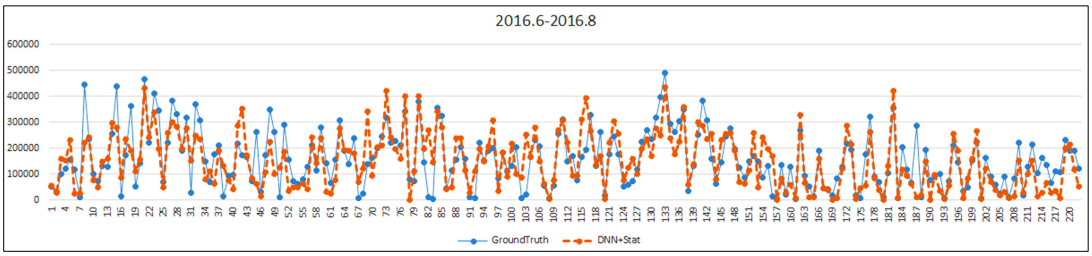
먼저 심층신경망을 이용한 매출 예측기의 성능을 평가하였다. 학습데이터의 구성과 관련해 14개월간의 데이터를 모두 학습에 사용한 경우와, 최근 3개월에 해당하는 2016년 6-8월의 데이터만을 이용해 학습한 경우로 구분하였다. 전자의 경우 학습에 많은 데이터가 사용된다는 장점이 있는 반면, 후자의 경우 최근의 매출을 집중적으로 반영함으로써 최근 트랜드와 계절 등이 강조된다는 장점이 있다. 두 경우 2016년 9월 매출을 예측한 후 실제 매출과 비교하였다.

제안하는 방법으로 매출을 예측한 결과는 그림 7 및 표 2와 같다. 그림 7에서 가로 축은 테스트 기간 중의 판매 시간을 나타내고, 세로 축은 각 시간대에 판매된 상품의 분당 매출을 나타낸다. 그림 7의 그래프는 제안하는 매출 예측 알고리즘이 실제 매출의 트랜드를 효과적으로 예측하고 있음을 보여준다. 그런데, 14개월간의 전체 데이터로 학습한 경우보다 최근 3개월만의 데이터로 학습한 경우가 근소하게 더 낮은 NMAE를 보였다. 이는 오래된 과거 데이터에는 예측대상일과 다른 환경에서의 데이터가 많이 포함되기 때문에 오히려 좋지 않은 영향을 주기 때문인 것으로 추정된다. 반면, 최근 3개월의 데이터만으로 학습할 경우 최근 트랜드 및 계절에 집중하기 때문에 더 좋은 결과를 낸 것으로 추정된다.

동일한 학습 데이터를 사용한 경우 심층신경망 예측기가 통계적 예측기보다 훨씬 좋은 성능을 보였다. 그러나, 심층신경망만 사용한 경우 구매 기록이 없는 상품-시간대 조합에 대하여 잘못된 결과를 출력하는 현상이 발견되었는데, 이는 T-커머스 방송 편성에 잘못된 영향을 끼칠 수 있다. 통계적 예측기의 경우 그러한 조합의



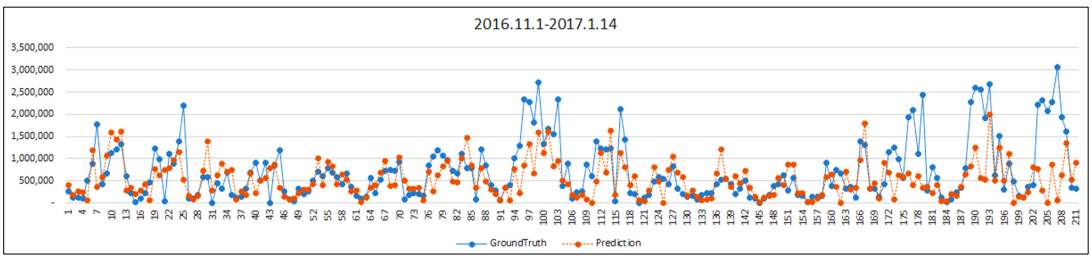
(a) Sales prediction trained on sales date for 14 mon



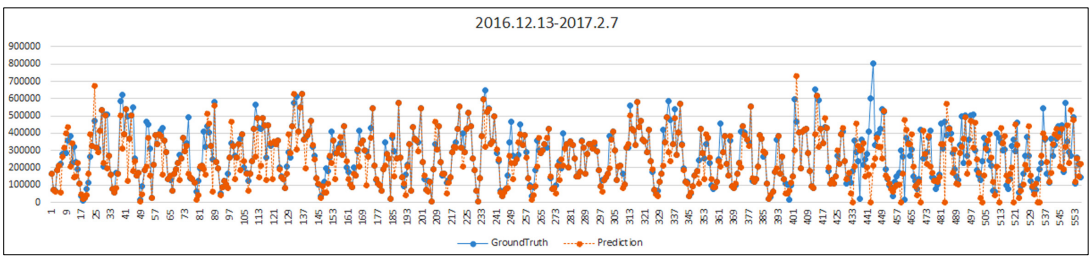
(b) Sales prediction trained on sales date for 3 mon

그림 7 매출 예측과 실제 매출 비교

Fig. 7 Comparison of sales prediction and actual sales



(a) Prediction of sales between 2017.1.15 and 2017.1.29



(b) Prediction of sales between 2017.2.28 and 2017.3.13

그림 8 다른 구간에서의 매출 예측과 실제 매출 비교

Fig. 8 Comparison of sales prediction and actual sales in other period

예상 매출을 낮춰주는 효과가 있어 심층신경망 예측기의 문제점을 보완해 주는 효과가 있었다. 심층신경망과 통계적 예측기를 함께 사용한 경우 심층신경망과 유사한 NMAE를 보이면서도 구매 기록이 없는 상품-시간대 조합에 대하여 낮은 예상 매출을 출력하여 좀 더 안정적인 동작을 보였다.

추가적인 검증을 위해 다른 구간의 데이터에 대하여 실험하였다. 그 결과는 그림 8과 같다. 학습을 위해서는

표 2 시간별 매출 예측기의 성능(NMAE)
Table 2 Accuracy of Hourly sales predictor in NMAE

Predictors	Training data	
	14 months	3 months
DNN predictor	0.13	0.12
Statistical predictor	0.19	0.18
DNN+Stat	0.13	0.12

각각 테스트 구간 이전의 3개월 매출 기록을 사용하였다. 그림 9의 결과에서도 제안하는 방법은 다른 구간에 대해서도 매출 트랜드를 효과적으로 예측하였다. 그림 10(a)의 중간과 우측에는 예측 값이 실제 매출보다 상당히 낮았다. 이는 해당 구간의 마지막 3일이 설 명절이었기 때문에 명절 기간 및 직전 주말에 매출이 크게 증가하였는데, 예측기에는 이러한 점이 잘 반영하지 않았기 때문으로 추정된다. 그러나, 그림 11(b)와 같이 명절이 없는 구간에는 비교적 안정적인 매출 성능을 보였다. 그림 12(a)와 (b)의 NMSE는 각각 0.63, 0.18이었다.

7. 결론 및 향후 개선 방향

본 논문은 T-커머스의 방송편성을 위한 상품별-시간대별 매출 예측 알고리즘을 제안하였다. 제안하는 방법은 상품, 판매시간, 휴일 여부, 날씨를 입력 받아 예상 매출을 출력하는 심층신경망 기반 예측기와 통계적 예측기의 조합으로 구성된다. 통계적 예측기에는 가우시안 평활화 및 SVD를 적용하여 데이터 희박성 문제를 완화하였다. 실험 결과 제안하는 방법은 T-커머스에서 각 상품의 시간별 매출 트랜드를 효과적으로 예측하는 것으로 나타났다. 심층신경망 예측기는 통계적 예측기보다 더 낮은 오차를 보였으며, 통계적 예측기는 판매기록이 없는 상품-시간대 조합에 대하여 낮은 예상 매출을 출력함으로써 심층신경망의 단점을 보완하는 효과를 보였다. 두 예측기를 제안하는 방법으로 결합한 결과 심층신경망과 유사한 수준의 NMAE를 보이면서도 판매기록이 없는 상품-시간대 조합에 대하여 안정적인 결과를 보였다.

제안하는 시스템은 두 가지 면에서 개선의 여지가 있다. 첫째, 동일 시간대 타 업체의 판매상품이나 일반 방송 프로그램이 T-커머스 매출에 영향을 끼침에도 안정적인 데이터의 획득 및 정량화의 어려움으로 인해 예측기의 입력 정보로 사용되지 못하고 있다. 또한, 명절 기간 등 특수한 기간에 대해서는 정확도가 저하되었는데, 명절의 판매 데이터는 양이 학습에 충분하지 않을 뿐 아니라, 명절 날짜가 음력을 사용하기 때문에 데이터 기반 방법으로 학습하는 것이 쉽지 않다. 이러한 부분이 개선된다면 더욱 정확한 예측이 가능할 것으로 기대된다.

References

[1] S. Park, et. al., 2012, "SERI Issue Paper: Effective Sales Predict and Examples," Samsung Economic Research Institute, [Online] Available: https://www.slideshare.net/girujang/seri20120303.

[2] I.-J. Kim, "Theory and Practice of Deep Learning Implementation," *KIISE Summer workshop on Pattern Recognition and Machine Learning*, 2016.

[3] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning*, Vol. 2, iss. 1, pp. 1-127, 2009.

[4] P. Smolensky, "Chapter 6: Information Processing in Dynamical Systems: Foundations of Harmony Theory," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 1: Foundations*, MIT Press, pp. 194-281, 1986.

[5] G. E. Hinton, R. R. Salakhutdinov, "Reducing the Dimensionality of Data with Neural Networks," *Science*, 313 (5786), pp. 504-507, 2006.

[6] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets, Neural Computation," Vol. 18, No. 7, pp. 1527-1554, 2006.

[7] Y. Bengio, et. al., "Greedy Layer-Wise Training of Deep Networks," *NIPS*, 2006.

[8] Singular Value Decomposition, Wikipedia, [Online]. Available: https://en.wikipedia.org/wiki/Singular_value_decomposition

[9] M. Kim, "Recommendation System: Centered on Collaborative Filtering," 8th ROSAEC workshop, 2012.

[10] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender System - A Case Study," *WebKDD-2000 Workshop*, 2000.

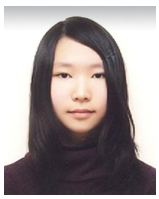
[11] Berry, M. W., Dumais, S. T., and O'Brian, G. W., "Using Linear Algebra for Intelligent Information Retrieval," *SIAM Review*, Vol. 37, No. 4, pp. 573-595, 1995.

김 인 중

1990년 2월~2001년 2월 KAIST 전산학과(학사, 석사, 박사). 2001년 2월~2006년 2월 (주)인지소프트 책임연구원. 2006년 3월~현재 한동대학교 전산전자공학부 교수. 2011년 12월~2013년 1월 Visiting Scholar at U.C.Irvine. 관심분야는 딥러닝, 머신러닝, 인공지능, 영상처리, 자연어처리

나 기 현

2013년 2월~2017년 2월 한동대학교 전산전자공학부 컴퓨터공학전공(학사). 2017년 6월~현재 대한민국 공군 장교 복무중(정보통신 특기). 관심분야는 머신러닝, 인공지능, 뇌공학



양 소 희

2014년 2월~현재 한동대학교 전산전자공학부 컴퓨터공학전공(학사). 관심분야는 딥러닝, 머신러닝, 인공지능

장 재 민

2011년 3월~2017년 2월 한동대학교 전산전자공학부 컴퓨터공학전공(학사). 2017년 3월~한컴GMD 포렌식연구소 연구원 관심분야는 머신러닝, 컴퓨터비전, 디지털포렌식

김 윤 종

2011년 2월~2017년 2월 한동대학교 전산전자공학부 컴퓨터공학전공(학사). 2017년 3월~현재 KAIST 지식서비스공학대학원 석사과정. 관심분야는 데이터마이닝, 머신러닝, 자연어처리

신 원 영

2012년 2월~2017년 2월 한동대학교 전산전자공학부 컴퓨터공학전공(학사). 2017년 3월~현재 KAIST 지식서비스공학대학원 석사과정. 관심분야는 머신러닝, 데이터마이닝

김 덕 중

1987년 2월~1993년 8월 한양대학교 전자계산학과(학사,석사). 1993년~2000년 (주)LG소프트 선임연구원. 2000년~2007년 (주)에어코드 수석연구원. 2007년~2008년 (주)유피온 연구소장. 2008년~현재 (주)인터렉티브비전 대표. 2015년~현재 더블유쇼핑 연구소장. 관심분야는 데이터방송, T-커머스, 빅데이터