

비정형 정보와 CNN 기법을 활용한 이진 분류 모델의 고객 행태 예측: 전자상거래 사례를 중심으로*

김승수

한양대학교 일반대학원 경영학과
(suwu@hanyang.ac.kr)

김종우

한양대학교 경영대학 경영학부
(kjw@hanyang.ac.kr)

최근 딥러닝 기술이 주목을 받고 있다. 대중들의 관심을 받았던 국제 이미지 인식 기술 대회(ILSVR)와 알파고(AlphaGo)에서 사용된 딥러닝 기술이 바로 합성곱 신경망(CNN; Convolution Neural Network)이다. 합성곱 신경망은 입력 이미지를 작은 구역으로 나누어 부분적인 특징을 인식하고 이것을 결합하여 전체를 인식하는 특징을 가진다. 이러한 딥러닝 기술이 우리의 생활에 있어 많은 변화를 야기할 것이라는 기대를 주고 있지만 현재까지는 이미지 인식과 자연어 처리 등에 그 성과가 국한되어 있다. 비즈니스 문제에 대한 딥러닝 활용은 아직까지 초기 연구 단계로 향후 마케팅 응답 예측이나 허위 거래 식별, 부도 예측과 같은 전통적 비즈니스 문제들에 대해 보다 깊게 활용되고 그 성능이 입증된다면 딥러닝 기술의 활용 가치가 보다 더 주목받게 될 것으로 기대된다. 이러한 때 비교적 고객 식별이 용이하고 활용 가치가 높은 빅데이터를 보유하고 있는 전자상거래 기업의 사례를 바탕으로 하여 딥러닝 기술의 비즈니스 문제 해결 가능성을 진단해보는 것은 학술적으로 매우 의미 있는 시도라 할 수 있겠다.

이에 본 연구에서는 전자상거래 기업의 고객 행태 예측력을 높이기 위한 방안으로 합성곱 신경망을 활용한 ‘이종 정보 결합(Heterogeneous Information Integration)의 CNN 모델’을 제시한다. 이는 정형과 비정형 정보를 결합하여 다층 퍼셉트론 구조의 합성곱 신경망에서 학습시키는 모델로서 최적의 성능을 발휘하도록 ‘이종 정보 결합’과 ‘비정형 정보의 벡터 전환’, 그리고 ‘다층 퍼셉트론 설계’로 하는 3개의 내부 아키텍처를 정의하고 각 아키텍처 단위로 구성되는 방식에 따른 성능을 평가하여 그 결과를 바탕으로 제안 모델을 확정하고 그 성능을 평가해보고자 한다. 고객 행태 예측을 위한 목표 변수는 전자상거래 기업에서 중요하게 관리하고 있는 재구매 고객, 이탈 고객, 고빈도 구매 고객, 고빈도 반품 고객, 고단가 구매 고객, 고할인 구매 고객 등 모두 6개의 이진 분류 문제로 정의한다.

제안한 모델의 유용성을 검증하기 위해서 국내 특정 전자상거래 기업의 실제 데이터를 활용하여 실험을 수행하였다. 실험 결과 정형과 비정형 정보를 결합하여 CNN을 활용한 제안 모델이 NBC(Naïve Bayes classification)과 SVM(Support vector machine), 그리고 ANN(Artificial neural network)에 비해서 예측 정확도와 F1 Measure가 높게 평가되었다. 또 NBC, SVM, ANN에서 정형 정보만을 사용할 때 보다 정형과 비정형 정보를 결합하여 입력 변수로 함께 활용한 경우에 예측 정확도가 향상되는 것으로 나타났다.

따라서 실험 결과로부터 비정형 정보의 활용이 고객 행태 예측의 정확도 향상에 기여한다는 점과 CNN 기법의 특징 추출 알고리즘이 VOC에 사용된 단어들의 분포와 위치 정보를 해석하여 문장의 의미를 파악하는데 효과적이라는 점을 실증적으로 확인하였다는데 그 의미가 있다고 할 수 있겠다. 이를 통해서 CNN 기법이 지금까지 소개된 이미지 인식이나 자연어 처리 분야 외에 비즈니스 문제 해결에도 활용 가치가 높다는 점을 확인하였다는데 이 연구의 의의가 있다 하겠다.

주제어 : 고객 행태 예측, 합성곱 신경망, 딥러닝, 고객의 소리

논문접수일 : 2018년 5월 21일 논문수정일 : 2018년 6월 14일 게재확정일 : 2018년 6월 16일

원고유형 : 일반논문 교신저자 : 김종우

* 이 논문은 2017년도 정부(과학기술정보통신부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (No.2016-0-00562, 상대방의 감성을 추론, 판단하여 그에 맞추어 대화하고 대응할 수 있는 감성지능 기술 연구개발)

1. 서론

딥러닝은 2012년 국제 이미지 인식 기술 대회인 ILSVR(Imagenet Large Scale Visual Recognition Challenge)에서 Hinton 교수의 토론토대학 팀이 출전하여 인식률을 85%로 끌어 올리며 2위 팀보다 10% 가까운 차이로 우승을 차지(Krizhevsky et al., 2012)하면서 주목을 받았다. 이후 2016년 3월 서울에서 열린 바둑 대국에서 Google의 DeepMind가 만든 바둑 인공지능 프로그램인 알파고(AlphaGo)가 이세돌 9단을 상대로 4대 1의 대승을 거두면서 일반 대중들에게 크게 관심을 받기 시작하였다. 앞선 사례에서 공통적으로 사용된 딥러닝 기법이 바로 합성곱 신경망(CNN; Convolution Neural Network)이다. 합성곱 신경망은 입력 이미지를 작은 구역으로 나누어 부분적인 특징을 인식하고 이것을 결합하여 전체를 인식하는 특징을 가진다. 국지적인 패턴과 이를 바탕으로 전반적인 형세를 파악 하는 것이 중요한 바둑이 이 기술이 사용된 적절한 예이다(Chu et al., 2016).

이러한 딥러닝 기술이 우리의 생활에 있어 많은 변화를 야기할 것이라는 기대를 주고 있지만 현재까지는 이미지 인식과 자연어 처리 등에 그 성과가 국한되어 있다. 비즈니스 문제에 대한 딥러닝 활용은 신문 기사를 그 주제에 맞는 카테고리 분류(Kim, 2014; Zhang et al., 2015) 하는 시도를 시작으로 마이크로 블로그 정보를 활용하여 고객의 통신 서비스 이탈 여부를 예측하는 초기 연구 단계에 머물러 있다. 향후 마케팅 응답 예측이나 허위 거래 식별, 부도 예측과 같은 전통적 비즈니스 문제에 보다 깊게 활용되어 그 성능이 입증된다면 딥러닝 기술의 활용 가치가 보다 주목받게 될 것으로 기대된다.

여기에 최근 우리의 생활과 산업 전반에 걸쳐 급속도로 진행되고 있는 디지털 트랜스포메이션(Digital transformation)의 영향으로 활용 가치가 높은 새로운 데이터 소스가 계속해서 늘어나고 있으며 딥러닝 기술의 진보로 과거에는 어려웠던 비정형 정보를 포함한 빅데이터에 대한 분석이 가능해졌다는 점에 주목해 볼 필요가 있다. 이러한 상황에서 비교적 고객 식별이 용이하고 활용 가치가 높은 빅데이터를 보유하고 있는 전자상거래 기업의 사례를 기반으로 한 딥러닝 기술의 비즈니스 문제 해결 가능성을 진단해보는 것은 학술적으로 매우 의미 있는 시도라 할 수 있겠다.

특히 전자상거래 기업에서는 최근 경쟁 환경이 급변하고 치열해 짐에 따라 수익 극대화를 위한 고객 행태 분석의 중요성이 점차 커지고 있고, 기업들은 환경변화와 경쟁에 대응하고 수익을 극대화하기 위해서 한 번 획득한 고객을 적극적으로 관리하고 지속적으로 유지하고자 노력하고 있다. 일반적으로 신규고객을 확보하는데 더 많은 비용이 발생하지만 어렵게 확보한 고객에게서 재구매가 이어지지 않는다면 수익 창출을 지속시키기 어렵다는 점에서 자사에 높은 수익을 가져다 주는 충성 고객을 확보해야 하는 필요성이 커지고 있다. 따라서 충성 고객을 선별하고 관리하기 위해서 기업은 보유하고 있는 데이터를 적극 활용하여 이탈 가능성과 고객 생애 가치 등과 같은 고객 행태를 연구하는 노력들을 기울이고 있는 실정이다.

이에 본 연구에서는 전자상거래 기업의 고객 행태 예측력을 높이기 위한 방안을 합성곱 신경망을 기반으로 제시하고자 한다. 고객 프로파일과 거래 데이터와 같은 정형 정보 외에 텍스트 데이터인 비정형 정보를 결합하고 딥러닝 기반

의 합성곱 신경망 기법을 활용하는 모델을 제안하여 예측력을 높이고자 한다. 본 연구의 구성은 다음과 같다. 2장에서는 기존의 고객 행태 예측과 그 기반 기술과 관련한 연구들을 살펴본다. 3장에서는 본 연구에서 제안하는 고객 행태 예측 방안을 위한 모델을 설명하고 실제 국내 전자상거래 기업의 데이터를 활용하여 실험한 결과를 제시하도록 한다. 마지막으로 4장에서는 결론을 제시한다.

2. 관련 연구

2.1 고객 행태 예측

최근 기업들에서는 고객가치 증대 및 고객과의 관계 형성을 위한 고객 행태 예측의 중요성이 점차 커지고 있다. 고객 행태 예측에는 기업의 수익 극대화를 목적으로 한 고객 이탈 방지가 주로 연구되고 있다(Kim et al., 2005). 이는 신규 고객을 유치하는 노력에 비해서 기존 고객을 잘 유지시키는 것이 기업 입장에서 보다 효과적이라는 인식에 기반하고 있다(Lee et al., 2007). 특히 온라인 분야에서는 그 특성상 고객 행태 정보를 활용하여 고객 세분화 및 상품 구성과 같은 기업의 마케팅 전략 수립에 기여(Lohse et al., 2000)하고자 하는 노력들이 있어 왔다. 이러한 이유는 회원 정보나 웹 로그 데이터 활용이 기술적으로 용이한 전자상거래 분야에서는 고객의 행동 패턴을 실시간 분석하여 앞으로의 고객 행태를 예측하는 것은 기업의 현재의 수익 증대에 기여할 뿐만 아니라 미래의 기업 경쟁력 확보를 위한 필수 요소이기 때문이다.

고객 행태 예측과 관련한 주요 선행 연구 사례

들을 살펴보면 다음과 같다. 먼저 로지스틱 회귀 분석으로 홈쇼핑 고객의 이탈 모델을 제안하여 고객의 이탈 확률을 예측하는 연구(Kim et al., 2005)와 은행 고객과 신용카드사 고객 이탈을 로지스틱 회귀분석과 인공신경망을 사용하여 예측하여 성능을 비교하는 연구(Lee, 2001, 2002), 그리고 다수의 분류 모델을 결합적으로 활용하여 신용카드 고객의 이탈을 예측한 연구(Lee et al., 2007)가 있다. 또한 자연어 처리와 관련해서는 신문사 고객 센터로 접수된 이메일 정보를 활용한 고객 이탈 예측 연구(Coussement et al., 2008)와 e-Commerce에서 고객의 VOC 정보를 이용한 고객 이탈 예측 연구(Yu et al., 2012)가 있다. 딥러닝 기법과 관련해서는 마이크로 블로그 정보를 활용하여 고객의 통신사 이탈을 예측하는 연구(Cridach et al., 2017)가 있다.

2.2 합성곱 신경망

합성곱 신경망(CNN; Convolution Neural Network)은 인공 신경망(ANN; Artificial Neural Network)에서 파생되어 나온 기법이다. 먼저 인공 신경망을 살펴보면, 인공 신경망은 디지털 컴퓨터에 인간 두뇌의 신경의 연결을 흉내 내도록 설계된 것으로 입력 계층, 은닉 계층, 출력 계층으로 구성된다. 은닉 계층에는 인공 뉴런이 있으며 입력 계층으로부터 입력 받은 신호들을 하나의 값으로 조합하고 이를 출력으로 변환시키는 것을 활성화 함수(Activation function)를 통해서 동작하게 된다. 활성화 함수에서는 조합된 입력들이 임계점을 넘기 전까지는 매우 낮은 값을 유지하다가 입력들이 합쳐져서 임계점을 넘기게 되면 활성화되고 출력이 높아지게 된다. 활성화 함수는 모든 입력을 하나의 값으로 통합시키는

결합 함수(Combination function)와 결합 함수의 값을 전환시키는 전환 함수(Transfer function)로 구분된다. 전형적인 전환 함수에는 시그모이드(Sigmoid), 선형(Linear), 하이포보릭 탄젠트(Hyperbolic tangent)가 있다. CNN은 인공 신경망에 합성곱 계층(Convolutional layer)과 풀링 계층(Pooling layer)이 추가된다. 합성곱 계층에서는 합성곱 연산을 처리하며 입력 데이터에 필터를 적용한다. 합성곱 연산은 필터의 윈도우(Window)를 일정 간격으로 이동해가며 입력 데이터에 적용한다. 이 때 편향(Bias)은 필터를 적용한 후의 데이터에 더해진다.

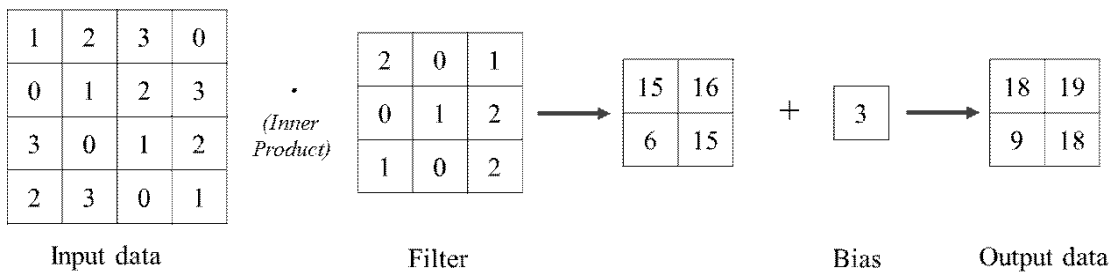
풀링은 가로·세로 방향의 공간을 줄이는 연산이다. 예를 들어 2x2 영역을 원소 하나로 집약하여 공간 크기를 줄이게 된다. 최대 풀링(Max pooling)은 최대값을 구하는 연산이며 2x2는 대

상 영역의 크기를 표현한다.

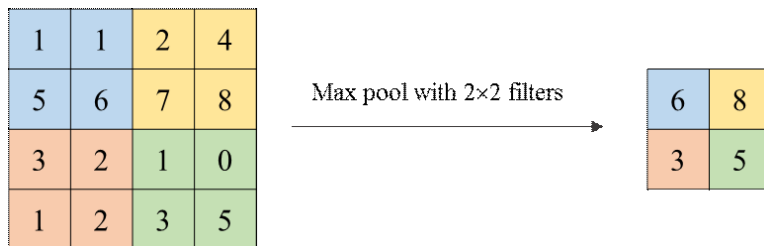
지금까지 기술한 합성곱 신경망(Convolution neural network)은 생물의 시각 처리 과정의 모방(Fukushima, 1980)으로, 이미지 인식 분야에서 성과를 보이며 다양한 분야에서 응용되고 있다.

2.3 워드 임베딩(Word Embedding)

워드 임베딩은 텍스트 정보를 분석이 가능한 숫자 정보로 변환하는 기술이며 주요 기법으로는 word2vec과 doc2vec, char2vec이 있다. 먼저 word2vec은 단어(Word)를 기준으로 하여 주변 단어들을 가지고 그 중심 단어를 예측하는 CBOW(Continuous Bag of Words)와 이와 반대로 특정 중심 단어를 가지고 그 주변 단어를 예측하는 Skip-Gram의 2가지 알고리즘이 있다. Skip-Gram 알고리즘에서는 중심 단어와 주변 단



〈Figure 1〉 Convolution Operation



〈Figure 2〉 2x2 Max Pooling Operation

어가 정해지면 아래 수식을 최대화하는 방향으로 학습을 진행하게 된다. 이를 수식으로 표현하면 <Formula 1>과 같이 중심 단어(w)가 주어졌을 때 주변 단어(w_0)가 나타날 조건부 확률이 된다. 여기서 v_w 와 v'_w 은 단어 w 의 입력, 출력 벡터이고 W 는 단어의 총 개수이다.

$$p(w_0|w_I) = \frac{\exp(v'_{w_0}{}^T v_{w_I})}{\sum_{w=1}^W \exp(v'_w{}^T v_{w_I})} \quad (1)$$

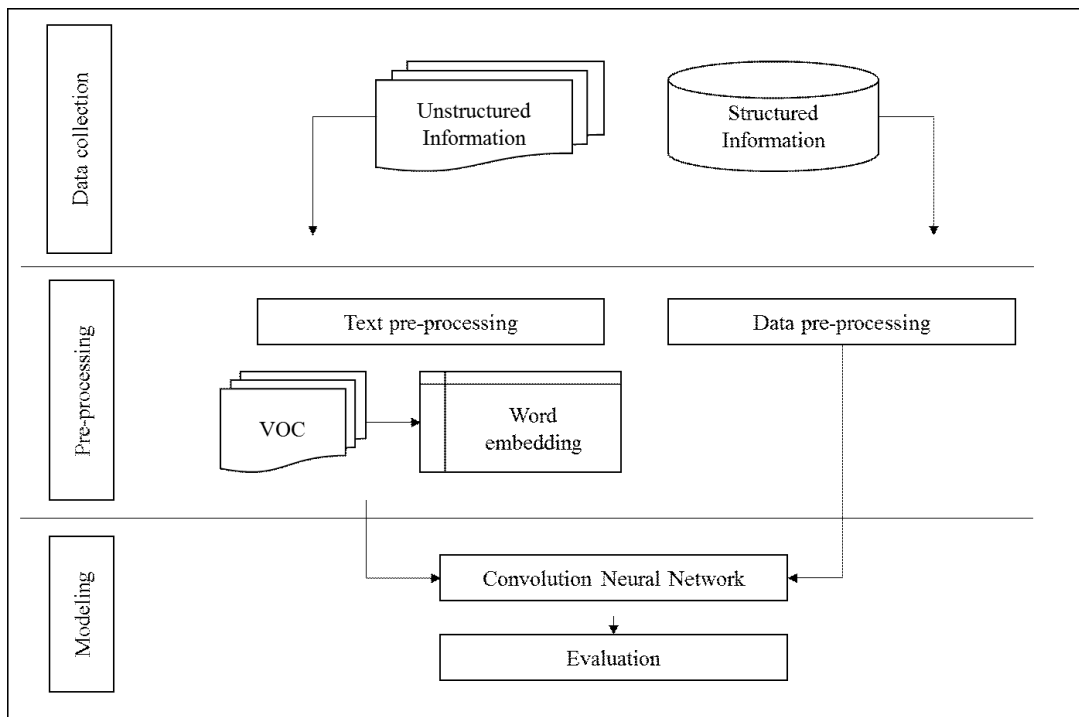
다음으로 doc2vec(또는 Paragraph2vec)은 문장이나 전체 문서의 단위를 비지도 학습하는 기법으로 Distributed memory(dm)와 Distributed bag of words의 2가지 알고리즘이 있다. 마지막으로

char2vec은 단어(Word)보다 더 작은 단위인 문자(Character)를 사용하는 기법이다.

3. 이진 분류 모델의 고객 행태 예측 방안

3.1 이종 정보 결합의 CNN 모델

본 연구에서는 정형 정보(Structured Information)와 비정형 정보(Unstructured Information)를 결합하고 이를 CNN 기법에서 지도 학습하는 ‘이종 정보 결합의 CNN 모델’을 제안한다. 제안 모델의 프로세스를 도식화하면 <Figure 3>과 같다.



<Figure 3> Process of Proposed Model

제안 모델에서 입력 정보의 결합 방식과 합성곱 신경망 네트워크 계층 설계는 모델 성능에 영향을 줄 수 있는 중요한 요소이다. 따라서 먼저 합성곱 신경망의 전체 구조를 정의하고 그 다음으로 정의된 합성곱 신경망의 성능과 관련이 있는 합성곱 신경망의 내부 구조를 함께 살펴보고자 한다. 앞서 기술한 합성곱 신경망의 전체 구조와 내부 구조를 본 연구의 4가지 아키텍처로 정의하고 이를 순서대로 정리하면 ‘합성곱 신경망 설계(Convolution Neural Network Design)’, ‘이종 정보 결합(Heterogeneous Information Integration)’, ‘비정형 정보 벡터 전환(Unstructured Information Vector Conversion)’, 그리고 ‘다층 퍼셉트론 설계(Multi-layer Perceptron Design)’이다.

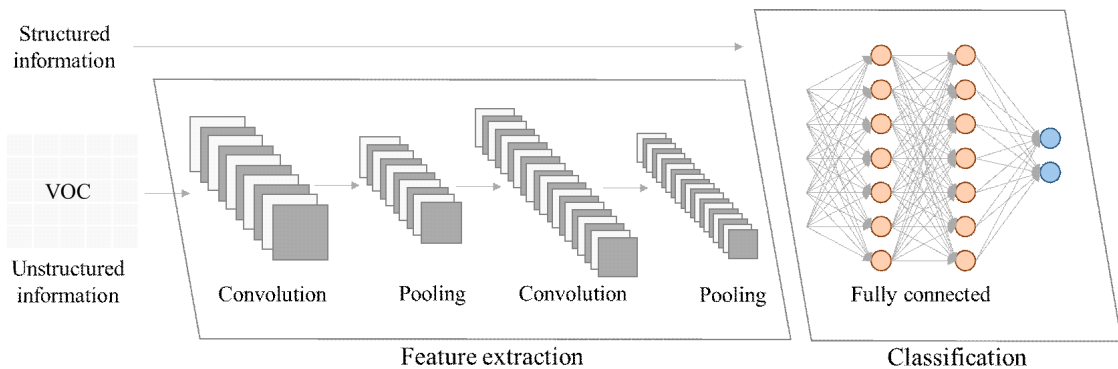
3.1.1 합성곱 신경망 설계

합성곱 신경망 설계(Convolution Neural Network Design)에서는 비정형 정보인 VOC를 입력 받아 합성곱 계층을 사용하여 특징을 추출하고 그 결과를 정형 정보와 결합하여 완전 연결의 다층 퍼셉트론에서 학습시키는 구조로 정의되며 이를 도식화하면 <Figure 4>와 같다.

비정형 VOC 데이터를 2차원의 행렬로 변환하는 방식, 정형 정보와 결합하는 구조, 그리고 다층 퍼셉트론 구조는 3.1.2부터 3.1.4에서 기술하는 3가지의 아키텍처에서 성능을 평가하고 그 결과에 의해서 결정하도록 한다. 그 외에 합성곱 신경망 내부는 2개의 합성곱 계층 구조에서 32개의 5×5 필터와 5×5 최대 풀링을 사용하며 5×5 최대 풀링으로 정의한 목적은 데이터의 공간적

<Table 1> Architecture

Index	Architecture	Description
3.1.1	Convolution Neural Network Design	Design of convolution filters and pooling
3.1.2	Heterogeneous Information Integration	Method of structured and unstructured information integration
3.1.3	Unstructured Information Vector Conversion	Method of converting unstructured VOC data to vector
3.1.4	Multi-layer Perceptron Design	Design of hidden layers and nodes



<Figure 4> CNN Design

축소가 원활하게 진행되도록 하기 위함이다.

3.1.2 이종 정보 결합

이종 정보 결합(Heterogeneous Information Integration)은 서로 다른 형식인 정형과 비정형 정보가 결합되는 위상 구조에 따라 5가지 유형으로 정의하고 각각에 대해서 성능을 평가하여 최적의 성능을 나타내는 유형을 선택하여 제안 모델에서 사용하고자 한다. 5가지 유형으로는 수평적 다층 결합(Horizontal Deep Integration), 이중 단층과 다층 결합(Double Wide & Deep Integration), 이중 다층과 단층 결합(Double Deep & Wide Integration), 수평적 단층 결합(Horizontal Wide Integration), 이중 다층 결합(Double Deep & Deep Integration)이 있다.

첫 번째 수평적 다층 결합은 정형과 비정형 정보를 구분 없이 동일 수준의 입력 변수로 처리하여 2개의 은닉 계층과 하나의 출력 노드로 연결시키는 구조이다. 다차원으로 변환된 비정형 정보를 정형 정보와 동일한 수준으로 사용한다는 특징을 가지고 있다.

두 번째 이중 단층과 다층 결합은 정형과 비정형 정보를 구분하여 입력 변수로 사용한다. 낮은 차원의 정형 정보는 출력 노드로 바로 연결시키고 다차원의 비정형 정보는 2개의 은닉 계층을 통과하여 출력 노드로 연결시키는 구조이다. 차원 수가 높은 비정형 정보만을 2개의 은닉 계층을 연결시킨다는 특징을 가진다.

세 번째 이중 다층과 단층 결합도 앞선 두 번째 결합과 마찬가지로 정형과 비정형 정보를 구분하여 입력 변수로 사용한다. 낮은 차원의 정형 정보를 2개의 은닉 계층을 통과시킨 후 출력 노드로 연결시키며 다차원의 비정형 정보는 출력

노드로 바로 연결시키는 구조이다. 차원 수가 낮은 정형 정보만을 2개의 은닉 계층을 연결시키는 특징을 가진다.

네 번째 수평적 단층 결합은 정형과 비정형 정보를 구분하여 입력 변수로 사용하는 점은 두 번째와 세 번째 결합과 동일하지만 정형과 비정형 정보를 각각 1차의 출력 노드로 연결시키고 이를 다시 2차의 출력 노드로 연결시키는 구조이다.

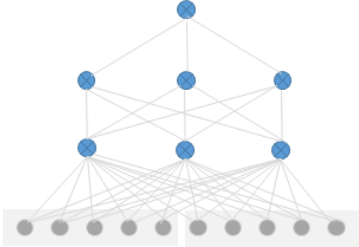
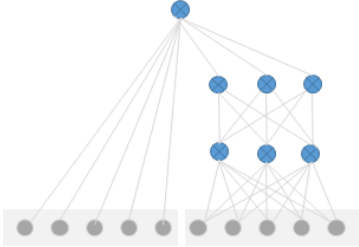
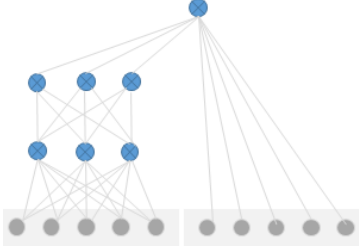
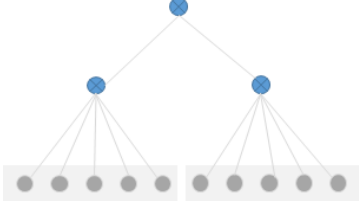
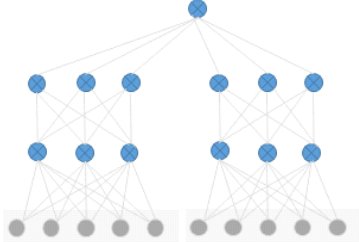
마지막 다섯 번째 이중 다층 결합 역시 정형과 비정형 정보를 구분하여 사용하며 각각에 대해서 2개의 은닉 계층을 통과시키고 그 결과를 출력 노드로 연결시키는 구조이다.

3.1.3 비정형 정보 벡터 전환

비정형 정보 벡터 전환(Unstructured Information Vector Conversion)은 비정형 정보인 VOC 데이터를 벡터로 변환시키는 과정을 다루게 된다. VOC를 벡터로 변화하는 과정에는 4가지 중요한 사항이 있으며 이를 순서대로 살펴보면 다음과 같다. 첫째 어떤 기술을 사용하는 지, 둘째 훈련 학습에 사용되는 코퍼스(Corpus, 말뭉치)의 기준 대상이 어떻게 되는지, 셋째 VOC가 가지고 있는 조사와 명사 등 모두 어휘를 사용하는지 아니면 가지고 있는 의미가 적은 조사와 같은 단어들은 찾아서 제외시키는지, 넷째, 몇 차원 벡터로 전환 하는지 이다. 이 4가지 사항에 대해서 최적의 성능을 나타내는 방식을 선정하여 제안 모델에서 사용하고자 한다.

먼저 임베딩 기술과 관련해서는 word2vec과 doc2vec, 그리고 char2vec 이렇게 3개의 임베딩 기법을 비교하고자 한다. 다음으로 훈련 학습에 사용하는 코퍼스 대상은 내부 도메인(In-Domain)

〈Table 2〉 Heterogeneous Information Integration

Integration Configuration	Description	Topology
Horizontal Deep Integration	The structured and unstructured information consists of two hidden layers using the same level input variables	
Double Wide & Deep Integration	The unstructured information consists of two hidden layers, the structured information is connected to the output nodes without hidden layers, and the structured and unstructured information are combined at the output nodes	
Double Deep & Wide Integration	The structured information consists of two hidden layers, the unstructured information is connected to the output nodes without hidden layers, and the structured and unstructured information are combined at the output nodes	
Horizontal Wide Integration	The structured and unstructured information are connected to the output nodes without hidden layers respectively, then both information are combined at next output nodes	
Double Deep & Deep Integration	structured and unstructured information go through two hidden layers and combine them at the output node	

과 외부 도메인(Out-of-Domain)으로 구분하여 비교하고자 한다. 여기서 내부 도메인은 실험 대상 기업의 VOC를 사용하여 훈련 학습을 시키고 그 결과를 가지고 하나 하나의 VOC를 벡터로 전환하는 것을 의미한다. 반면 외부 도메인은 VOC와 관련 없는 텍스트 데이터셋으로 먼저 학습을 시키고 그 결과를 가지고 VOC를 벡터로 전환하게 된다. 세번째 VOC로부터 사용하는 대상 어휘를 모든 단어를 그 대상으로 사용하는 경우와 전체 VOC를 기준으로 출현 빈도가 높은 상위 100개의 단어를 제외시키는 경우 이렇게 2가지 경우로 구분하고자 한다. 마지막으로 변환되는 벡터의 차원 수와 성능과의 영향을 검토하기 위해서 100차원(100D), 300차원(300D), 500차원(500D)의 3가지의 경우로 구분하고자 한다.

3.1.4 다층 퍼셉트론 설계

다층 퍼셉트론 설계(Multi-layer Perceptron Design)에서는 은닉 계층과 은닉 노드의 수를 가지고 모두 3가지의 경우에 대해서 살펴보고자 한다. 첫 번째 경우에 해당하는 Case1은 128개의 은닉 노드를 가지는 1개의 은닉 계층으로 구성된다. 두 번째 Case2는 64개와 32개의 은닉 노드를 각각 가지는 2개의 은닉 계층으로 구성되며 마지막 세 번째 Case3는 128개와 64개의 은닉 노드를 각각 가지는 2개의 은닉 계층으로 정의하여 성능을 비교해보고자 한다. 또 학습 횟수(Epoch)는 10회, 50회, 100회로 구분하여 성능을 검토하고자 한다.

은닉 계층과 은닉 노드, 그리고 학습 횟수는 그 수가 증가하면 일반적으로 복잡한 패턴을 학습할 수 있다는 장점이 있는 반면 훈련 데이터에만 최적화되는 과적합(Overfitting)의 단점을 가질 수 있다. 따라서 다층 퍼셉트론 설계에서 이를 검토하여 최적의 성능을 갖는 제안 모델을 찾고자 한다.

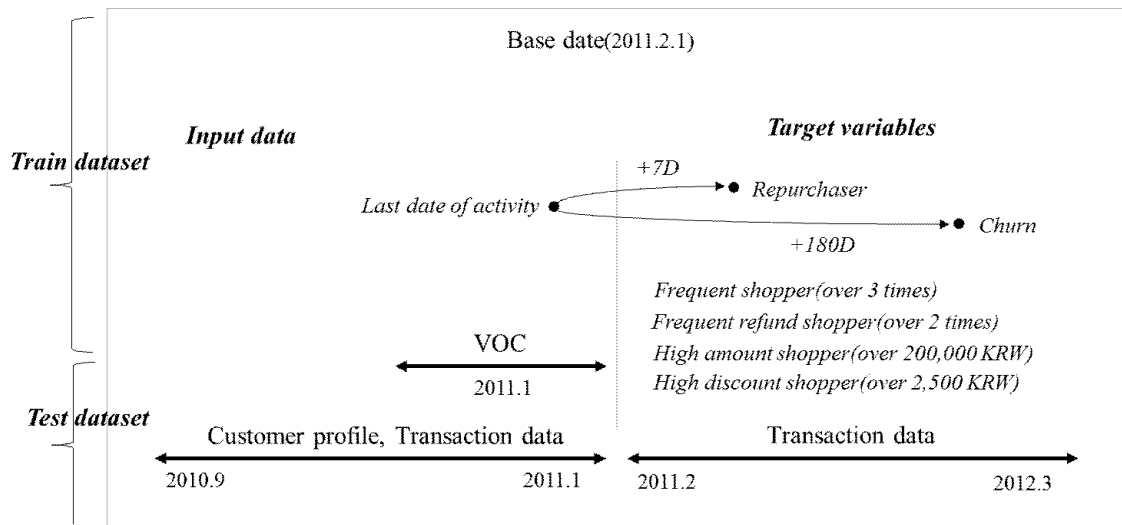
3.2 실험

3.2.1 이진 분류 모델의 정의

전자상거래 기업이 효율적인 마케팅 활동을 하기 위한 고객 세분화 방안에 부합할 수 있도록 기준으로 수립하고자 한다. 이러한 원칙에 의해서 고객의 구매 가능성과 기업에 제공하는 기여 가치를 큰 축으로 하여 이진 분류 모델을 정의하면 다음과 같다. 총 6가지의 이진 분류 모델의 목표 변수는 재구매 고객(Repurchaser), 이탈 고객(Churn), 고빈도 구매 고객(Frequent shopper), 고빈도 반품 고객(Frequent refund shopper), 고단가 구매 고객(High amount shopper), 고할인 구매 고객(High discount shopper)이다. 이진 분류의 목표 변수 중에서 재구매 고객과 이탈 고객 등 2가지 변수에 대해서는 고객의 마지막 활동 일자를 기준으로 하여 그 이후 기간의 구매 여부를 가지고 결정되며 그 외에 4가지 목표 변수는 기준일(2011.2.1) 이후의 구매 특성으로 판단한다. 6개 목표 변수의 세부 정의는 <Table 3>에서와 같이 기술되며 이를 <Figure 5>로 도식화 할 수 있다.

〈Table 3〉 Definition of Target Variables

Target Variables	Definition
Repurchaser	Customers purchasing within 7 days after the last date of activity(last date of purchase date or VOC registration date)
Churn	Customers who will not make purchases within 180 days(6 months) after the last date of activity(last date of purchase date or VOC registration date)
Frequent shopper	Customers purchasing more than 3 times in total after base date ※ The cumulative ratio of 50% customers is 3 times
Frequent refund shopper	Customers who return more than 2 times in total after Base date ※ The cumulative ratio of 75% customers is 2 times
High amount shopper	Customers who purchase more than KRW 200,000 as the total sum after Base date ※ The cumulative ratio of 50% customers is KRW 216,875
High discount shopper	Customers who purchase at a discount of KRW 2,500 or more after base date ※ The cumulative ratio of 75% customers is KRW 2,500



〈Figure 5〉 Experiment dataset

〈Table 4〉 Statistic of Target Variables

Target Variables	T/F	Frequency	Ratio
Repurchaser	T	7,504	31.76%
	F	16,121	68.24%
Churn	T	20,877	88.37%
	F	2,748	11.63%
Frequent shopper	T	11,429	48.38%
	F	12,196	51.62%
Frequent refund shopper	T	6,117	25.89%
	F	17,508	74.11%
High amount shopper	T	11,567	48.96%
	F	12,058	51.04%
High discount shopper	T	5,758	24.37%
	F	17,867	75.63%

3.2.2 실험 데이터

국내 A 온라인 쇼핑몰의 고객과 거래 정보, 그리고 텍스트로 작성된 VOC 정보를 실험 데이터로 하며 데이터 추출 기준은 다음과 같다. 2011년 1월(1개월) 중 VOC를 한 건 이상 등록한 모든 고객 47,947명을 대상으로 정의하고 이 대상 고객들의 고객 프로파일과 2010년 9월부터 2012년 3월까지 총 19개월간의 거래 데이터, 그리고 1개월간 등록된 VOC 데이터이다. 고객 프로파일에는 성별(Gender), 자녀 유무(Status of children), 결혼 여부(Marital status), 나이(Age)로 구성된다. 이 중 성별은 여성과 남성이 73.2%와 약 26.1%, 자녀 유무는 자녀 있음과 없음이 20.8%와 약 79.2%, 결혼 여부는 결혼한 경우와

안한 경우가 약 33.8%와 약 66.2%, 나이는 30대가 가장 많은 53.2%와 그 다음으로 40대와 20대가 25.2%와 14.7%로 각각 분포된다. 거래 정보는 구매 횟수(Number of purchase), 반품 횟수(Number of refund), 구매 금액(Amount of purchase), 할인 금액(Amount of discount), B2C 구매 횟수(Number of purchase in B2C)와 B2E(임직원) 구매 횟수(Number of purchase in B2E)로 구성된다. 이 중 구매 횟수의 평균은 4.69회, 반품 횟수의 평균은 1.68회, 구매금액의 평균은 506,802원, 할인 금액의 평균은 2,196원, B2C와 B2E 구매 횟수는 각각 3.91회와 0.79회를 나타냈었다. VOC는 고객이 질문 형식으로 직접 인터넷 상에 등록하는 텍스트이며 주요 내용은 배송, 반품, 취소, 상품과 관련된 질문이나 불만이었다.

〈Table 5〉 Demographic variables

Variables	Component	Frequency	Ratio
Gender	Male	12,513	26.1%
	Female	35,102	73.2%
	Nan	332	0.7%
Status of children	T	9,983	20.8%
	F	37,964	79.2%
Marital status	T	16,213	33.8%
	F	31,734	66.2%
Age	Teenager	126	0.3%
	20s	7,077	14.7%
	30s	25,526	53.2%
	40s	12,213	25.5%
	50s	2,318	4.8%
	More than 60	355	0.7%
	Nan	332	0.7%

〈Table 6〉 Transaction variables

Variables	Statistics		Variables	Statistics	
Number of purchase	mean	4.69	Number of refund	mean	1.68
	std.	17.32		std.	3.15
	min	1.0		min	0.0
	25%	1.0		25%	0.0
	50%	3.0		50%	1.0
	75%	5.0		75%	2.0
Amount of purchase	mean	506,802	Amount of discount	mean	2,196
	std.	2,349,575		std.	4,595
	min	1		min	0
	25%	90,058		25%	0
	50%	216,875		50%	0
	75%	517,120		75%	2,500
Number of purchase in B2C	mean	3.91	Number of purchase in B2E	mean	0.79

3.2.3 실험 구성

본 연구의 실험은 2단계로 구분된다. 먼저 1단계에서는 제안 모델 성능의 영향을 주는 3가지 아키텍처인 ‘이중 정보 결합’과 ‘비정형 정보 벡터 전환’, 그리고 ‘다층 퍼셉트론 설계’를 평가한다. 이 평가는 인공 신경망(ANN) 기법을 기준으로 하여 성능 평가를 진행한다. 앞의 평가 결과로부터 최적의 성능을 나타내는 방식을 채택하여 ‘합성곱 신경망 설계’의 아키텍처를 확정하는 절차로 하여 본 연구의 제안 모델을 확정한다. 다음 2단계에서는 확정된 제안 모델을 가지고 이진 분류의 목표 변수에 대한 성능을 평가한다. 실험 과정에서 과대적합이나 과소적합이 발생할 수 있어 오버 샘플링(Over Sampling) 기법을 사용하여 각 목표 변수별로 빈도수가 높은 쪽으로 균형을 맞추도록 빈도수가 낮은 쪽에서 무작위 반복 샘플링하고 실험별로 5회의 교차 검증(k-fold cross validation)을 실행하였다.

3.3 실험 결과

3.3.1 아키텍처 성능 평가

3가지 아키텍처의 성능 평가 결과를 살펴보면 다음과 같다. 첫째로 ‘이중 정보 결합’에서는 수평적 다층 결합(Horizontal Deep Integration)의 성능이 가장 높게 평가되었다. 그 다음 순으로는 이중 단층과 다층 결합(Double Wide & Deep Integration)과 이중 다층 결합(Double Deep & Deep Integration)의 성능이 높게 평가되었다. 반면 이중 다층과 단층 결합(Double Deep & Wide Integration)과 수평적 단층 결합(Horizontal Wide Integration)의 성능이 가장 낮게 평가되었다. 이러한 결과는 다차원의 비정형 정보를 다층 퍼셉트론 구조에서 학습시킨 결과가 그렇지 않은 경

우에 비해 상대적으로 높은 성능을 나타낸다는 점을 보여주고 있으며 이는 다중 처리 계층(Multiple processing layers)이 여러 수준의 추상화를 통해 데이터의 표현을 학습하게 되고 역전파 알고리즘(Back propagation algorithm)을 사용하여 대용량 데이터 집합의 복잡한 구조를 발견하는데 용이한 기법이라는 이론적 배경을 지지하는 실험 결과라 할 수 있겠다.

둘째로 ‘비정형 정보 벡터 전환’의 성능은 외부 도메인(Out-of-Domain)의 코퍼스(Corpus)에서 doc2vec 기법으로 500D로 전환한 경우가 가장 높게 평가되었다. 코퍼스 관점에서는 내부 도메인(In-Domain)의 코퍼스에서 사용되는 단어(Word)나 문자(Character)가 전자상거래에 국한되는데 비해 외부 도메인이 보다 다양한 단어와 문자를 사용하고 있어 더 높은 성능을 나타낸 것으로 해석된다. 워드 임베딩 기법 관점에서는 문서(Document) 단위의 벡터 전환 기법인 doc2vec이 단어나 문자 단위의 word2vec과 char2vec보다 높은 성능을 나타냈으며 word2vec과 char2vec을 비교해보면 char2vec의 성능이 word2vec에 비해서 성능이 높은 것으로 평가되었다. 다만 합성곱 신경망 특성상 입력값을 2차원의 벡터로 구성하기에는 doc2vec 기법이 한계를 가지고 있어 이를 배제하고 char2vec 기법을 채택한다. 하지만 비교대상 기법인 NBC, SVM, ANN에서는 가장 높은 성능을 나타낸 doc2vec 기법을 사용하여 평가하도록 한다. 마지막으로 전환되는 벡터의 차원이 높을수록 성능이 향상되는 경향을 보였지만 300D에서 500D로 차원 증가할 때는 그 증가 폭이 둔화된 것으로 확인되었다. 셋째로 ‘다층 퍼셉트론 설계’는 계층과 노드 수가 가장 큰 Case3(Layer: 2, Node: 128, 64)의 성능이 가장 높게 평가되었다. 그 다음으로는 계층 수는 적지만

<Table 7> Architectures of The Proposed Model

Index	Configuration mode	Optimal performance
3.1.1	Convolution Neural Network Design	char2vec, 32 5×5 Filters, 5×5 Max-pooling, 2 Convolution layers
3.1.2	Heterogeneous Information Integration	Horizontal Deep Integration
3.1.3	Unstructured Information Vector Conversion	Out-of-Domain, 300D ※ NBC, SVM , ANN : doc2vec
3.1.4	Multi-layer Perceptron Design	Case3(Layer: 2, Node: 128, 64)

노드의 수가 많은 Case1(Layer: 1, Node: 128)이 Case2(Layer: 2, Node: 64, 32) 보다 상대적으로 높은 성능을 보였다. 위와 같은 3가지의 아키텍처 성능 평가 결과를 기준으로 다음과 같은 구조를 가지도록 ‘합성곱 신경망 설계’ 아키텍처를 확정하였다.

3.3.2 이진 분류 모델의 목표 변수 평가

본 연구의 제안 모델인 이중 정보 결합의 CNN 모델을 가지고 6개 목표 변수로 하여 실험을 진행한 결과 비교 대상인 NBC, SVM, 그리고 ANN을 활용한 경우에 비해서 예측 정확도가 향상된 것으로 평가되었으며 세부 실험 결과를 정리하면 다음과 같다.

첫째, 입력 변수로 정형과 비정형 정보를 결합하여 사용한 경우가 정형 정보만을 활용한 경우에 비해서 NBC, SVM, ANN 기법 모두에서 예측 정확도가 향상된 것으로 평가되었다. <Table 8>에서 재구매 고객(Repurchaser)을 예측한 결과를 보면 ANN 기법으로 정형 정보만을 입력 변수로 사용했을 경우가 57.18%인 것에 비하여 정형과 비정형을 결합하여 사용한 경우에는 66.42%로 향상되었다. 이는 정형 정보와 함께 비정형 정보를 함께 활용하는 것이 정형 정보만을 기반으로

한 분석 접근 방법에 비해서 보다 효과적이라는 사실을 보여준다.

둘째, 정형과 비정형 정보를 결합하여 사용한 경우에서 NBC나 SVM 기법에 비해서 ANN과 CNN의 예측 정확도가 높게 평가되었으며 특히 CNN이 ANN에 비해서도 높은 예측 정확도를 나타낸 것으로 평가되었다. <Table 9>에서 이탈 고객(Churn)을 예측한 결과로 살펴보면 NBC가 57.56%, SVM이 58.57%, ANN이 74.16%, CNN이 77.89%로 평가되었다. 이 결과는 다차원의 입력 변수로 변환된 비정형 정보를 학습하는데 있어 다층 퍼셉트론 구조가 보다 효과적이라고 해석할 수 있겠다. 또 CNN의 예측 정확도가 ANN에 비해서 높게 평가된 것은 CNN 기법의 특징 추출 알고리즘이 VOC에 사용된 단어들이 분포하고 있는 위치 정보를 활용하여 문장의 의미를 파악하는데 효과적으로 작용되었다고 해석될 수 있겠다.

셋째, 6개 목표 변수를 F1 Measure로 측정된 결과를 살펴보면 <Table 14>에서와 같이 6개 목표 변수 모두에서 제안 모델인 CNN의 성능이 높게 평가되었다. <Table 14>의 고빈도 고객(Frequent shopper)의 F1 Measure로 살펴보면 ANN은 70.01%, CNN은 73.78%로 평가되었다.

〈Table 8〉 Accuracy of Repurchaser

Methods	Structured Information	Unstructured Information	Both
NBC	50.83%(+/-0.34%)	50.38%(+/-0.53%)	51.85%(+/-0.51%)
SVM	51.63%(+/-1.01%)	53.63%(+/-1.83%)	55.48%(+/-1.47%)
ANN	57.18%(+/-1.22%)	61.59%(+/-1.49%)	66.42%(+/-0.83%)
CNN	-	62.99%(+/-1.20%)	67.07%(+/-1.17%)

〈Table 9〉 Accuracy of Churn

Methods	Structured Information	Unstructured Information	Both
NBC	56.12%(+/-1.36%)	51.75%(+/-0.67%)	57.56%(+/-0.40%)
SVM	57.15%(+/-1.31%)	55.32%(+/-1.23%)	58.57%(+/-1.40%)
ANN	68.23%(+/-0.53%)	66.48%(+/-1.43%)	74.16%(+/-1.05%)
CNN	-	67.49%(+/-1.12%)	77.89%(+/-0.18%)

〈Table 10〉 Accuracy of Frequent shopper

Methods	Structured Information	Unstructured Information	Both
NBC	53.33%(+/-2.01%)	50.29%(+/-0.26%)	57.46%(+/-0.14%)
SVM	57.59%(+/-1.17%)	54.14%(+/-1.11%)	63.22%(+/-1.07%)
ANN	68.54%(+/-0.70%)	67.83%(+/-1.82%)	71.02%(+/-0.65%)
CNN	-	69.71%(+/-1.59%)	73.55%(+/-1.17%)

〈Table 11〉 Accuracy of Frequent refund shopper

Methods	Structured Information	Unstructured Information	Both
NBC	50.85%(+/-0.46%)	50.28%(+/-0.27%)	52.07%(+/-0.51%)
SVM	58.12%(+/-0.80%)	57.59%(+/-1.40%)	58.34%(+/-1.19%)
ANN	72.51%(+/-0.67%)	71.01%(+/-1.44%)	76.59%(+/-1.24%)
CNN	-	73.64%(+/-1.65%)	78.23%(+/-1.42%)

〈Table 12〉 Accuracy of High amount shopper

Methods	Structured Information	Unstructured Information	Both
NBC	51.29%(+/-0.36%)	51.83%(+/-1.14%)	56.65%(+/-1.03%)
SVM	58.93%(+/-0.27%)	60.83%(+/-1.32%)	66.11%(+/-1.53%)
ANN	75.22%(+/-0.52%)	78.77%(+/-1.43%)	85.30%(+/-1.49%)
CNN	-	82.88%(+/-1.64%)	87.66%(+/-2.16%)

〈Table 13〉 Accuracy of High discount shopper

Methods	Structured Information	Unstructured Information	Both
NBC	50.84%(+/-0.29%)	51.15%(+/-0.32%)	52.11%(+/-0.48%)
SVM	54.58%(+/-1.61%)	52.07%(+/-1.30%)	58.40%(+/-1.51%)
ANN	64.57%(+/-0.91%)	62.88%(+/-1.21%)	77.40%(+/-1.69%)
CNN	-	67.52%(+/-1.52%)	79.18%(+/-1.21%)

〈Table 14〉 F1 Measure of All Target Variables

Target Variables	ANN (Structured and Unstructured)	CNN (Structured and Unstructured)
Repurchaser	53.66%(+/-1.37%)	55.03%(+/-1.42%)
Churn	78.86%(+/-1.21%)	85.45%(+/-1.51%)
Frequent shopper	70.01%(+/-1.58%)	73.78%(+/-1.29%)
Frequent refund shopper	60.57%(+/-1.26%)	64.12%(+/-1.52%)
High amount shopper	80.97%(+/-1.61%)	85.16%(+/-1.33%)
High discount shopper	63.17%(+/-1.43%)	65.72%(+/-1.55%)

4. 결론

본 연구에서는 전자상거래 고객 행태를 효과적으로 예측하기 위해 정형 정보와 비정형 정보를 결합하고 이를 합성곱 신경망(CNN) 기법으로 활용하는 ‘이중 정보 결합의 CNN 모델’을 제시하였다. 최적의 제안 모델을 정의하기 위해서

정형과 비정형 정보의 결합과 VOC 데이터를 벡터로 전환, 그리고 다층 퍼셉트론의 계층과 노드 구성을 대상으로 하는 3가지 아키텍처를 ‘이중 정보 결합’, ‘비정형 정보 벡터 전환’, 그리고 ‘다층 퍼셉트론 설계’라고 정의하고 각각의 아키텍처 단위로 실험을 통해서 최적의 성능을 나타내는 구조를 확인하여 합성곱 신경망에 사용하는

방식으로 제안 모델을 확정하였다. 제안 모델을 가지고 재구매 고객과 이탈 고객 등 총 6개의 이진 분류 목표 변수를 정의하고 성능을 평가하였다. 평가 결과 정형과 비정형 정보를 결합하여 합성곱 신경망 기법으로 활용한 제안 모델이 NBC, SVM, ANN에 비해 예측 정확도와 F1 Measure가 높게 평가되었다.

연구 결과를 바탕으로 확인한 본 연구의 의의를 다음과 같이 정리할 수 있다. 첫째, 비정형 정보의 활용이 고객 행태 예측 정확도 향상에 기여한다는 점을 확인하였다. 둘째, 다차원으로 변환된 비정형 정보는 다층 퍼셉트론을 기반으로 한 ANN과 CNN을 활용하는 것이 보다 적절한 접근 방법이라는 점과 특히 텍스트로 작성된 VOC 데이터에서 문맥의 의미를 파악하고 해석하는데 있어 CNN 기법이 보다 효과적이라는 사실을 실험 결과를 통해서 확인했다는 데 의의가 있다고 하겠다. 셋째, 전자상거래 기업의 실제 데이터를 바탕으로 한 실증적인 연구를 통해서 고객이 직접 텍스트 형식으로 작성한 VOC 데이터로부터 고객 행태 예측에 있어 매우 유의미한 정보를 추출할 수 있다는 점을 입증한데 그 의미가 크다고 하겠다. 넷째, 여러가지 실험을 통해서 제안 모델을 구성하는 아키텍처에서의 파라미터 선택과 그에 따른 성능과의 관계를 정리함으로써 향후 이와 관련된 연구를 진행하는데 있어 유익한 정보를 제공했다고 할 수 있겠다.

이번 연구를 계기로 향후에는 다중 분류와 연속형 목표 변수를 대상으로 하여 연구 범위를 확대한다면 보다 의미가 있을 것으로 보인다. 다음으로 금융, 통신 등과 같은 다양한 분야와 음성 통화 기록과 같은 다양한 형식의 데이터로 실증 검증의 범위를 확대하여 일반화하면 연구 가치가 높을 것으로 판단된다.

참고문헌(References)

- Ahn, S., "Deep learning architectures and applications," *Journal of Intelligence and Information Systems*, 22(2), (2016), 127-142.
- Chu, H., S. Ahn, and S. Kim, "AlphaGo's artificial intelligence algorithm analysis", *Software Policy & Research Institute*, (2016).
- Coussement, K., D. Van den Poel, "Integrating the voice of customers through call center emails into a decision support system for churn prediction," *Information & Management*, 45(3), (2008), 164-174.
- Gridach, M., H. Haddad, and H. Mulki, "Churn identification in microblogs using convolutional neural networks with structured logical knowledge," *Paper presented at the Proceedings of the 3rd Workshop on Noisy User-Generated Text*, (2017), 21-30.
- Kim, K., B. Lee, and J. Kim, "Feasibility of Deep Learning Algorithms for Binary Classification Problems," *Journal of Intelligence and Information Systems*, 23(1), (2017), 95-108.
- Kim, S., J. Song, and K. Lee, "A Study of customer churn by analysing CRM customer data," *Asia Marketing Journal*, 7(1), (2005), 21-42.
- Kim, Y., "Convolutional neural networks for sentence classification," *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (2014), 1746-1751
- Krizhevsky, A., I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Paper presented at the Advances in Neural Information Processing Systems*, (2012),

- 1097-1105.
- Le, Q., and T. Mikolov, "Distributed representations of sentences and documents," *Paper presented at the International Conference on Machine Learning*, (2014), 1188-1196.
- LeCun, Y., Y. Bengio, and G. Hinton, "Deep learning," *Nature*, 521(7553), (2015), 436-444.
- Lee, J., J. Kim, "Integrated use of classification and association rule for real-time CRM: Application of predicting credit card customer churn," *KMIS International Conference*, (2007), 135-140.
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Paper presented at the Advances in Neural Information Processing Systems*, (2013), 3111-3119.
- Schmidhuber, J. "Deep learning in neural networks: An overview," *Neural Networks*, 61, (2015), 85-117.
- Yiğit, İ. O., A. F. Ateş, M. Güvercin, H. Ferhatosmanoğlu, and B. Gedik, "Call center text mining approach," *Paper presented at the Signal Processing and Communications Applications Conference (SIU)*, 2017 25th, (2017), 1-4.
- Yu, E., J. Kim, C. Lee, and N. Kim, "Using ontologies for semantic text mining," *Journal of Information Systems*, 21(3), (2012), 137-161.
- Zhang, X., J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," *Paper presented at the Advances in Neural Information Processing Systems*, (2015), 649-657.

Abstract

Customer Behavior Prediction of Binary Classification Model Using Unstructured Information and Convolution Neural Network: The Case of Online Storefront

Seungsoo Kim* · Jongwoo Kim**

Deep learning is getting attention recently. The deep learning technique which had been applied in competitions of the International Conference on Image Recognition Technology(ILSVR) and AlphaGo is Convolution Neural Network(CNN). CNN is characterized in that the input image is divided into small sections to recognize the partial features and combine them to recognize as a whole. Deep learning technologies are expected to bring a lot of changes in our lives, but until now, its applications have been limited to image recognition and natural language processing.

The use of deep learning techniques for business problems is still an early research stage. If their performance is proved, they can be applied to traditional business problems such as future marketing response prediction, fraud transaction detection, bankruptcy prediction, and so on. So, it is a very meaningful experiment to diagnose the possibility of solving business problems using deep learning technologies based on the case of online shopping companies which have big data, are relatively easy to identify customer behavior and has high utilization values. Especially, in online shopping companies, the competition environment is rapidly changing and becoming more intense. Therefore, analysis of customer behavior for maximizing profit is becoming more and more important for online shopping companies.

In this study, we propose 'CNN model of Heterogeneous Information Integration' using CNN as a way to improve the predictive power of customer behavior in online shopping enterprises. In order to propose a model that optimizes the performance, which is a model that learns from the convolution neural network of the multi-layer perceptron structure by combining structured and unstructured information, this

* Dept. of Business Administration, Graduate School, Hanyang University

** Corresponding author: Jongwoo Kim

School of Business, Hanyang University

222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Korea

Tel: +82-2-2220-1067, Fax: +82-2-2220-1169, E-mail: kjw@hanyang.ac.kr

model uses 'heterogeneous information integration', 'unstructured information vector conversion', 'multi-layer perceptron design', and evaluate the performance of each architecture, and confirm the proposed model based on the results. In addition, the target variables for predicting customer behavior are defined as six binary classification problems: re-purchaser, churn, frequent shopper, frequent refund shopper, high amount shopper, high discount shopper.

In order to verify the usefulness of the proposed model, we conducted experiments using actual data of domestic specific online shopping company. This experiment uses actual transactions, customers, and VOC data of specific online shopping company in Korea. Data extraction criteria are defined for 47,947 customers who registered at least one VOC in January 2011 (1 month). The customer profiles of these customers, as well as a total of 19 months of trading data from September 2010 to March 2012, and VOCs posted for a month are used. The experiment of this study is divided into two stages. In the first step, we evaluate three architectures that affect the performance of the proposed model and select optimal parameters. We evaluate the performance with the proposed model.

Experimental results show that the proposed model, which combines both structured and unstructured information, is superior compared to NBC(Naïve Bayes classification), SVM(Support vector machine), and ANN(Artificial neural network). Therefore, it is significant that the use of unstructured information contributes to predict customer behavior, and that CNN can be applied to solve business problems as well as image recognition and natural language processing problems. It can be confirmed through experiments that CNN is more effective in understanding and interpreting the meaning of context in text VOC data. And it is significant that the empirical research based on the actual data of the e-commerce company can extract very meaningful information from the VOC data written in the text format directly by the customer in the prediction of the customer behavior. Finally, through various experiments, it is possible to say that the proposed model provides useful information for the future research related to the parameter selection and its performance.

Key Words : Customer Behavior Prediction, Deep Learning, Convolution Neural Network(CNN), Voice of Customer(VOC)

Received : May 21, 2018 Revised : June 14, 2018 Accepted : June 16, 2018

Publication Type : Regular Paper Corresponding Author : Jongwoo Kim

저 자 소개



김 승 수

현재 신한금융투자(주) 빅데이터센터에 재직 중이다. 한양대학교 일반대학원 경영학과 박사과정을 수료하였고, 주요 관심분야는 데이터마이닝, 기계학습 및 딥러닝 기법의 응용, 빅데이터, 상품추천기술 등이다.



김 종 우

현재 한양대학교 경영대학 경영학부 교수로 재직 중이다. 서울대학교 수학과에서 학사를 마쳤으며, 한국과학기술원에서 경영과학으로 석사학위를, 산업경영학으로 박사학위를 취득하였다. 주요 연구 관심분야는 데이터마이닝 기법과 응용, 기계학습과 딥러닝, 오피니언 마이닝, 상품추천기술, 지능형 정보시스템, 집단지성, 사회 네트워크 분석, 클라우드 컴퓨팅 서비스 등이다.