디지털산업정보학회 논문지 제15권 제3호-2019년 9월
http://dx.doi.org/10.17662/ksdim.2019.15.3.197

혼성 표본 추출과 적층 딥 네트워크에 기반한 은행 텔레마케팅 고객 예측 방법

이 현 진*

A Method of Bank Telemarketing Customer Prediction based on Hybrid Sampling and Stacked Deep Networks

Lee Hyunjin

〈Abstract〉

Telemarketing has been used in finance due to the reduction of offline channels. In order to select telemarketing target customers, various machine learning techniques have emerged to maximize the effect of minimum cost. However, there are problems that the class imbalance, which the number of marketing success customers is smaller than the number of failed customers, and the recall rate is lower than accuracy. In this paper, we propose a method that solve the imbalanced class problem and increase the recall rate to improve the efficiency. The hybrid sampling method is applied to balance the data in the class, and the stacked deep network is applied to improve the recall and precision as well as the accuracy. The proposed method is applied to actual bank telemarketing data. As a result of the comparison experiment, the accuracy, the recall, and the precision is improved higher than that of the conventional methods.

Key Words : Bank Telemarketing, Deep Learning, Stacked Networks, Class Imbalance

Ⅰ. 서론

은행에서는 금융 상품을 판매할 때 인건비 및 비용 절감 등의 이유로 오프라인 상품판매방식에서 벗어나 다양한 방식들을 도입하고 있다. 정보통신기술이 발전함에 따라 텔레마케팅(Telemarketing), 이메일(e-mail), 챗봇(chatbot)등 비대면 마케팅 방법으로 상품 가입 권유가 활발하게 이루어지고 있다[1]. 텔레마케팅은 전화를 이용한 마케팅 방법으로 마케팅 캠페인을 위해 많이 사용되는 방법 중 하나이다[2]. 텔레마케팅 방법으로는 콜센터(Call Center)의 상담사가 직접 고객에게 전화를 하는 아웃바운드(Outbound) 방식과 고객이 콜센터에 전화했을 때 상품을 권유하는 인바운드(Inbound) 방식이 있다. 텔레마케팅은 전화를 하는 것과 통화가 이루어지는 것 모두가 비용이 소요되는 행위이다. 통화가 이루어진다고 해서 상품 가입 권유가 성공하는 것은 아니기 때문에 상품 가입 권유를 성공할 수 있는 적절한 목표 고객을 선정하는 것이 중요하다. 목표 고객을 선정하기 위하여 다양한 방법이 사용되는데, 마케터의 직관에 의한 규칙 기반 시스템(Rule-based System)과 데이터웨어하우스와 데

* 숭실사이버대학교 ICT공학과 부교수

이터 마이닝을 활용한 비즈니스 인텔리전스 시스템이 있다[3].

텔레마케팅은 고객과 직접 연결이 이루어지기 때문에 고객의 의견을 직접 들을 수 있고, 고객의 반응을 실시간으로 인지할 수 있어서 기업의 입장에서는 중요한 고객과의 접점 수단이다. 하지만, 원하지 않는 고객에게 접촉 하거나, 필요 없는 상품을 추천한다면 고객의 불만이 증가하는 단점이 발생 할 수 있다. 따라서, 텔레마케팅 대상 고객을 선정할 때 전화 통화에 대해 불만을 가지지 않으면서, 상품 가입의 가능성이 높은 고객을 선정하는 것이 중요하다.

본 논문에서는 혼성 표본 추출과 적층 딥 네트워크에기반한 텔레마케팅 대상 목표 고객을 선정하는 방법을 제안 한다. 텔레마케팅을 수행했을 때 상품판매에 성공한 고객의 수는 실패한 고객의 수보다 적을 수밖에 없기 때문에 축적된 데이터는 성공한 고객의 수와 실패 고객의 수가 불균형을 이루는 구조이다. 따라서 혼성 표본 추출(Hybrid Sampling)에 의한 데이터 균형화를 통하여 성공한 고객의 수와 실패한 고객의 수에 대한 균형 잡힌 데이터를 생성한다. 그리고 적층 딥 네트워크(Stacked Deep Networks) 모델을 적용하여 텔레마케팅 구매 가망 고객의 포함을 향상시키는 방법을 제안한다. 본 논문의 구성은 다음과 같다. 2장에서는 텔레마케팅 분석에 대한 기존 연구들을 살펴보고, 3장에서는 제안하는 혼성 표본 추출과 적층 딥 네트워크에 기반한 텔레마케팅 구매 가망 고객 예측 모델에 대해 설명한다. 4장에서는 실제 텔레마케팅 데이터를 이용하여 제안하는 모델의 성능을 살펴보고, 마지막 5장에서 결론과 향후 연구 방향에 대해서 논의한다.

Ⅱ. 관련연구

텔레마케팅은 고객에게 직접적으로 1대 1로 하는 다이렉트 마케팅(Direct marketing)의 한 분야로 다이렉트 마케팅의 큰 비중을 차지하고 있다. 소매업(retail), 이커머스(e-commerce), 은행상품 추천 등에서 다이렉트마케팅의 효율을 높이기 위하여 기계 학습(Machine Learning)방법이 활용되고 있다[4-5].

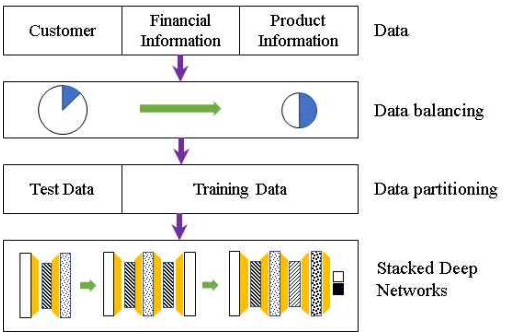
김승수 등은 이커머스 상의 고객 행태 예측력을 높이기 위하여 딥러닝 기술을 이용한 이종 정보 결합의 합성곱 신경망 모델을 제시하였다[6]. Kaefer 등은 해당 브랜드에 대한 다이렉트 마케팅에 반응할 고객을 예측하는데 신경 회로망 기반의 기계 학습 기법을 사용하여, 특정 브랜드에 대한 고객의 충성도를 예측하려고 하였다[7]. Liao 등은 미용 용품에 대한 이커머스를 대상으로 K-means 군집화 알고리즘과 연관 규칙(Association rule)을 이용하여 다이렉트 마케팅의 대상 고객을 선정하는 방법으로 고객의 구매 패턴을 기반으로 고객을 유의미하게 분류를 하였다[8].

Moro 등은 로지스틱 회귀 분석(Logistic Regression), 서포트 벡터 머신(Support Vector Machine) , 신경회로망(Neural Networks) 등을 이용하여 텔레마케팅의 가망 고객을 결정하는 방법을 제안하였다[2]. Javaheri 등은 텔레비전이나 라디오와 같은 매스미디어에 대한 마케팅 캠페인이 은행의 신규 상품 판매에 미치는 영향에 대해 연구하였다[9]. 캠페인이 종료된 후 신규 상품 구매 고객은 캠페인에 영향을 받았다고 가정한 후 서포트 벡터 머신알고리즘을 적용하여 전체 고객 중 40%에 대한 마케팅을 통해 신규 상품 구매 고객의 80%를 특정할 수 있었다. Ladyzinski 등은 개인 대상 금융 회사에서 다이렉트 마케팅의 적중률 향상을 위하여 시계열 기반의 데이터 전처리와 딥러닝 기반의 딥 빌리프 네트워크(Deep Belief Network, DBN)와 랜덤 포레스트(Random Forest, RF)를 결합하는 방법을 제안하였다[10].

Ⅲ. 제안하는 텔레마케팅 구매 가망 고객 예측 모델

3.1 전체 시스템 구성

제안하는 시스템의 전체 구성은 <그림 1>과 같다. 딥러닝 모델을 설계하는 흐름은 먼저 텔레마케팅 성공 고객과 실패 고객의 균형도를 맞추기 위한 데이터 균형화를 수행하고, 모델을 생성하기 위해 학습 데이터와 검증 데이터로 분할한다. 그 후 제안하는 적층 딥 네트워크로 모델을 생성하는 과정으로 이루어져 있다.



<그림 1> 제안하는 방법의 흐름도

3.2 혼성 표본 추출 데이터 균형화

분류할 대상의 균형이 한쪽으로 치우친 경우에는 분류의 정확도가 분류 대상이 많은 범주에 영향을 받아 결과는 대규모 범주(major class)의 정확도를 높이는 방향으로 분류가 이루어진다. 텔레마케팅 예측 모델의 목적은 텔레마케팅 대상 고객을 정확하게 선정하는 것이지만, 텔레마케팅 대상 고객 중 성공 고객은 소규모 범주(minor class)에 속하기 때문에 모델의 결과가 시스템의 목적과는 다른 방향으로 진행된다. 따라서, 소규모 범주에 대한 분류 정확도를 높이는 방법이 필요하며, 이를 위해 적용하는 방법이 데이터 균형화이다.

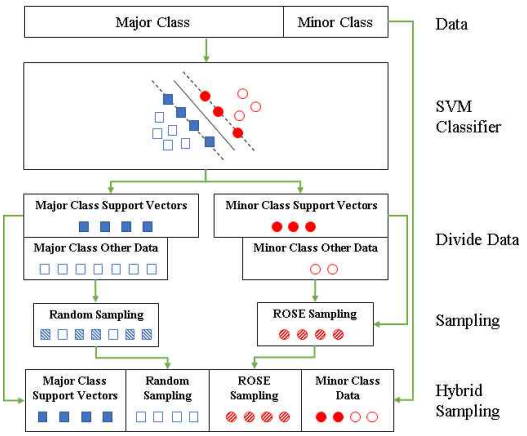
데이터 균형화를 위한 방법은 표본 추출이 주로 사용되며, 대표적인 방법으로는 대규모 범주의 데이터 개수를 줄이는 축소 표본 추출(under sampling) 방법과 소규모 범주의 데이터 개수를 늘리는 확대 표본 추출(over sampling) 방법, 그리고 이 두 가지 방법을 혼합한 혼성 표본 추출(hybrid sampling) 방법이 있다.

축소 표본 추출 방법은 대규모 범주에 속한 데이터들은 중복되는 속성을 가지고 있다는 것을 전제로 임의로 소규모 범주의 데이터 개수만큼 선택하는 임의 표본 추출(random sampling)과 다른 범주와의 경계에 위치한 데이터가 더 중요한 정보를 가지고 있다는 것을 가정으로 한 서포트 벡터 머신 기반의 표본 추출(SVM based under sampling) 방법이 있다. 이 방법들은 범주간의 균형을 보장하면서 데이터의 개수를 줄여서 모델의 성능을 항상 시키면서 모델링 시간을 단축시킬 수 있는 장점이 있다. 하지만, 대규모 범주에서 선택한 데이터가 해당 범주를 대표한다는 것을 보장하지 못하므로 분류 성능이 저하될 가능성이 있다.

확대 표본 추출은 소규모 범주에 속한 데이터의 개수를 대규모 범주에 속한 데이터의 개수만큼 늘리는 방법이다. 중복 확대 표본 추출(duplicated over sampling)은 소규모 범주의 데이터를 복제하는 방법으로 가장 단순한 방법이지만 데이터의 중복에 의해 중복된 데이터에 더 민감하게 반응하는 문제가 있다. 합성 소수 추가 표준 추출 기법(Synthetic Minority Oversampling Technique, SMOTE)[11]과 무작위 추가 표본 추출(Random Over Sampling Examples, ROSE) [12]은 소규모 범주에 속할 가능성이 있는 데이터를 생성하여 데이터의 개수를 증가시키는 방법이다. 범주의 균형도를 보장하면서 중복된 데이터도 없어서 높은 성능을 보장할 수 있지만, 데이터 양이 늘어나 모델링 시간이 오래 걸리고, 현실에 존재할

수 없는 의미 없는 데이터가 다수 생성될 가능성이 증가 한다.

본 논문에 적용하는 방법은 혼성 표본 추출 방법이다. 대규모 범주에서는 전체 분포를 고려하여 데이터들을 추출하고, 소규모 범주에서는 일부 새로운 데이터를 생성하여, 축소 표본 추출과 확대 표본 추출의 강점을 유지하면서, 단점은 최소화하는 방법이다. 혼성 표본 추출 방법의 흐름도는 <그림 2>와 같다.



<그림 2> 혼성 표본 추출 방법의 흐름도

전체 흐름에 대한 상세한 내용은 다음과 같다. 전체 데이터는 D 이고, 대규모 범주에 속한 데이터는 D_{ma}, 소규모 범주에 속한 데이터는 D_{mi}이다.

$$D = D_{ma} + D_{mi} \tag{1}$$

전체 데이터에 SVM 분류기를 적용하여 범주의 경계를 의미하는 서포트 벡터를 구한다. 서포트 벡터(SV)중 대규모 범주에 속한 서포트 벡터는 SV_{ma} , 소규모 범주에 속한 서포트 벡터는 SV_{mi}이다.

$$SV = SV_{ma} + SV_{mi} \tag{2}$$
$$SV_{ma} \in D_{ma}\,,\; SV_{mi} \in D_{mi} \tag{3}$$

식(4)와 같이 대규모 범주에 속한 데이터 중 서포트 벡터가 아닌 데이터 D_{mas}에 대해 식(5)와 같이 임의 추출을 적용하여 D_{mar}을 구한다. 그 개수는 식(6)과 같이 대규모 범주의 서포트 벡터의 개수를 넘지 않는다.

$$D_{mas} = D_{ma} - SV_{ma} \tag{4}$$
$$D_{mar} = rand(D_{mas}) \tag{5}$$
$$n(D_{mar}) < n(SV_{ma}) \tag{6}$$

소규모 범주에 속한 데이터에는 식(7)과 같이 무작위 추가 표본 추출(ROSE) 방법을 적용하여, 새로운 대규모 범주의 데이터와 새로운 소규모 범주의 데이터의 개수가 균형을 이루도록 한다.

$$D_{mir} = ROSE(D_{mi}) \tag{7}$$
$$n(D_{mir}) + n(D_{mi}) = n(D_{mar}) + n(SV_{ma}) \tag{8}$$

새로운 데이터는 대규모 범주의 서포트 벡터와 임의 표본 추출 데이터, 소규모 범주에 속한 데이터와 무작위 추가 표본 추출에 의해 생성된 데이터를 결합하여 구성한다.

$$D_n = SV_{ma} + D_{mar} + D_{mi} + D_{mir} \tag{9}$$

혼성 표본 추출 방법에 의해 생성된 데이터는 범주간에 균형을 이루고, 데이터의 개수는 원 데이터의 개수보다는 적기 때문에 모델링 시간이 적게 소요된다. 무작위 추가 표본 추출에 의해 새로 생기는 데이터의 개수도 원 소규모 범주의 데이터와 비슷하기 때문에 비현실적인 데이터가 생성될 가능성이 감소된다.

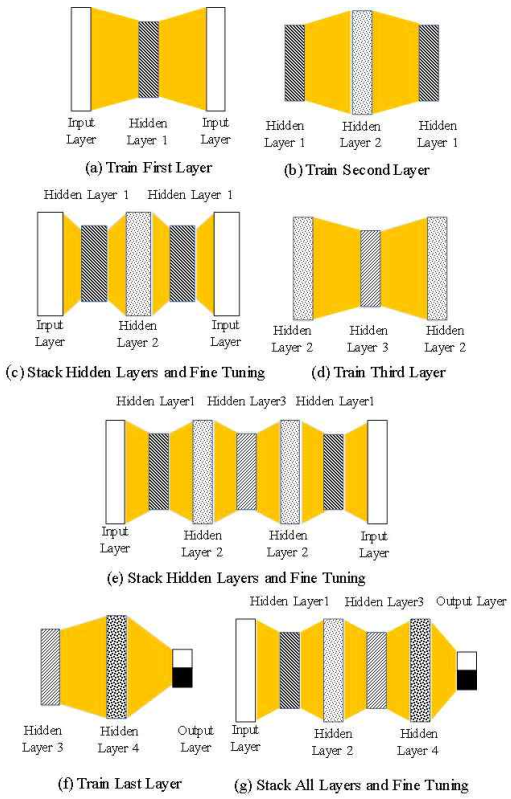
3.3 적층 딥 네트워크

신경회로망은 사람의 두뇌를 모사한 것으로 데이터를 분류하고 예측하는데 높은 성능을 보이고 있다.

하지만, 네트워크의 구조를 해석하기 불가능한 블랙박스(blackbox)이고, 모델링 시간이 오래걸리고, 성능 향상이 크지 않기 때문에 기업에서는 많이 사용되지 못했다. 하지만, GPU(Graphical Processing Unit)의 발전에 의해 모델링 시간이 단축되고, 특징 추출(Feature Extraction)과 다층 퍼셉트론을 결합한 합성곱 신경회로망(Convolutional Neural Networks, CNN)이 이미지와 비디오 데이터에 높은 성능을 보이면서, 신경회로망의 일종인 딥러닝(Deep learning)이 많이 사용되고 있다[13-14]. 음성 인식과 같은 연속형 데이터에 대해서는 순환 신경회로망(Recursive Neural Networks, RNN)이 많이 활용되고 있다.

대표적인 딥러닝 알고리즘인 합성곱 신경회로망과 순환 신경회로망은 기업 데이터에 직접 적용하기에는 데이터의 형태가 한정적이라는 문제가 있다. 그래서 다층퍼셉트론의 층(Layer)을 늘려 딥 네트워크(Deep networks)를 구성하는 방법이 사용되는데, 네트워크의 층을 늘려도 성능은 크게 증가하지 않는 문제가 발생한다. 이는 신경회로망의 학습방법으로 네트워크의 가중치를 조정할 때 학습에 영향을 미치는 인자가 네트워크가 깊어지면 영향도가 0이 되는 문제(Vanishing Gradient)에 기인한다. 이 문제는 네트워크의 학습은 출력층(Output layer)에 가까운 1 또는 2개의 층에서만 이루어져서 네트워크의 층을 증가 시켜도 학습 시간은 크게 증가하지만, 성능에는 큰 변화가 없는 것이다. 이 문제를 해결하기 위하여 적층 딥 네트워크(Stacked Deep Networks: SDN)를 구성한다. 적층 딥 네트워크는 딥 네트워크를 한 번에 학습하는 것이 아니라 한층씩 학습해서 쌓아가는 방법으로 합성곱 신경회로망의 학습과정을 모사한 것이다.

적층 딥 네트워크의 학습 방법은 <그림 3>과 같다. 적층 딥 네트워크는 은닉층(Hidden layer) 학습과 파인 튜닝(Fine Tuning)의 두 단계로 이루어져 있다. 먼저 은닉층을 학습하는 단계에서는 한 번에 하나의 은닉층을 학습하여 쌓아가게 된다. 이 때 사용하는 알



<그림 3> 적층 딥 네트워크의 학습 과정

고리즘은 오토인코더(Autoencoder)이다. 오토인코더는 입력층과 출력층이 동일한 비교사 학습(Unsupervised learning) 알고리즘으로 학습의 결과로 생긴 모델의 은닉층은 입력 데이터가 내포하고 있는 속성은 유지하면서 특징의 개수를 조절하는 특징 추출을 수행한다. 은닉층의 노드의 개수를 늘리거나 줄임으로써 특징의 개수를 늘리거나 줄일 수 있다. 또한 은닉층의 출력이 입력 데이터에 대한 은닉 특징(Hidden features)을 의미하게 된다. 은닉층의 한 노드는 입력 데이터의 모든 노드에 연결되어 있는 완전 연결의 형태이 때문에 입력 데이터에 대한 비선형 변환이 되어 새로운 은닉 특징이 추출 된다.

학습과정은 단계별로 진행된다. 은닉층이 한 개인 오토인코더를 반복적으로 구성하여 은닉 특징을 추

출한다. 오토인코더의 수식은 식(10)과 같다. 입력 데이터 X에 인코더 $E(X)$와 디코더 $D(H)$를 적용하고 학습을 통해 최적의 인코더와 디코더를 생성한다.

$$X = D_1 \cdot E_1(X) \quad (10)$$

$$E_1(X) = H_1 \quad (11)$$

은닉특징은 식(11)과 같이 표현되며, 이는 인코더의 출력을 의미한다. 두번째 은닉층은 입력 데이터로 첫번째 은닉층의 출력인 식(12)의 은닉 출력 H_1을 사용하여 오토인코더 $E_2(H_1)$를 식(13)과 같이 학습한다.

$$H_1 = D_2 \cdot E_2(H_1) \quad (12)$$

$$E_2(H_1) = H_2 \quad (13)$$

그 후 이 네트워크를 결합하여 식(14)와 같이 파인튜닝을 수행한다.

$$X = D_1 \cdot D_2 \cdot E_2 \cdot E_1(X) \quad (14)$$

이 과정을 원하는 은닉층의 개수만큼 반복하고 난 후 식(15)와 같이 퍼셉트론을 적용하여 출력 데이터에 대해 학습한다.

$$O = F(H_n) \quad (15)$$

마지막으로 전체 네트워크에 대한 파인 튜닝을 식(16)과 같이 수행하여 최종 딥러닝 모델의 학습을 수행한다.

$$O = F \cdot E_n \cdot E_{n-1} \cdot \ldots \cdot E_2 \cdot E_1(X) \quad (16)$$

Ⅳ. 실험 및 결과

4.1 실험 데이터

실험에 사용한 데이터는 UCI 기계 학습 저장소(machine learning repository)에서 제공하는 은행 마케팅 데이터집합(bank marketing data set)이다[15]. 이 데이터는 2008년 5월에서 부터 2010년 11월까지에 포르투갈의 은행에서 텔레마케팅으로 장기 예금 상품을 판매했던 결과를 정리한 데이터이다. 상품 판매는 캠페인 기간에 콜센터 상담원이 특정 고객에게 직접 전화한 것과 해당 기간에 고객이 콜센터에 전화하였을 때 상품을 제안하는 방법으로 진행되었다. 캠페인 결과는 성공과 실패의 두가지 상태로 저장된다. 데이터의 속성은 모두 20개로 통화 상태, 이자율 등 상품 정보, 연령 등 고객 기본 정보와 고객의 신용 상태 정보를 포함한다.

<표 1> 실험 데이터 속성

		# Cases	Imbalance	# Fail
ALL	Train	30,891	0.1592	4,917
	Test	10,297	0.1592	1,640
HS	Train	12,000	0.5	6,000

<표 1>과 같이 전화 통화는 모두 41,188건이다. 이 중 15.92%의 통화만 성공인 불균형한 데이터이다. 전체 데이터 중 75%는 학습 데이터로 25%는 테스트 데이터로 구분하였다. 학습은 데이터를 모두 사용한 것(ALL)과 제안하는 혼성 표본 추출 방법에 의해서 범주간의 균형을 조정한 집합(HS)을 사용하였다. 혼성 표본 추출 방법에 의해서 데이터는 12,000건으로 성공 비율은 50%인 데이터로 구성하였다.

4.2 성능 평가

텔레마케팅 대상 고객을 선정하는 이유는 전체 고객을 대상으로 마케팅을 진행하지 않고, 구매로 이어질 수 있는 구매 가망 고객을 대상으로 마케팅을 진행하여 비용을 절감하기 위해서이다. 전체 고객을 대상으로 진행했을 때에 성공할 고객은 포함시키면서 실패할 고객은 배재하여 마케팅 효율을 극대화하는

것을 목적으로 한다. 따라서 성능 평가를 할 때는 정확도보다 재현율(Recall)과 정밀도(Precision)가 더 중요한 성능 평가요소이다. 성능 평가를 위한 정확도, 재현율, 정밀도는 <표 2>의 오차 행렬(Confusion Matrix)로 확인할 수 있다.

<표 2> 오차 행렬

		Predicted condition	
	Total population	Prediction positive	Prediction negative
Actual condition	Actual positive	True Positive(TP)	False Negative (FN)
	Actual negative	False Positive (FP)	True Negative (TN)

정확도는 전체 데이터 중 참, 거짓을 맞춘 것을 의미하고 식(17)로 계산된다.

$$Accuracy = \frac{(TP + TN)}{Total\,Population} \tag{17}$$

재현율은 실제 참인 것 중 맞춘 것을 의미하며, 식(18)로 계산된다.

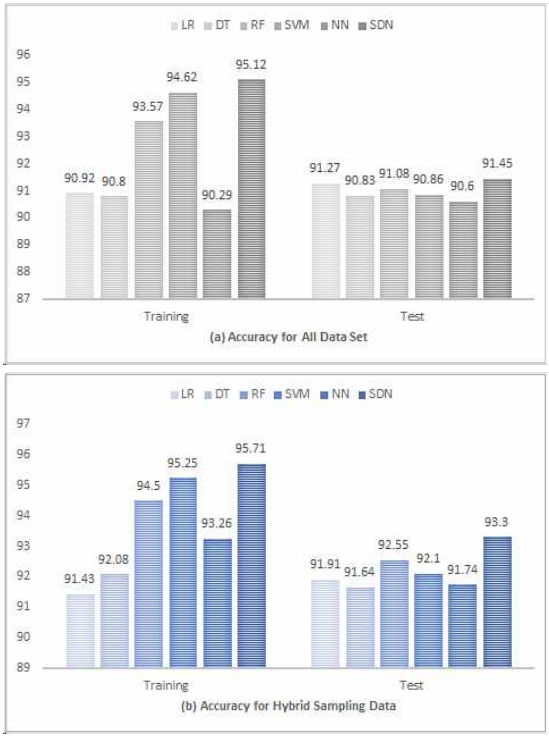
$$Recall = \frac{TP}{Actual\,Positive} \tag{18}$$

정밀도는 참이라고 예측한 것 중 실제 참인 것을 의미하며, 식(19)로 계산된다.

$$Precision = \frac{TP}{Prediction\,Positive} \tag{19}$$

4.3 실험 결과

제안하는 적층 딥 네트워크(SDN)의 성능을 비교하기 위하여 기계 학습 알고리즘인 로지스틱 회귀분석(LR)과 의사 결정 나무(DT), 랜덤 포레스트(RF), 서포트 벡터 머신(SVM), 신경회로망(NN) 의 성능을 비교하였다.



<그림 4> 정확도 비교 실험 결과

<그림 4>에서는 전체데이터 집합과 혼성 표본 추출에 의한 데이터 집합에 대하여 학습 데이터와 테스트 데이터에 대해 알고리즘들의 정확도(Accuracy)를 비교하였다.

<그림 4>의 (a) 전체 데이터에 대한 정확도를 살펴보면, 학습 데이터에 대해서는 비교 대상 알고리즘 중 신경회로망이 90.29%로 가장 낮았고, SVM이 94.62 %로 가장 높았다. 제안하는 적층 딥 네트워크는 95.12%로 가장 높은 성능을 보였다. 테스트 데이터를 살펴보면, 비교 대상 알고리즘 중 신경회로망이 가장 낮았고, 로지스틱 회귀 분석이 가장 높았으며,

제안하는 알고리즘이 91.45%로 가장 높았다. 하지만, 그 차이는 0.18 %로 데이터의 편중화에 의해서 대규모 범주에 대해서는 정확도는 향상되지만, 소규모 범주에 대해서는 정확도가 저하되어 일반화 성능은 크게 향상되지 않았다.

<그림 4>의 (b) 혼성 표본 추출 방법으로 데이터의 균형도를 개선한 경우를 살펴보면, 학습 데이터에 대해서는 신경회로망을 제외하고는 모든 알고리즘이 0.6% 정도의 성능 향상을 보였다. 테스트 데이터에 대해서는 일반화 성능이 증가하여 모든 알고리즘에서 전체 데이터를 사용한 경우보다 1%이상의 성능 향상을 보였다. 제안하는 적층 딥 네트워크는 전체 데이터를 사용한 기존 알고리즘보다 학습 데이터에 대해 1%, 테스트 데이터에 대해서는 2%의 성능 향상을 보였다.

<표 3>에서는 테스트 데이터에 대해서 각 알고리즘 별 정확도, 재현율, 정밀도를 비교하였다. 테스트 데이터에 대한 정확도는 <그림 4>의 결과와 동일하고, 재현율과 정밀도를 추가로 비교하였다. 텔레마케팅 대상 고객을 선정시에는 구매 가망 고객이 많이 포함되어야 마케팅 효과가 높기 때문에 정확도보다 재현율, 정밀도가 더 중요하다.

전체 데이터에 대하여 정확도는 90% 이상으로 높지만, 재현율과 정밀도는 40%에서 60% 대로 낮았다. 학습 데이터에 마케팅 실패가 더 많이 포함된 불균형한 데이터이기 때문에 마케팅 실패를 더 잘 예측하도록 학습이 진행되기 때문에 정확도보다 재현율은 크게 떨어질 수밖에 없다. 하지만, 제안하는 혼성 표본 추출 방법으로 데이터의 균형화를 하였을 때 재현율은 비교 대상 알고리즘에 비하여 7%에서 13%로 평균 10%, 정밀도는 0.3%에서 9%로 평균 5% 정도 향상되었다.

제안하는 방법은 정확도, 재현율, 정밀도에서 우수한 성능을 보였다. 하지만 정확도의 성능 향상도는 낮았고, 재현율과 정밀도의 성능 향상도는 높았다. 제안하는 혼성 표본 추출 방법에 의해 데이터의 편중화가 개선되면서, 상대적으로 낮았던 소규모 범주에 속한 데이터의 재현율과 정밀도가 향상되어 전체 재현율과 정밀도가 향상되었다.

Ⅴ. 결론

은행의 영업 실적을 위하여 적극적으로 고객을 유치하고자 하는 다이렉트 마케팅에 대한 수요는 지속적으로 늘어나고 있다. 은행의 입장에서는 다이렉트 마케팅의 대상 고객에 상품 구매 가망 고객을 되도록 많이 선정하고, 상품 비구매 가망 고객을 줄여서 캠페인 비용은 줄이면서, 높은 효과를 얻으려고 한다. 본 논문에서는 혼성 표본 추출과 적층 딥 네트워크를

<표 3> 알고리즘별 정확도, 재현율, 정밀도 비교 실험

Algorithms	Test Data					
	ALL			HS		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
LR	91.27 %	40.66 %	68.42 %	91.91 %	50.65 %	68.75 %
DT	90.83 %	55.08 %	59.75 %	91.64 %	62.29 %	62.67 %
RF	91.08 %	43.44 %	65.19 %	92.55 %	56.56 %	70.92 %
SVM	90.86 %	44.66 %	62.84 %	92.10 %	55.78 %	67.86 %
NN	90.60 %	46.92 %	60.20 %	91.74 %	57.08 %	64.79 %
SDN	91.45 %	58.91 %	62.49 %	93.30 %	66.72 %	71.44 %

결합하여 은행의 텔레마케팅 가망 고객을 선정하는데 있어서 정확도, 정밀도, 재현율에서 우수한 성능을 보이는 모델을 제안하였다. 캠페인 성공 고객이 적은 불균형한 데이터로 인해 정확도에 비해 재현율과 정밀도가 저하되는 문제를 해결하기 위하여 혼성 표본 추출로 범주간 데이터 개수의 균형을 맞추어 성능을 향상시키고, 적층 딥 네트워크를 통하여 구매 가망 고객 선정의 분류의 성능을 높이는 방법을 제안하였다. 제안하는 방법을 은행의 텔레마케팅 데이터에 적용하여 실험한 결과 구매 가망 고객의 선정에 있어서 정확도, 재현율, 정밀도에서 우수한 성능을 보였다.

본 연구에서 사용한 은행 마케팅 데이터집합(bank marketing data set)은 정제된 데이터이다. 적층 딥 네트워크는 특징 추출 기능이 있어서 정제되지 않은 데이터에도 적용할 수 있다. 따라서 다양한 속성을 가지고 있는 정제되지 않은 마케팅 데이터에 대해서 제안하는 모델을 적용할 수 있을 것이다.

참고문헌

[1] P. Kotler, K. L. Keller, "Framework for Marketing Management, 6th edition," Pearson, 2015.

[2] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decision Support Systems, Vol. 62, 2014, pp. 22-31.

[3] E. Turban, R. Sharda, and D. Delen, " Decision Support and Business Intelligence Systems, 9th edition," Pearson, 2010, pp. 2-35.

[4] S. L. France, and S. Ghose, "Marketing analytics: Methods, practice, implementation, and links to other fields," Expert Systems with Applications, Vol. 119, 2019, pp. 456-475.

[5] G. Marinakos, and S. Daskalaki, "Imbalanced customer classification for bank direct marketing," Journal of Marketing Analytics, Vol. 5, Issue 1, 2017, pp. 14-30.

[6] 김승수 · 김종우, "비정형 정보와 CNN 기법을 활용한 이진 분류 모델의 고객 행태 예측," 지능정보연구, Vol. 24, No. 2, 2018, pp. 221-241.

[7] F. Kaefer, C. M. Heilman, and S. D. Ramenofsky, "A neural network application to consumer classification to improve the timing of direct marketing activities," Computers & Operations Research, Vol. 32, No. 10, 2005, pp. 2595-2615.

[8] S. Liao, Y. Chen, and H. Hsieh, "Mining customer knowledge for direct selling and marketing," Expert Systems with Applications, Vol. 38, 2011, pp. 6059-6069.

[9] S. H. Javaheri, M. M. Sepehri, and B. Teimourpour, "Response modeling in direct marketing: a data mining based approach for target selection," Data Mining Applications with R, 2014, pp. 153-178.

[10] P. Ladyzinski, K. Zbikowski, and P. Gawrysiak, "Direct marketing campaign in retail banking with the use of deep learning and random forests," Expert Systems with Applications, Vol. 134, 2019, pp. 28-35.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, Vol. 16, 2002, pp. 321-357.

[12] G. Menardi, and N. Torelli, "Training and assessing classification rules with imbalanced data,"Data Mining and Knowledge Discovery, Vol. 28, No. 1, 2014, pp. 92-122.

[13] 김창식 · 김남규 · 곽기영, "머신러닝 및 딥러닝 연

구동향 분석: 토픽모델링을 중심으로," 디지털산업정보학회 논문지, 제15권, 제2호, 2019, pp.19-28.

[14] 주명길 · 윤성욱, "워드 임베딩과 CNN을 사용하여 영화 리뷰에 대한 감성 분석," 디지털산업정보학회 논문지, 제15권, 제1호, 2019, pp.87-97.

[15] S. Moro, P. Cortez, and P. Rita, "A Data-Driven Approach to Predict the Success of Bank Telemarketing," Decision Support Systems, Vol. 62, 2014, pp. 22-31.

■ 저자소개 ■

이 현 진
Lee, Hyun Jin

2003년 3월~현재 숭실사이버대학교 ICT 공학과 부교수
2002년 8월 연세대학교 컴퓨터과학과 (공학박사)
1998년 8월 연세대학교 컴퓨터과학과 (공학석사)
1996년 8월 순천향대학교 전산학과(공학사)

관심분야 : 머신러닝, 빅데이터 처리, 사이버교육
E-mail : hjlee@mail.kcu.ac

논문접수일 : 2019년 8월 2일
수 정 일 : 2019년 8월 29일
게재확정일 : 2019년 9월 3일