Journal of The Korea Institute of Information Security & Cryptology ISSN 1598-3986(Print)
VOL.28, NO.6, Dec. 2018 ISSN 2288-2715(Online)
https://doi.org/10.13089/JKIISC.2018.28.6.1499

Filter Method와 Classification 알고리즘을 이용한 전자상거래 블랙컨슈머 탐지에 대한 연구*

이 태 규,† 이 경 호‡
고려대학교 정보보호대학원

Black Consumer Detection in E-Commerce Using Filter Method and Classification Algorithms*

Taekyu Lee,† Kyung Ho Lee‡
Institute of Cyber Security & Privacy (ICSP), Korea University

요 약

빠른 속도로 성장하고 있는 전자상거래 시장이 기업들에게 고객층을 넓혀나갈 좋은 기회를 제공하고 있는 반면에 블랙컨슈머로 인한 기업들의 피해 사례 또한 늘어나고 있다. 본 연구는 전자상거래 고객 데이터를 통해 전자상거래 상의 블랙컨슈머를 탐지해내는 머신 러닝 모델을 구축하고 최적화하는 것을 목표로 한다. Feature selection의 filter method와 4개의 classification 알고리즘을 이용한 실험을 통해 F-measure 0.667의 정확도로 블랙컨슈머를 탐지하는 모델을 구축하였으며 F-measure에서 11.44%, AURC에서 10.51%, TPR에서 22.87%의 성능 향상을 확인 할 수 있었다.

ABSTRACT

Although fast-growing e-commerce markets gave a lot of companies opportunities to expand their customer bases, it is also the case that there are growing number of cases in which the so-called 'black consumers' cause much damage on many companies. In this study, we will implement and optimize a machine learning model that detects black consumers using customer data from e-commerce store. Using filter method for feature selection and 4 different algorithms for classification, we could get the best-performing machine learning model that detects black consumer with F-measure 0.667 and could also yield improvements in performance which are 11.44% in F-measure, 10.51% in AURC, and 22.87% in TPR.

Keywords: Machine Learning, Supervised Learning, Fraud Detection, User Classification, Feature Selection

I. 서 론

인터넷과 스마트폰의 보급화로 인해 전자상거래 시장은 지난 10년간 매우 빠른 속도로 성장해왔으며 기존 상거래 시장의 규모에 까지 근접해 가고 있다. 이러한 성장 과정에서 소비자 보호 관련법과 규제는 계속해서 늘어가고 있으며 소비자의 안전한 구매가 가능한 환경으로 발전되어 가고 있다. 반면에 이러한 상황을 악용하는 블랙컨슈머라 불리는 소비자들 또한 늘어나고 있다. 악성을 뜻하는 블랙과 소비자를 뜻하는 컨슈머의 합성어로 고의적, 상습적으로 기업에 피해를 주는 소비자를 뜻하는 말이다[1]. 전자상거래에서도 신용카드 차지백 악용, 환불 정책 악용 등 비정상적인 행위를 통해 기업에 큰 피해를 끼치는 블랙

Received(10. 15. 2018), Modified(11. 08. 2018), Accepted(11. 08. 2018)
* 본 연구는 방위사업청과 국방과학연구소의 지원으로 수행되었습니다 (UD060048AD)
† 주저자, xorb4123@korea.ac.kr
‡ 교신저자, kevinlee@korea.ac.kr(Corresponding author)

컨슈머 들이 늘어나고 있다. 이러한 상황에 대한 예방책으로서 블랙컨슈머 탐지 및 대응책이 필요한 상황이다.

본 연구는 이렇게 급증하고 있는 전자상거래상의 블랙컨슈머를 탐지하기 위해 고객 주문 데이터를 이용해 지도 학습 머신러닝 모델을 구축하고 filter method와 classification 알고리즘을 이용한 실험을 통해 최적화하는 것을 목표로 한다.

II. 관련연구

2.1 Anomaly 탐지

Anomaly 탐지를 위한 머신 러닝 연구는 과거부터 꾸준히 진행 되어오던 연구이며 여러 가지 케이스 스터디 들이 진행되어 신용 카드 사기, 보험 사기, 내부자 위협 탐지, 적군 감시 등의 여러 분야에서 연구 되고 있다. Anomaly 탐지란 데이터 속에서 일반적으로 기대되는 범위를 벗어난 패턴을 찾아내는 것이다. Anomaly 탐지가 중요한 이유는 많은 분야에서 데이터 속 anomaly는 매우 치명적이고 중대한 영향을 줄 수 있는 상황을 나타내기 때문이다. 예를 들면 네트워크 속 anomaly 트래픽은 해커의 침입을 의미할 수 있으며 병원에서 촬영한 MRI 이미지 속의 anomaly는 악성 종양을 의미할 수도 있다[2]. Anomaly 탐지 모델은 clustering 또는 classification을 이용한 머신 러닝을 통해 일반적인 데이터에서 기대되는 범위를 벗어난 anomaly를 탐지해내도록 구현한다.

김태호 등은 내부자 위협 탐지를 위한 비지도 학습에서의 속성선택 최적화에 대해 연구하였다. Filter method를 통해 feature selection을 했으며 EM, k-means, canopy, density based등의 clustering 알고리즘을 이용하여 실험을 진행했다. 각 실험값에서 최적화된 모델을 찾기 위해 accuracy, AUC, TPR 값을 이용하여 실험값을 검증했다[3].

RG Stafford등은 신경망 시스템을 이용해 의료용 이미지 상에서의 anomaly를 탐지해내는 방법을 구현하였다. 의료용 이미지를 같은 크기의 작은 부분으로 나누어 신경망 시스템에 학습시키고 anomaly가 탐지되는 부분을 찾아 전체 이미지에서 어떤 부분인지 확인하는 방식이다[4].

M Ahmed 등은 네트워크상의 Anomaly 탐지 기술을 크게 classification, statistical, clustering, information theory 이렇게 4분류로 나눠서 정리하였다[5].

2.2 신용 카드 사기 탐지

신용 카드 사기 탐지(credit card fraud detection)는 anomaly 탐지를 신용 카드 거래 내역 데이터에 적용한 연구 분야이다. 신용 카드 사기 탐지는 데이터 마이닝, 머신러닝, 인공 지능 등의 분야에서 여러 가지 방법을 통해 연구되어 오고 있다.

이호진 등은 신용 카드 사기 탐지를 위한 clustering 알고리즘에서의 여러 가지 feature selection method들을 이용하여 실험 한 후 결과값을 비교하여 효율성과 정확도를 높이는 연구를 진행하였다[6]. 해당 연구에서는 6개의 비지도 학습 알고리즘, 10개의 feature evaluators 그리고 11개의 search methods를 이용하여 성능을 비교하는 실험을 진행하였다.

Sam Maes등은 신용 카드 사기 탐지를 위해 인공 신경망(artificial neural networks)과 베이지언망(bayesian belief networks)을 사용하여 머신 러닝 모델을 구축한 후 ROC(the receiver operating curve)를 이용해 결과를 비교하였다[7].

A Srivastava등은 은닉 마르코프 모델(hidden markov model)을 사용하여 신용 카드 거래프로세스를 모델링하고 카드 사용자 데이터를 학습해 신용카드 사기를 탐지한 연구하였다[8].

2.3 고객 이탈 예측

머신 러닝을 이용해서 고객을 분류하는 연구는 고객 이탈 예측(customer churn prediction) 분야에서도 연구되고 있다. 고객 이탈 예측은 기존 고객이 기업의 서비스를 탈퇴 또는 구독 취소 등의 이유로 이탈하는 것을 머신 러닝 기술을 이용해 예측하는 것을 말한다. 고객의 행동을 포함한 고객 데이터를 통해 특정 고객들을 분류해 낸다는 점에서 본 연구의 블랙컨슈머 탐지 모델 연구과 유사하다고 볼 수 있다.

2008년 Xia Guo-en과 Jin Wei-dong의 연구에서는 구조적 위험 최소화(structural risk minimization)에 기반을 둔 support vector

machine 알고리즘을 사용하여 고객 이탈 예측을 위한 머신 러닝 모델을 구축한 후 artificial neural network, decision tree, logistic regression, naive bayesian classifier등의 다른 알고리즘과 성능을 비교하는 실험을 진행했다. 성능 비교에는 각 알고리즘의 accuracy rate, hit rate, covering rate, lift coefficient 값을 이용했으며 support vector machine 알고리즘이 고객 이탈 예측에서 가장 좋은 성능을 보여주었다[9].

2015년 T. Vafeiadis, K.I. Diamantaras, G. Sarigiannidis, K.Ch. Chatzisavvas의 연구에서는 고객 이탈 예측 모델에 여러 가지의 알고리즘을 적용해 비교한 후 boosting 알고리즘을 적용하여 예측 모델의 성능을 향상 시켰다. AdaBoost와 같은 boosting 알고리즘을 적용한 모델에서 accuracy는 1-4%, F-measure은 4.5-15% 향상되어 일반 모델에 비해 확실한 성능 개선을 확인 할 수 있었다[10].

2017년 Coussement, K., Lessmann, S., Verstraeten, G.의 연구에서는 고객 이탈 예측 모델 구축에서의 data preparation 기술에 대해 분석하였다. 이전 연구들이 알고리즘의 성능에 집중을 한 반면에 이 연구는 data preparation 기술이 예측 모델 성능에 어떠한 영향을 주는지 실험하였다. 실험 결과 data preparation 기술에 따라 성능에는 확연한 차이가 있었으며 AURC에서 최고 14.5%, TDL에서 최고 34%의 성능 향상을 확인할 수 있었다[11].

2.4 본 연구의 차별성

신용 카드 사기 탐지 연구들에서 사용된 데이터들은 전자상거래 플랫폼 상에서의 고객의 행동이나 주문 특성들은 포함하기 어려운 점이 있다. 본 연구에서는 고객의 행동과 주문 상의 특성을 포함한 데이터 셋을 이용하여 어떤 특성을 갖고 있는 데이터들이 블랙컨슈머로 분류되었는지 학습하여 탐지하는 지도 학습 머신러닝 모델을 구축하고자 한다. Clustering을 사용한 비지도 학습 탐지 모델은 전체의 데이터 중에 anomaly를 찾아내는 데에는 좋은 방법일 수 있지만 과거에 수집된 데이터에 대해 비슷한 데이터를 탐지해 내기 위해서는 labeled된 아웃풋이 존재하는 데이터 셋을 이용한 지도 학습 classification 알고리즘을 사용하는 것이 더 적합하다.

본 연구에서는 filter method로 측정한 feature의 중요도 랭킹에 따라 feature를 하나씩 줄여가며 각 classification 알고리즘에서의 최적의 feature set을 찾아가는 방식으로 filter method와 classification 알고리즘을 이용한 머신 러닝 모델의 최적화 방법을 블랙컨슈머 탐지 모델에 적용해볼 것 이다.

아직까지 많은 연구가 진행되지 않은 전자상거래 상의 블랙컨슈머 탐지 분야에서 본 연구는 새로운 연구 방향 제시 및 실제 적용에 대한 의미가 있으며 본 연구에서 진행한 블랙컨슈머 탐지 모델 구축 과정과 실험 결과는 실제 기업의 운영에 적용하였을 때에 실용적인 도움을 줄 것으로 기대된다. 기업에서 과거에 수집한 블랙컨슈머들의 데이터를 학습해 새로운 고객의 데이터가 입력되었을 때 기업의 고객 특성에 맞게 학습된 머신러닝 모델이 고객 타입을 분류하여 일반 고객인지 블랙컨슈머인지 예측하여 혹시 모를 블랙컨슈머로부터의 피해에 미리 대처할 수 있는 기회를 줄 수 있다. 이러한 기능은 기업이 블랙컨슈머로 인한 피해를 예방하고 줄이는 데에 많은 도움을 줄 것이다.

III. 탐지 모델 구현

본 연구에서는 WEKA(the waikato environment for knowledge analysis)[12]를 이용하여 머신러닝 모델을 구현한다. 머신러닝 모델 구현은 다음과 같은 과정을 통해 이루어진다.

Fig. 1. Machine Learning Model Process

3.1 Data Pre-Processing

실험에 사용될 데이터는 국내에서 실제로 운영되고 있는 전자상거래 기업의 고객 주문 데이터 셋이며 결제 시도 횟수 등의 고객 행동 정보를 포함하고 있다. 총 4179개의 주문 고객 데이터 중 35개의 데이터가 블랙컨슈머로 분류 되어 있으며 전체 주문 고객 데이터의 0.8375%가 블랙컨슈머이다. 블랙컨슈머 고객 데이터는 앞서 언급한 전자상거래 기업에서 과

거에 발생한 블랙컨슈머 피해 데이터를 통해 수집되었다. 지도 학습 머신 러닝에 적합하게 고객 분류는 type이라는 항목으로 일반 고객과 블랙컨슈머로 labeled되어있다.

본 연구에서 구축하고자 하는 모델의 목적은 새로운 주문 고객 데이터가 입력되었을 때 기업에 의해 분류 되어있는 블랙컨슈머와 같은 고객 type으로 예측되는 잠재적 블랙컨슈머를 탐지해 내는 것이다. 그러므로 해당 기업이 과거에 어떤 기준으로 블랙컨슈머를 분류했는가에 따라 탐지 모델의 잠재적 블랙컨슈머 예측 결과도 달라진다. 본 연구의 실험 방식은 특정 데이터 셋에 의존적이지 않기 때문에 다른 기업의 새로운 항목을 갖고 있는 데이터 셋과 그 기업에서 발생한 블랙컨슈머 데이터에 대해서도 본 연구의 실험 과정은 동일하게 적용가능하다.

Table 1은 본 연구에서 쓰일 데이터 셋의 feature 리스트이다. 중복되는 항목, 의미 없는 항목들이 포함되어 있고 feature들의 format도 데이터 전처리가 필요하다.

Table 1. List of features before pre-processing

Feature	Format
Accepts marketing	object
Currency	object
Subtotal	float
Shipping	float
Taxes	float
Total	float
Lineitem requires shipping	bool
Lineitem taxable	bool
Payment method	object
Source	object
Presentment currency	object
Payment attempts	int
Shipping and IP location match	bool
Billing country and shipping country match	bool
Web proxy	bool
Type	int

마지막 항목인 Type은 고객 타입을 나타내며 0이 일반 고객, 1이 블랙컨슈머로 분류되어 있다. 나머지 항목들은 마케팅 수신 여부, 결제 시도 횟수, 웹 프록시 사용 여부, 주문 결제 금액 등 여러 가지 고객 정보를 담고 있다.

머신러닝에 사용하기 위해서 데이터 전처리를 통해 데이터 셋의 중복 항목 또는 실험에 관련 없는 무의미한 항목을 제거하고 항목들의 포맷을 머신러닝 과정에 사용 가능한 integer로 치환했다. Table 2는 데이터 전처리 후의 데이터 셋의 feature 리스트를 정리한 표이다.

Table 2. List of features after pre-processing

Feature	Format
Accepts marketing	int
Subtotal	int
Lineitem taxable	int
Payment method	int
Source	int
Payment attempts	int
Shipping and IP location match	int
Billing country and shipping country match	int
Web proxy	int
Type	int

데이터 전처리 후 데이터 항목간의 다중공선성 존재 여부에 대해 NumXL 프로그램을 통해 측정해보았다. 독립변수들 간에 상호의존성 또는 상관계수가 일정 수준 이상 높은지 측정하는 방식이다. 계산 결과 데이터 전처리 이후 각 항목들 간에 다중공선성은 존재하지 않아 본 연구 실험에 적합한 데이터 셋임을 확인하였다.

Multicollinearity Test				
Variable	Tol.	R^2	VIF	Present?
1	96.8%	3.2%	1.03	FALSE
2	99.7%	0.3%	1.00	FALSE
3	81.6%	18.4%	1.23	FALSE
4	83.0%	17.0%	1.20	FALSE
5	71.2%	28.8%	1.40	FALSE
6	83.0%	17.0%	1.21	FALSE
7	92.7%	7.3%	1.08	FALSE
8	96.3%	3.7%	1.04	FALSE
9	60.3%	39.7%	1.66	FALSE

Fig. 2. Multicollinearity test result

3.2 Feature Selection

Feature selection은 데이터 셋의 여러 가지 항목 중 머신 러닝 모델이 학습하고 결과 값을 도출 해낼 때에 가장 큰 연관성을 갖고 있는 항목을 골라내는 과정이다. 연관성이 적은 항목을 제외시키고 연관성이 큰 항목들을 골라내 데이터 셋의 차원수를 줄여

학습 성능을 향상시키고 불필요한 데이터를 없애 과대적합(overfitting)문제를 해결할 수 있다[13].

Feature selection 방법은 filter, wrapper, embedded method 이렇게 총 3가지 방법으로 나뉜다[14].

Filter method는 각 feature의 중요도를 평가한 랭킹을 기반으로 가장 중요도가 낮은 feature 부터 하나씩 제거해 나가며 가장 좋은 feature set을 찾아내는 방법이다. 비교적으로 과대적합 문제가 발생할 가능성이 더 낮으며 더 큰 규모의 데이터 셋에도 적합한 방법이다. Wrapper method처럼 해당 러닝 알고리즘에서의 최고의 성능을 내는 feature subset을 결과로 주는 방식이 아닌 러닝 알고리즘과 독립적으로 각 feature의 중요도를 측정해내는 방법으로 러닝 알고리즘에 종속되지 않는다.

Wrapper method는 각 러닝 알고리즘에서 어떠한 feature subset이 가장 좋은 성능을 갖는지 찾아내는 방식으로 모든 subset의 성능을 평가해 가장 좋은 subset을 찾아내는 방법이다. 특정 알고리즘에서는 높은 정확도의 결과를 주지만 데이터의 feature 수가 많고 규모가 클 경우 비효율적이며 과대적합 문제가 잘 발생하는 단점이 있다.

Embedded method는 filter method와 wrapper method의 장점을 결합한 방법이다. Wrapper method와 같이 각 러닝 알고리즘에서의 최적의 feature set을 구해내지만 보다 더 효율적으로 찾아내는 방법이다.

본 연구에서는 과대적합 문제가 덜 발생하고 러닝 알고리즘으로부터 독립적인 filter method를 이용해 각 feature의 연관성을 측정한다. 연관성 랭킹이 낮은 순으로 feature 개수를 하나씩 제외하며 각 알고리즘에서의 머신 러닝 결과 값을 비교해 최적의 feature set을 찾아내도록 할 것이다.

본 연구에서는 WEKA에서 제공하는 filter method인 correlation ranking filter를 이용해 각 feature의 중요도 랭킹을 측정했으며 payment attempts, shipping and IP location match, billing country and shipping country match, web proxy, accepts marketing, subtotal, payment method, source, line item taxable 순으로 결과가 나왔다.

최적의 feature set은 지도 머신러닝 모델의 알고리즘 최적화와 함께 실험하여 도출해낼 것이다.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 10 Type):
        Correlation Ranking Filter
Ranked attributes:
 0.51468   6 Payment Attempts
 0.39652   7 Shipping and IP Location Match
 0.19361   8 Billing Country and Shipping Country Match
 0.16834   9 Web Proxy
 0.09474   1 Accepts Marketing
 0.08997   2 Subtotal
 0.05677   4 Payment Method
 0.02725   5 Source
 0.00163   3 Lineitem taxable

Selected attributes: 6,7,8,9,1,2,4,5,3 : 9
```

Fig. 3. Correlation ranking filter result

3.3 Supervised Learning

지도 학습(supervised learning)은 인풋과 아웃풋이 존재하는 데이터 셋을 통해 머신러닝 모델을 학습시킨 후 새로운 인풋에 대한 아웃풋을 예측해내는 머신러닝 방법이다[15]. 지도 학습은 보통 classification과 regression에 사용된다. 본 연구에서 구축하고자 하는 머신러닝 모델은 classification 모델이다. Classification은 모델의 아웃풋 데이터들이 분류되어 나누어질 수 있는 경우를 말하며 본 연구에서는 일반고객과 블랙컨슈머로 분류된다.

본 연구에서는 WEKA에서 제공하는 classification 알고리즘 4가지의 실험 결과를 비교하여 최적의 모델을 구축하고자 한다. 실험에 사용할 알고리즘은 random forest, J48, naive bayes, SMO이다.

Random Forest는 Leo Breiman에 의해 개발된 알고리즘으로 classification 성능을 높이기 위해 여러 개의 decision tree들을 하나로 묶은 forest를 형성하여 decision tree의 과대적합 문제를 해결해 더 정확한 결과를 얻을 수 있다[16].

J48은 Ross Quinland에 의해 개발된 C4.5 decision tree의 오픈 소스 알고리즘이다[17]. Decision tree는 나무와 같은 모양의 그래프로 표현되는 의사 결정 방식의 한 종류이다. Decision tree는 class의 종류가 미리 주어지는 지도 학습 classification에서주로 쓰이며 데이터 마이닝, 정보 추출, 패턴 인식 등의 분야에서도 매우 유용하게 쓰이고 있다[18].

Naive Bayes는 Bayes theorem을 적용한 알

고리즘으로 모든 feature들이 독립적이라고 가정을 하고 계산을 하는 방식이다[19]. 독립적이라는 가정이 제대로 된 가정이 아니라 현실적으로 정확하게 알맞지 않을 수 있지만 실제로 다른 classification 알고리즘들과 비교했을 때 좋은 성능을 보여주는 경우가 많다.

SMO(sequential minimal optimization)는 John Platt에 의해 개발된 알고리즘이다. SMO는 Support Vector Machines에서 재기된 최적화 문제를 해결하기 위해 개발되었으며 학습 과정에서 재귀적으로 문제를 나눠 해결한다[20].

3.4 Validation

Feature selection에서의 correlation ranking filter를 통해 얻은 결과 값에서 연관성 랭킹이 낮은 feature들을 하나씩 제거하며 각 feature set에 앞서 언급한 4가지의 classification 알고리즘을 적용하여 가장 최적의 feature set과 classification 알고리즘을 찾아내도록 한다. 각 실험에서 classification 알고리즘은 10-fold cross-validation을 통해 한 번의 실험에 각기 다른 10가지 경우의 test set, train set을 적용하여 더 정확한 결과 값을 도출해낸다[21].

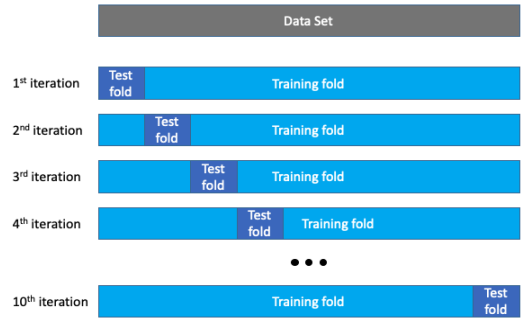

Fig. 4. 10-fold cross-validation example

결과 값을 평가하기 위해서 각 실험 결과 값을 confusion matrix로 정리하고 결과 값에 대한 F-measure, AURC(area under ROC curve), TPR을 평가 기준으로 사용하기로 한다.

Confusion Matrix는 classification 결과 값을 실제 클래스와 예측된 클래스를 기준으로 정리한 표이다. TP는 실제 클래스와 예측된 클래스 모두 positive인 경우, FP는 실제 클래스는 negative이지만 positive로 예측한 경우, FN은 실제 클래스는 positive이지만 negative로 예측한 경우, 마지막으로 TN은 실제 클래스와 예측된 클래스 모두 negative인 경우를 나타낸다[22].

Table 3. Confusion matrix

Confusion matrix		Predicted class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

True Positive Ratio (TPR)은 실제 positive 데이터 중에 positive로 알맞게 예측된 비율을 말한다.

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

False Positive Ratio (FPR)은 실제 negative 데이터 중에 positive로 잘못 예측된 비율을 말한다.

$$FPR = \frac{FP}{TN + FP} \tag{2}$$

Accuracy는 전체 데이터 중에 알맞게 예측된 결과 값의 비율을 말한다.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3}$$

AURC(area under ROC curve)은 TPR과 FPR의 관계를 나타낸 것으로 classification 알고리즘의 효율성을 나타낸다[23].

F-measure은 classification의 결과 값에서 실제 클래스와 예측된 클래스 사이의 정확도를 precision과 recall값을 이용해 계산한 조화평균(harmonic mean) 값이다[24]. Precision과 recall 값이 고려되어 계산되는 값이기 때문에 단순 accuracy를 이용한 평가보다 실제 모델이 하고자 하는 탐지를 제대로 수행하는지 더 정확한 평가가 가능하다.

Precision은 positive로 예측된 모든 값 중에 실제 positive 데이터의 비율을 나타낸다. 본 연구에서는 블랙컨슈머로 예측된 고객 중에 실제 블랙컨

슈머의 비율을 나타낸다.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

Recall은 실제 positive 데이터 중에 제대로 positive로 예측된 데이터의 비율을 나타낸다. 본 연구에서는 실제 블랙컨슈머 중에 어느 정도의 비율을 모델이 탐지했는지를 나타낸다.

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

F-measure는 precision과 recall 함께 계산하여 두 가지 정확도를 모두 고려한 정확도를 측정할 수 있다. Recall과 precision의 곱에 2배를 recall과 precision의 합으로 나눈 값이다. F-measure 값이 1에 가까울수록 더 정확한 모델이라고 평가할 수 있다.

$$F-Measure = \frac{2(Recall \times Precision)}{Recall + Precision} \quad (6)$$

3.5 Result and Optimization

본 연구에서는 실험 결과 값의 F-measure, AURC, TPR 값을 비교하여 모델의 성능을 평가할 것이다. 평가 항목의 priority(우선순위)는 탐지 능력을 가장 잘 보여주는 F-measure, 효율성을 보여주는 AURC, 블랙컨슈머 중에 실제 탐지된 비율을 보여주는 TPR 순으로 평가하여 가장 높은 성능을 보이는 최적화된 모델을 찾도록 한다.

3.5.1 실험 결과

Table 4는 본 연구에서 실험에 사용한 4가지 classification 알고리즘들에서 feature selection을 하기 전의 결과 값이다.

Table 4. Each algorithms' results before feature selection

Classification algorithm	F-measure	AURC	TPR
Random Forest	0.667	0.901	0.600
J48	0.552	0.739	0.457
Naive Bayes	0.481	0.685	0.371
SMO	0.608	0.945	0.686

Filter method를 이용하여 각 알고리즘에서 feature selection을 진행한 실험의 결과 값들을 Table 5에 하나의 표로 정리하였다.

Table 5. Test results with filter method

# of features	Classification algorithm	F-measure	AURC	TPR
9	Random Forest	0.667	0.901	0.600
	J48	0.552	0.739	0.457
	Naive Bayes	0.481	0.685	0.371
	SMO	0.608	0.945	0.686
8	Random Forest	0.645	0.912	0.571
	J48	0.552	0.739	0.457
	Naive Bayes	0.608	0.948	0.686
	SMO	0.509	0.699	0.400
7	Random Forest	0.667	0.928	0.600
	J48	0.586	0.739	0.486
	Naive Bayes	0.623	0.954	0.686
	SMO	0.500	0.685	0.371
6	Random Forest	0.621	0.796	0.514
	J48	0.621	0.796	0.514
	Naive Bayes	0.650	0.946	0.743
	SMO	0.528	0.700	0.400
5	Random Forest	0.544	0.905	0.514
	J48	0.645	0.795	0.571
	Naive Bayes	0.535	0.908	0.543
	SMO	0.500	0.685	0.371
4	Random Forest	0.468	0.862	0.314
	J48	0.409	0.795	0.257
	Naive Bayes	0.543	0.873	0.543
	SMO	0.108	0.529	0.057
3	Random Forest	0.431	0.863	0.314
	J48	0.409	0.795	0.257
	Naive Bayes	0.543	0.873	0.543
	SMO	0.108	0.529	0.057
2	Random Forest	0.444	0.863	0.343
	J48	0.409	0.795	0.257
	Naive Bayes	0.543	0.876	0.543
	SMO	0.054	0.514	0.029

4개의 classification 알고리즘에서 최고의 성능을 내는 결과 값을 얻기 위해 feature 개수를 correlation ranking filter의 결과 상 랭킹이 낮은 순으로 총 9개에서 하나씩 줄여가며 각 알고리즘에서 어

떠한 성능을 보여주는지 실험해 보았다. 각 알고리즘에서 가장 좋은 결과를 낸 feature 개수는 다음과 같다.

Random forest 알고리즘에서는 9개와 7개의 feature로 실험했을 때에 동일하게 F-measure 0.667이 나왔다. 이 둘의 성능을 비교하기 위해 AURC를 추가로 참고하였고 7개의 feature로 진행한 실험에서 가장 좋은 결과가 나왔다.

J48 알고리즘에서는 F-measure 0.645가 나온 5개의 feature로 진행한 실험에서 가장 좋은 결과가 나왔다.

Naive Bayes 알고리즘에서는 F-measure 0.650이 나온 6개의 feature로 진행한 실험에서 가장 좋은 결과가 나왔다.

SMO 알고리즘에서는 F-measure 0.608이 나온 9개 feature 모두를 사용했을 때에 가장 좋은 결과가 나왔다.

결과적으로 연관성이 가장 적은 2개의 feature를 제외시킨 총 7개의 feature로 random forest 알고리즘을 이용하여 학습한 머신러닝 모델이 가장 좋은 성능을 보여주었다. 아래의 Table 6는 해당 모델의 결과 값을 confusion matrix로 표현한 표이다. 실제 35개의 블랙컨슈머 고객 데이터 중 21개의 데이터를 제대로 탐지해 냈으며 4144개의 일반 고객 중에서는 단 7개의 데이터만을 블랙컨슈머로 오탐하여 실제 기업 운영에서 블랙컨슈머로 부터의 피해 예방에 실용적인 도움이 될 수 있는 의미 있는 수준의 TPR과 FPR을 갖는 결과 값이 나왔다.

Table 6. Detection model's confusion matrix

Confusion matrix		Predicted type	
		Black(1)	Normal(0)
Actual type	Black(1)	21	14
	Normal(0)	7	4137

Table 7. Performance comparison

Results	F-measure	AURC	TPR
Average of results before feature selection	0.577	0.818	0.529
Average of the best performing models of each algorithms	0.643	0.904	0.650

위의 Table 7은 feature selection 전후의 결과 값들의 평균값을 정리한 표이다. Filter method를 통해 feature selection을 진행하여 각 classification 알고리즘 별 가장 좋은 feature set을 찾은 탐지 모델의 결과 값들의 평균값이 feature selection을 진행하기 전의 각 알고리즘의 탐지 모델들의 결과 값들의 평균값에 비해 F-measure에서 11.44%, AURC에서 10.51%, TPR에서 22.87%의 성능 향상을 보여주었다.

IV. 결 론

본 연구에서는 지도 학습 머신러닝 알고리즘을 이용해 고객 타입을 분류하여 블랙컨슈머를 탐지해내는 머신러닝 모델을 구축하였다. Filter method를 이용하여 주요한 feature들을 추려낼 수 있었고 4가지의 classification 알고리즘의 성능을 비교하여 더 효과적이고 정확한 모델을 구축할 수 있었다. 그렇게 진행한 실험을 통해 F-measure 0.667의 정확도로 블랙컨슈머를 탐지하는 모델을 구축하였으며 F-measure에서 11.44%, AURC에서 10.51%, TPR에서 22.87%의 성능 향상을 확인 할 수 있었다.

이번 연구에서 사용된 탐지 모델 구축 과정은 다른 feature로 구성된 새로운 고객 데이터 셋에서도 탐지 모델 구축을 위해 쓰일 수 있다. 해당 데이터 셋에서도 feature selection과 classification 알고리즘들의 성능을 같은 과정의 실험을 통해 최적화해 각 기업의 고객 데이터 셋과 블랙컨슈머 분류 기준에 맞는 블랙컨슈머 탐지 모델을 구축할 수 있을 것이다.

향후 블랙컨슈머 탐지 모델의 성능을 향상시키기 위해 더 큰 규모의 데이터 셋, 새로운 머신러닝 알고리즘과 feature selection method 등을 이용하여 모델 개선 방안에 대해 연구하여 적용할 예정이다.

References

[1] Jae Wook Shin, Min Cheol Shin, "The Effects of Consumers` Psychological Characteristics on Dysfunctional Consumer Behavior and Life Satisfaction", The Korean Journal of Consumer and Advertising Psychology,

15(3), pp. 409-433, Aug. 2014
〔2〕 Chandola, V., Banerjee, A. and Kumar, V., "Anomaly detection: A survey", ACM computing surveys (CSUR), vol. 41, no. 3, p.15, Jul. 2009
〔3〕 Tae-ho Kim and Kyung-ho Lee, "Feature Selection Optimization in Unsupervised Learning for Insider Threat Detection", KSII The 13th Asia Pacific International Conference on Information Science and Technology (APIC-IST), June 2018
〔4〕 Stafford, Richard G., et al., "Application of neural networks as an aid in medical diagnosis and general anomaly detection", U.S. Patent No 5,331,550, 1994
〔5〕 Ahmed, M., Mahmood, A. N., & Hu, J., "A survey of network anomaly detection techniques. Journal of Network and Computer Applications", vol. 60, pp. 19-31., Jan. 2016
〔6〕 Lee, Hojin, et al., "Feature Selection Practice For Unsupervised Learning of Credit Card Fraud Detection", Journal of Theoretical & Applied Information Technology, vol. 96, no. 2, pp. 408-417, Jan, 2018
〔7〕 Maes, S., Tuyls, K., Vanschoenwinkel, B. and Manderick, B., "Credit card fraud detection using Bayesian and neural networks", In Proceedings of the 1st international naiso congress on neuro fuzzy technologies, pp. 261-270, Jan. 2002
〔8〕 Srivastava, A., Kundu, A., Sural, S. and Majumdar, A., "Credit card fraud detection using hidden Markov model", IEEE Transactions on dependable and secure computing, vol. 5, no. 1, pp.37-48, Jan. 2008
〔9〕 Guo-en Xia, Wei-dong Jin, "Model of customer churn prediction on support vector machine", Systems Engineering-Theory & Practice, vol.28, no.1, pp. 71-77, Sep. 2008
〔10〕 Vafeiadis, T., Diamantaras, K. I., Sarigiannidis, G., Chatzisavvas, K. C., "A comparison of machine learning techniques for customer churn prediction", Simulation Modelling Practice and Theory, vol. 55, pp. 1-9, Jun. 2015
〔11〕 Coussement, K., Lessmann, S., Verstraeten, G., "A comparative analysis of data preparation algorithms for customer churn prediction: A case study in the telecommunication industry", Decision Support Systems, vol. 95, pp. 27-36, Mar. 2017
〔12〕 Garner, Stephen R., "Weka: The waikato environment for knowledge analysis", In Proceedings of the New Zealand computer science research students conference, pp. 57-64. May 1995
〔13〕 Vipin K., et al., "Feature selection : a literature review", SmartComputing Review, vol. 4, no. 3, Jun. 2014
〔14〕 Guyon, Isabelle, and André Elisseeff., "An introduction to feature extraction", Feature extraction, Studies in Fuzziness and Soft Computing, vol, 207, pp. 1-25, 2006
〔15〕 Ghahramani, Zoubin, and Michael I. Jordan., "Supervised learning from incomplete data via an EM approach", In Advances in neural information processing systems, pp. 120-127, 1994
〔16〕 Breiman, L., "Random forests", Machine learning, vol.45, no.1, pp.5-32, Oct. 2001
〔17〕 Bhargava, N., Sharma, G., Bhargava, R., & Mathuria, M., "Decision tree analysis on j48 algorithm for data mining", Proceedings of International Journal of Advanced Research in Computer Science and Software

Engineering, vol. 3, no. 6, Jun 2013
〔18〕 Patil, T.R. and Sherekar, S.S., "Performance analysis of Naive Bayes and J48 classification algorithm for data classification", International journal of computer science and applications, vol. 6, no. 2, pp.256-261, Apr. 2013
〔19〕 Dimitoglou, G., Adams, J.A. and Jim, C.M., "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability", arXiv preprint arXiv: 1206.1121, Jun. 2012
〔20〕 Platt, J.C., "12 fast training of support vector machines using sequential minimal optimization", Advances in kernel methods, pp.185-208, Aug. 1999
〔21〕 Kohavi, R., "A study of cross-validation and bootstrap for accuracy estimation and model selection", the International Joint Conference on Articial Intelligence (Ijcai), vol. 14, no. 2, pp. 1137-1145, Aug. 1995
〔22〕 Davis, J. and Goadrich, M., "The relationship between Precision-Recall and ROC curves", In Proceedings of the 23rd international conference on Machine learning, pp. 233-240, Jun. 2006
〔23〕 Hanley, J.A. and McNeil, B.J., "The meaning and use of the area under a receiver operating characteristic (ROC) curve", Radiology, vol. 143, no. 1, pp.29-36, Apr. 1982
〔24〕 Sasaki, Y., "The truth of the F-measure", Teach Tutor mater, vol.1, no.5, pp.1-5, Oct. 2007

〈저자소개〉

이 태 규 (Taekyu Lee) 정회원
2015년 8월: 고려대학교 컴퓨터공학과 졸업
2015년 9월~현재: 고려대학교 정보보호학과 석사과정
〈관심분야〉 정보보호, 위험관리

이 경 호 (Kyung Ho Lee) 종신회원
1989년 8월: 서강대학교 수학과 학사
1997년 8월: 서강대학교 정보통신대학원 석사
2009년 8월: 고려대학교 정보보호대학원 박사
1994년 2월~현재: 삼성그룹, nhn, 시큐베이스 등 근무
2011년 9월~현재: 고려대학교 정보보호대학원 조교수, 부교수
〈관심분야〉 위험관리, 정보보호컨설팅, 정보보호 및 개인정보보호정책