Data Mining

2017 Term Project

담당교수 : 백준걸 교수님

산업경영공학부

2013170833

권 기 준

<목차>

1 Introduction

1.1 프로젝트 동기

최근 피싱 웹사이트로 인한 피해가 적지 않은 가운데, 미지의 웹사이트에 대하여 해당 웹사이트가 피싱사이트인지, 아니면 안전한 웹사이트인지 구분하는 방법을 찾고 싶었다. 그래서 각 사이트의 정보 수집 단계부터 피싱 여부 판단까지의 데이터 마이닝 작업을 수행하는 계획을 세우게 되었다.

1.2 프로젝트 목표

웹사이트들이 어떠한 Attribute 를 갖고 있으며, 해당 attribute 들이 피싱사이트 여부와 어떠한 관계가 있는지 알아볼 것이다. 그리고 그 결과를 이용하여 새로이 수집된 웹사이트(New Instance)가 피싱의 위험이 있는지 판단 또는 예측하는 단계를 거치게 될 것이다. 또한 이번 프로젝트를 수행함으로써, 앞서 제시한 문제를 겪고 있는 웹 사용자들에게 긍정적인 영향을 줄 수 있기를 바란다.

1.3 Data 수집

- Phishtank 데이터 아카이브 www.phishtank.com 에서 수집한 2456 개의 웹사이트 데이터 (http://archive.ics.uci.edu/ml/datasets/Website+Phishing)
- Neda Abdelhamid 의 Phishing detection based Associative Classification data mining 의 연구 진행과정에서 수집된 데이터를 인용하였다.
- 변수 설명

SFH(Server Form Handler)	사용자가 정보를 제출했을 때, 웹사이트가 정보를 보내는 장소(서버)
PopUp window	사용자에게 자격 증명의 제출을 요구하는지 여부
Having sub domain	URL 내 Domain part 의 dot 개수

Request URL	페이지 내의 개체가 웹 페이지와 다른 서버에서 호출 되는 비율
URL of anchor	웹 페이지 내에서, 다른 웹페이지로 통하는 링크 비율
Web traffic	해당 웹사이트의 트래픽 순위(Alexa company 의 Alexa database 내 한정)
URL length	URL 의 길이
Age of domain	도메인 생성 후 경과시간
Having IP address	URL 도메인 이름에 IP 주소가 포함되어 있는지 여부
Result	Phishing 여부

2 Preprocessing

2.1 Variable Handling

총 30 개의 속성 중 변수들을 9 개의 속성에 대하여 모두 {1,0,-1} 또는 {1,-1} 로 이산화 한다. (나머지 21 개의 속성은 유의미한 결과를 도출하지 못할 것이라 판단하여 제거하였다.)

- SFH

  : -1 (SFH 가 'about : blank' or empty)

  : 0 (Different domain)

  : 1 (Otherwise)

- Popup Window

  : -1 (Pop-up window contains text fields)

  : 1 (Otherwise)

- Having sub domain

  : 1 (Dots in domain part = 1)

  : 0 (Dots in domain part = 2)

  : -1 (Otherwise)

- Request URL

  : 1 (Request URL 비율 < 22%)

  : 0 (22% ~ 61%)

  : -1 (Otherwise)

- URL of anchor

  : 1 (URL of Anchor 비율 <31%)

  : 0 (31% ~ 67%)

  : -1 (Otherwise)

- Web traffic

  : 1 (Rank < 100,000)

  : 0 (Rank > 100,000)

  : -1 (Alexa 데이터베이스에 없을 경우)

- URL length

  : 1 (Length < 54)

  : 0 (54 ~ 75)

  : -1 (Otherwise)

- Age of domain

  : 1 (Age > 6 months)

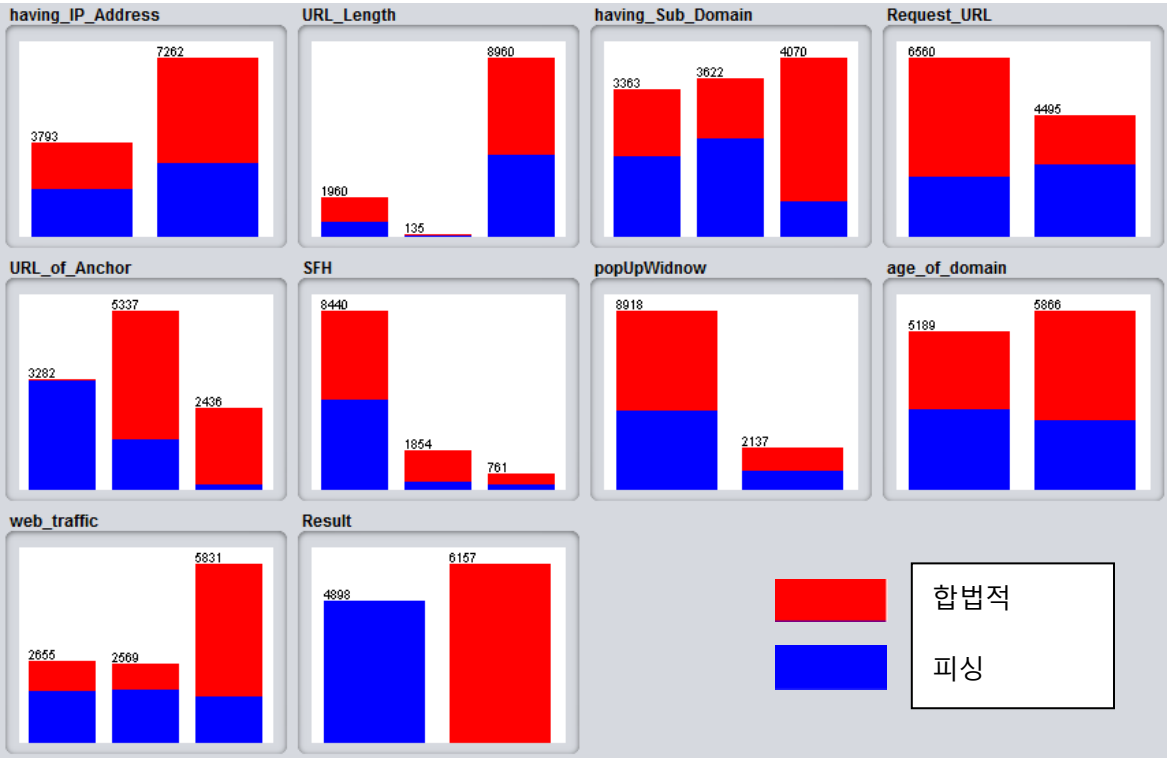
  : -1 (Otherwise)

- Having IP address

  : 1 (IP Address 포함)

  : -1 (Otherwise)

2.2 Basic Analysis



가장 현저하게 구분되는 것은 URL of Anchor와 web traffic 부분이다. Anchor 비율이 67%가 초과되었을 때 대부분의 변수가 피싱사이트로 분리되었으며, Alexa database 내의 web_traffic 순위에서 100,000위 이내를 기록한 것 중 약 4분의 3은 합법적인 사이트로 구분되었다. (가로축은 왼쪽부터 -1, 0, 1 또는 -1, 1을 나타낸다.)

3 Classification

3.1 Naïve Bayes Classifier

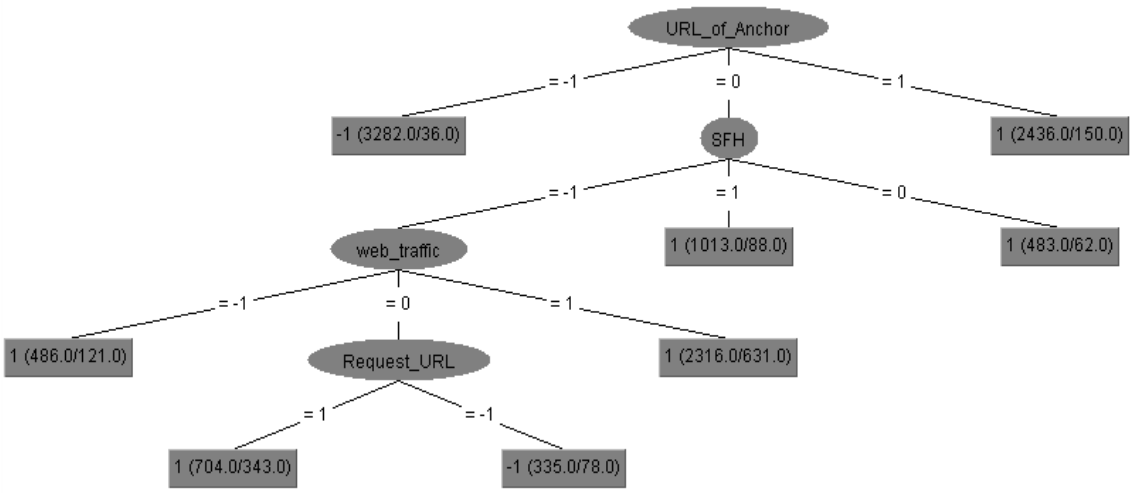
```
Correctly Classified Instances        9667               87.4446 %
Incorrectly Classified Instances      1388               12.5554 %
Kappa statistic                          0.7421
Mean absolute error                      0.1725
Root mean squared error                  0.3036
Relative absolute error                 34.9555 %
Root relative squared error             61.1212 %
Total Number of Instances           11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.793    0.061    0.912      0.793    0.848      0.747    0.939     0.941     -1
                 0.939    0.207    0.851      0.939    0.893      0.747    0.939     0.943     1
Weighted Avg.    0.874    0.142    0.878      0.874    0.873      0.747    0.939     0.942
```

Attribute 들 간의 독립성 정도를 알 수 없기 때문에, 독립이라고 가정할 수 있는 Naïve Bayes rule 을 사용하였으며, 이 알고리즘을 사용하였을 때, 87.4%의 정확도를 얻을 수 있었다.

3.2 Decision Tree



```
Correctly Classified Instances        9530               86.2053 %
Incorrectly Classified Instances      1525               13.7947 %
Kappa statistic                          0.7132
Mean absolute error                      0.2012
Root mean squared error                  0.3177
Relative absolute error                 40.7781 %
Root relative squared error             63.9617 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.720    0.025    0.958      0.720    0.822      0.733    0.911     0.912     -1
               0.975    0.280    0.814      0.975    0.887      0.733    0.911     0.898     1
Weighted Avg.  0.862    0.167    0.878      0.862    0.858      0.733    0.911     0.904
```

위의 decision tree 는 앞서 preprocessing 과정에서 시각적으로 두드러지는 영향을 보였던 4 개의 속성에 대해서만 의사결정 트리 분석을 실시한 것이다.

첫 번째로 나누어진 기준은 URL of Anchor(anchor 의 비율)였다. 처음 preprocessing 과정에서 예상할 수 있었듯이, anchor 의 비율이 가장 중요한 속성으로 나타났고, 그 다음은 SFH 속성이 두번째로 큰 비중을 차지했다. 다음은 web traffic, request url 의 순서로 구분되었다.

9 개 속성 전부를 대상으로 decision tree 를 분석했을 때는 분류의 정확도가 89.1%로 도출되었는데, 앞서 실시한 4 개 속성을 대상으로 한 분석에서 나왔던 86.2%와 비교한다면, 그렇게 큰 정확도를 가지고 있다고 보기 어렵다. (9 개 속성 중 나머지 5 개의 속성은 Instance 를 분류하는데 큰 기여를 하지 못한다고 판단할 수 있다.)

결론적으로, 피싱 사이트를 구별함에 있어서 anchor 의 비율, SFH, traffic 순위, request 비율의 순서대로 중요한 속성임을 알 수 있다.

3.3 RIPPER (JRip)

```
Correctly Classified Instances        9804               88.6839 %
Incorrectly Classified Instances      1251               11.3161 %
Kappa statistic                          0.767
Mean absolute error                      0.1891
Root mean squared error                  0.3083
Relative absolute error                 38.3068 %
Root relative squared error             62.0717 %
Total Number of Instances            11055

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.795    0.040    0.940      0.795    0.862      0.775    0.885     0.900     -1
                 0.960    0.205    0.855      0.960    0.904      0.775    0.885     0.848     1
Weighted Avg.    0.887    0.132    0.893      0.887    0.885      0.775    0.885     0.871
```

다음의 RIPPER 알고리즘을 이용한 분석에서는 분류의 정확도가 88.7%로, Bayes 알고리즘 보다 약간 상향된 수치가 측정되었다. RIPPER 는 Decision tree model 과 비슷한 구조를 갖는데, 실험 결과를 본다면 Decision tree 가 성능이 더 우수하다고 볼 수 있으며, 의사결정을 내리는데 상대적으로 비중이 적을 것이다.

4 Evaluation

4.1 k-fold cross-validation

앞선 분석에서는 모두 test-mode 를 10-fold cross-validation 으로 설정하고 분석을 실시했다. 그 결과, RIPPER 알고리즘이 88.7%, Naïve Bayes 가 87.4%, Decision Tree 가 89.1%의 분류 정확도를 보였다.

또한 1000-fold validation 으로 분석하였을 때, 각각 88.6%, 87.6%, 89.2%의 정확도를 보였는데, Naïve Bayes 와 Decision Tree 두 알고리즘은 정확도가 향상된 반면, RIPPER 알고리즘은 정확도가 오히려 낮아짐을 발견할 수 있었다.

마지막으로, k=N 으로 설정하는 Leave-one-out approach 를 사용하였을 때(Instance 수:11055)는 각각 88.6% 87.6%, 89.2% 분류 정확도를 보였다. 일정 수준 이상의 k 를 설정하면 Leave-one-out approach 와 차이가 거의 없다는 것을 확인할 수 있었다.

4.2 Model Comparison

Naïve Bayes, Decision tree 와 RIPPER 는 각각 Precision 에서 0.879, 0.894, 0.891 의 수치를 보였으며, TP-Rate(Recall) 에서는 각각 0.876, 0.892, 0.886 로 측정되었다. F-Measure 값을 살펴보면, 0.874, 0.891, 0.884 의 값을 기록하였다.

위의 결과 값들과 앞선 정확도 수치를 종합해보면, Decision Tree 의 분류 성능이 가장 우수한 것을 확인할 수 있다.

5 결론 및 고찰

앞선 분석을 통해 내가 수집한 data set 에서는 세가지 알고리즘 중 Decision Tree 알고리즘을 적용하는 것이 가장 좋은 방법이라는 결론을 내릴 수 있었다. 또한 RIPPER 알고리즘이 Naïve Bayes Rule 보다 성능이 우수하고, 통계 기반보다는 규칙 기반 알고리즘의 성능이 더 좋다고 판단할 수 있다.

Naïve Bayes Rule 은 모든 속성들이 단순히 서로 독립적이고 중요하다고 간주하기 때문에, 모두 의사 결정에 기여하도록 하는 알고리즘이다. 그래서 만약 분류에 사용된 속성들이 상호 관련되어 있다면, 분류기의 성능이 저하될 것이므로 신뢰하기 어려운 점이 있다.

Decision Tree model 에서 보았듯이, 피싱 사이트를 구별하기 위해서는 anchor 의 비율과, SFH, Traffic 순위, Request 의 비율이 중요한 것으로 보여졌다. 하지만 Preprocessing 의 이산화 과정에서 anchor 의 경우에는 impurity 가 낮도록 이산화 구간이 잘 산정되었는데, Request URL 이나 URL Length 등의 변수들은 구간 분할이 잘 되지 못했기 때문에 분류하는데 있어 비중이 상대적으로 낮았을 것이라 여겨진다. 또한 이런 상황이라면 각 속성들의 비중을 잘 반영 할 수 있는 ANN 알고리즘을 사용하는 것이 도움이 되었을 것이다. Preprocessing 에서 이런 부분이 조금 아쉬웠던 것 같고, 좀 더 나은 분석을 할 수 있는 여지가 있음을 확인했다.

마지막으로 Decision Tree 분석 시에, 대략적인 나의 주관을 바탕으로 4 개 속성을 선정했기 때문에 분석에 가장 영향력이 있는 속성들이라고 단정 짓기는 어렵다. 본문에서는 9 개 모든 속성에 대해 Decision Tree 를 나타내는데 한계가 있어 분석을 간소화하였지만, 가능하다면 모든 속성을 대상으로 분석을 실시한

후 속성들의 비중을 매기는 것이 분석을 향상시키는 또 하나의 방법이 될 것이다.

많은 부분이 미흡한 분석이었지만, 가장 우수한 모델을 찾을 수 있었고, Decision Tree 를 중심으로 어떠한 변수가 의미가 있는지 개략적으로 알아볼 수 있었다. 그 결과, URL of anchor 의 비율이 피싱사이트를 예측하는 데 가장 유의한 속성이라는 것을 알 수 있었다. 이러한 모델을 통해 여러 웹 사이트 사용자들이 피싱사이트로 인한 피해를 방지하는 데에 간접적으로 도움이 될 수 있는 지표가 되기를 기대한다.