AD OR NON-AD PREDICTION

NIKHIL SARAOGI
Master of Computer Application [MCA]
VIT UNIVERSITY CHENNAI, INDIA
(nikhilsarawgi9616@gmail.com)

Under guidance of
Dr. SATHIS KUMAR B.
School of Computer Science and Engineering
VIT UNIVERSITY CHENNAI, INDIA
(sathiskumar.b@vit.ac.in)

Abstract – The dataset information comes from the paper Learning to remove Internet advertisements. Number **of possible advertisements on Internet pages is showed by the dataset. The dataset store the image's geometry as well as phrases from the URL, the image's URL and alt text, the anchor text, and words that appear near the anchor text. In dataset, class labels (Binary Classification): "ad" and "nonad". The exciting part about this data is that as part of some pre-processing operation, someone could want to filter the WebPages for unnecessary ads (useful for later website categorisation, for example.). Data mining is used to detect whether an image is an advertisement ("ad") ("nonad"). Because I had not provided any exact training/test split, I opted to adopt an acceptable way of measuring performance. After completing feature reduction to reduce the number of features, I used several data mining methods for prediction.**

Index Terms – Different Data Mining Algorithms, Chi-Square Method, Evaluating measures.

I. Introduction

Because of the widespread usage of the internet, digital marketing has gotten a lot of attention recently. Every day, billions of people visit websites like Facebook, Google, and Instagram. This issue has prompted many businesses to turn to the Internet for marketing; they place adverts on websites in the same way that advertisements are placed in magazines.

Companies typically keep their marketing budgets secret to avoid rivalry; yet, rival companies are eager in learning about their competitors' marketing budgets in order to better their own marketing tactics. As a result, they estimate their competitors' marketing expenses by manually identifying their adverts; however, because the number of advertisements has grown dramatically, this method requires time and effort. Furthermore, processing the photos manually is nearly impossible. To detect adverts in a fair amount of time, an efficient computerised system is required.

Many websites make money via third-party advertising, which are frequently shown as images on the site's pages. If users find these so-called "banner advertisements" attractive or relevant, they can click on them to proceed to the advertiser's own site. Some people might rather not see advertisements like these. Users connecting over sluggish links will notice that advertisements considerably slow down their browsing because graphics take up the bulk of a page's entire download time. Others resent paying for services indirectly through ads and want to pay for them directly. Finally, some people are opposed to the concept of advertising on the open Internet.

In this paper, I'll use data mining methods such as Random Forest, Decision Tree, Boosting Algorithm, Logistic Regression, and others to determine whether or not a picture is an advertising ("ad") or not ("nonad"). I performed feature reduction to considerably reduce the amount of features because I am not conducting any explicit training/test split.

DATA SET –

- **Attributes**: There are 1558 characteristics in the data (3 continuous; others binary;) In 28% of the cases, one or more of the three continuous traits is absent. It contains 19 caption features, 111 alt features, 495 base URL features, 472 destination URL features, and 457 in mage

URL features, as well as height, width, and aspect ratio.

- **Number of records**: The data has 3279 instances (2821 non-ads, 458 ads).

The number of instances for each class is out of proportion. In addition, the number of characteristics is quite large in proportion to the dataset's size, suggesting that effective feature reduction is critical. In 28% of the data, one or more of the three continuous characteristics is missing.

The remainder of the paper is organized as follows. Methodology is presented in section II with complete details. In section III, discuss result with comparison respectively. Conclusions, future work and acknowledgment are provided Section IV, V, and VI.

II. Methodology

In this study, Convolution neural networks were used to recognize ads in scanned photos.. For Prediction, the data set is binary so we used binary classification algorithms with feature extraction like Chi-square method. Below Figure is showing the steps involved in the implementation of the task.

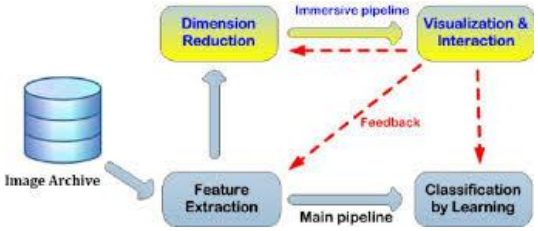


Figure 1- Data Mining Task

Data Exploration -

- First three columns are important as most of the variation between results are because of first three continuous variables
- column1-> height of image
- column2->width of image
- column30->aspect ratio

```
dataset.info () # data set information
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3279 entries, 0 to 3278
Columns: 1559 entries, height to ad-nonad
dtypes: int64(1554), object(5)
memory usage: 39.0+ MB
```

Information of First three continuous variables

- 28% data is missing for three continuous attributes.

```
column[0] has missing values -903
column[1] has missing values -901
column[2] has missing values -910
column[3] has missing values -15
```

Exploratory Data Analysis (E.D.A)

Statistical approach

- data is right skewed

	0	1	2	3
count	32 are 79.000000	3279.000000	3279.000000	3279.000000
mean	64.021886	155.344828	3.911953	0.768222
std	46.703695	110.729038	5.136153	0.422032
min	1.000000	1.000000	0.001500	0.000000
25%	32.500000	90.000000	1.279400	1.000000
50%	64.021886	150.000000	3.911953	1.000000
75%	64.021886	155.344828	3.911953	1.000000
max	640.000000	640.000000	60.000000	1.000000

```
dataset[[0,1,2,3]].describe()
```

Feature Reduction

Feature selection, also known as attribute selection, is the process of selecting the most essential properties in a dataset and subsequently improving

the model's performance using machine learning algorithms. A large number of irrelevant features exponentially increases training time and raises the risk of over-fitting.

Chi-Square Distribution

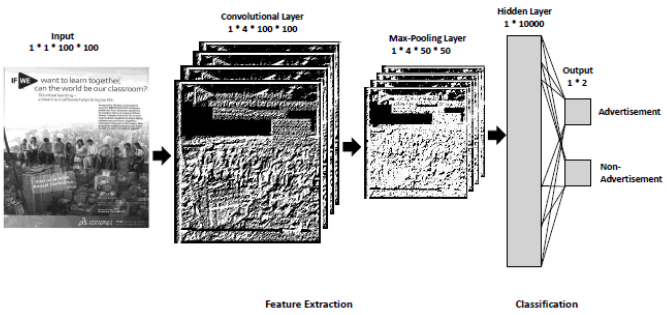
The Chi-square test is used to find categorical characteristics in a dataset. The Chi-square between each feature and the goal is computed, and the features with the highest Chi-square scores are selected. It determines if the sample's relationship between two categorical variables is indicative of the population's true relationship.

The chi-square distribution is a degree-of-freedom continuous distribution. It's a term for the distribution of a set of squared random variables. It's also used to assess if data series are independent, as well as to estimate confidence intervals for a random variable from a normal distribution's variance and standard deviation. In addition, the chi-square distribution is a gamma distribution version.
Chi- square score test is given by:

$$\chi^2 = \frac{(Observed\ frequency - Expected\ frequency)^2}{Expected\ frequency}$$

where –
Observed frequency = Number of class observation.
Expected frequency = If there was no link between the feature and the target, the number of predicted class observations would be zero.

Chi2 score function. 30 selected attributes and their fit scores
[{'width': 24645.94459079523},
{'ancurl*com': 1027.5898024704795},
{'url*ads': 850.0980613559348},
{'ancurl*click': 814.7100844130008},
{'alt*click': 624.5390113319156},
{'ancurl*redirect': 616.1754708351916},
{'ancurl*ng': 604.5578788480207},
{'alt*click+here': 598.4761298369539},
{'alt*here': 576.6482672053518},
{'aratio': 543.7864596628834},
{'ancurl*adid': 524.7351933744342},
{'ancurl*adid': 524.7351933744342},
{'ancurl*event': 469.4820957609907},
{'ancurl*ng+type': 463.3434324155194},
{'ancurl*ng+type': 463.3434324155194},
{'ancurl*ng+type': 463.3434324155194},
{'ancurl*ng+type': 463.3434324155194},
{'ancurl*ng+type': 463.3434324155194},
{'ancurl*2f': 457.2048988802811},
{'ancurl*redirect+http': 451.0665001479771},
{'ancurl*redirect+http': 451.0665001479771},
{'ancurl*http+www': 440.4651919441344},
{'alt*for': 359.3938429981922},
{'ancurl*2f+2fwww': 309.9374465556694},
{'ancurl*2f+2fwww': 309.9374465556694},
{'alt*here+for': 295.3774167026522},
{'ancurl*familyid': 294.9019607843137},
{'ancurl*familyid': 294.9019607843137},
{'ancurl*familyid': 294.9019607843137},
{'url*ad': 278.0520523785073}]



Evaluating Measures

Evaluation metrics are frequently used to assess categorization performance. This is done with the most popular accuracy measure. A classifier's accuracy on a test dataset is the percentage of datasets that it correctly classifies. Because the accuracy measure is never enough to obtain the proper result in the text mining technique, I included some extra metrics to evaluate classifier performance. Three often used essential metrics are precision, recall, and F-measure.

Before we talk about different metrics, there are a few terms we need to understand:

- The number of successfully identified data is shown by True Positive (TP).
- The number of correct data that has been misclassified is referred to as False Positive (FP).
- The number of correct data that has been misclassified is referred to as False Positive (FP).
- The number of inaccurate data classed as True Negative (TN).

Precision: The accuracy of a classifier is determined by how many accurate documents are returned. A greater precision suggests a reduced amount of false positives, whereas a lower accuracy indicates a higher number of false positives. The ratio of properly categorised examples to total instances is known as precision (P).It can be defined as follows:

```
P = TP/(TP+ FP)
```

Recall: The number of positive data that a classifier returns determines its sensitivity. Fewer false negatives are associated with higher recall. The recall is the ratio of properly categorised examples to the total number of anticipated occurrences. This can be demonstrated as follows:

```
R = TP/(TP + FN)
```

F-Measure: The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall.

```
F = 2 * ((P * R)/(P + R))
```

Accuracy: A classifier's accuracy determines how frequently it gets it right. The accuracy of a prediction is defined as the number of correct predictions divided by the total number of forecasts.

$$Accuracy = \frac{Correct\ prediction}{Total\ cases} * 100\%$$

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} * 100\%$$

III. Result

In my study, I employed Gaussian Naive Bayes ,Support vector Machine Classifier (SVM), Gradient Boosting, Logistic Regression, XG boost, AdaBoost, Multilayer Perceptron (MLP) Random Forest, and Decision Tree as machine learning methods. I have done cross validation methods and K fold gave the best accuracy.

Classifier	Accuracy	F-Score	Mean	Standard Deviation
Knn	98.47%	99.13 %	96.49 %	0.65%
SVM	96.64%	98.11 %	94.81 %	0.98%
Gaussian NB	96.03%	97.75 %	94.92 %	0.72%
Logistic Regression	96.64%	98.10 %	94.81 %	0.875%
Decision Tree	99.39%	99.65 %	95.95 %	0.871%
Random Forest	99.39%	99.65 %	96.95 %	0.832%
AdaBoost	95.88%	97.70 %	95.76 %	0.400%
Gradient Boosting	98.47%	99.13 %	97.10 %	0.484%
MLP	98.32%	99.04 %	96.56 %	0.762%

Decision Tree and Random Forest achieved improved accuracy in dataset, as evidenced by all of the studies. Because the greater number of trees in the forest leads to higher accuracy and prevents the problem of over-fitting. A decision tree is quick and easy to use on big data sets; especially linear data sets.The maximum accuracy was 99.39 percent, based on these findings.

IV. Conclusion

I explained the model's core theory, the methodologies I employed in my research, and the performance metrics for the experiment. I also looked at a variety of research publications on Internet Ad dataset. With the F1 measure, we were able to obtain accuracy of over 95%, precision, and recall of over 95%. For comparing diverse amounts of data, I used K-fold cross validation, and multiple feature extraction processes to reach promising results. In most cases, ten-fold accuracy was preferable, whereas Decision Tree and Random Forest delivered the best classification results.

As part of my research, I attempted to improve all of the extraction methods and pre-processing phases in order to achieve the highest level of accuracy. The use of various pre-processing techniques assisted in the removal of superfluous words. Finally, by using the best features collected from the datasets and learning through correct classifiers, more accuracy might be achieved.

V. Future Work

In the future, I'll look into different aspects of ads, such as semantics, and experiment with various filters. My long-term aims include classifying photos as advertising or non-advertisements, and then further categorising those that are classified as advertisements by source (such as magazine advertisements, newspaper advertisements, etc.). I also want to figure out what commercials are about and categorise them by industry (food, tourism, vehicles, etc.).

VI. Acknowledgment

I would like to start by thanking my supervisor, Professor Dr. Sathis Kumar B, for his help in the study topics and methods. Your informative remarks encouraged me to improve my thoughts and raise the quality of my work. The work described in this paper was supported by the Vellore Institute of Technology Chennai India.

References

[1.] Kushmerick N. Learning to remove internet advertisements. InProceedings of the third annual conference on Autonomous Agents 1999 Apr 1 (pp. 175-181).

[2.] Almgren, Khaled, Murali Krishnan, Fatima Aljanobi, and Jeongkyu Lee. "AD or Non-AD: A Deep Learning Approach to Detect Advertisements from Magazines." *Entropy* 20, no. 12 (2018): 982.