CASE STUDY:

DATA MINING TECHNIQUES FOR SPAM EMAIL CLASSIFICATION: A COMPARISON OF CART AND KNN ALGORITHMS

Iris Analytics : Luke Vincent Samson // Ferdinand Lomerio // John Paul Del Rosario // Carlo Ramirez // Francis Mendaros

ABSTRACT

The use of email as a primary communication channel has increased tremendously, making spam emails a pervasive and significant problem. Therefore, detecting and filtering spam emails is crucial for maintaining email security and privacy. In this study, we employed data mining techniques for classifying spam emails using decision tree (CART) and k-nearest neighbors (KNN) algorithms. We used a publicly available dataset containing attributes related to email content and metadata. Our analysis showed that both classification techniques performed well in detecting spam emails, with KNN outperforming CART in terms of accuracy and recall. Our findings suggest that the combination of data mining techniques and classification algorithms can effectively classify spam emails and provide a framework for developing automated spam email filtering systems.

METHODOLOGY

Using the CRISP-DM framework, it provided a structured approach to developing a spam email classification model using data mining techniques and comparing the performance of CART and KNN algorithms. The methodology allowed us to address the business problem and develop a classification model that can be used to combat spam emails.

CRISP - DM

01. BUSINESS UNDERSTANDING

02. DATA UNDERSTANDING

03. DATA PREPARATION

04. MODELING

05. EVALUATION

01. BUSINESS UNDERSTANDING

The first step of the CRISP-DM framework involves identifying the business problem and understanding the objectives of the study. In this case, the objective is to develop a spam email classification model using data mining techniques and compare the performance of CART and KNN algorithms.

02. DATA UNDERSTANDING

The second step involves collecting and understanding the data. We obtained a publicly available dataset of spam emails containing attributes related to email content and metadata. The dataset consisted of 5,172 emails, with 3002 columns and approximately 3600 of the emails being classified as spam.

Open the Dataset

THE FIRST COLUMN

Indicates Email name. The name has been set with numbers and not recipients' name to protect privacy.

THE LAST COLUMN

Has the labels for prediction :
1 for spam
0 for not spam.

THE REMAINING 3000 COLUMNS

are the 3000 most common words in all the emails, after excluding the non-alphabetical characters/words.

Close Dataset na

03. DATA PREPARATION

The third step involves data preprocessing, cleaning, and transformation. We performed several preprocessing steps on the dataset

PRE PROCESSING

REMOVING UNNECESSARY COLUMNS

Email.No. column was removed as the researchers believed it held no weight as to determine whether an email is spam

FIXING DATA TYPES

Converting character values to numbers to make sure all objects contain numerical data

RENAMING THE BINARY CLASSES TO THEIR RESPECTIVE LABELS

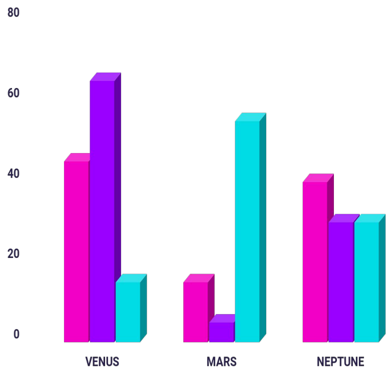
1 for SPAM

0 for NOT SPAM.

PRE PROCESSING

SAMPLING

Since Data is imbalanced, Stratified Sampling was used to sample an equal amount of objects between the two classes, a sample of 1200 objects each was implemented



LOADING...

04. MODELING

The fourth step involves selecting and applying appropriate data mining techniques to develop a classification model. We employed two classification algorithms, decision tree (CART) and k-nearest neighbors (KNN) setup with 80/20 training and testing data split, 1000 best features selected and default parameters for both algorithms. Researchers used the rpart package for R to implement the algorithms.

05. EVALUATION

The fifth step involves evaluating the performance of the classification models using several metrics, including accuracy and precision by using caret, a package for R

Open Caret Confusion Matrix for CART and KNN

CARET CONFUSION MATRIX

CART

```
Confusion Matrix and Statistics

          Reference
Prediction NOT SPAM SPAM
  NOT SPAM      207    6
  SPAM           33  234

               Accuracy : 0.9188
                 95% CI : (0.8906, 0.9416)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8375

 Mcnemar's Test P-Value : 3.136e-05

            Sensitivity : 0.8625
            Specificity : 0.9750
         Pos Pred Value : 0.9718
         Neg Pred Value : 0.8764
             Prevalence : 0.5000
         Detection Rate : 0.4313
   Detection Prevalence : 0.4437
      Balanced Accuracy : 0.9187

       'Positive' Class : NOT SPAM
```

KNN

```
Confusion Matrix and Statistics

          Reference
Prediction NOT SPAM SPAM
  NOT SPAM      133    2
  SPAM          107  238

               Accuracy : 0.7729
                 95% CI : (0.7328, 0.8097)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5458

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5542
            Specificity : 0.9917
         Pos Pred Value : 0.9852
         Neg Pred Value : 0.6899
             Prevalence : 0.5000
         Detection Rate : 0.2771
   Detection Prevalence : 0.2812
      Balanced Accuracy : 0.7729

       'Positive' Class : NOT SPAM
```

RESULTS

In this case study, we demonstrated how CART and KNN can be used to classify email messages as spam or non-spam. Both models were developed and evaluated using the CRISP-DM process.

The CART model achieved an accuracy of 91.8% on the testing set, while the KNN model achieved an accuracy of 77.2%. Although both models performed well, the CART model outperformed the KNN model in terms of accuracy.