

Part 2: Machine Learning

Patrick Collins
1900609

Table of Contents

Context	4
Using Machine Learning Algorithms to Categorise Data	5
Random Forest.....	5
What is this algorithm?.....	5
Strengths of RF.....	6
Weaknesses of RF.....	6
Limitations of RF	6
Naïve Bayes.....	7
What is this algorithm?.....	7
Strengths of NB.....	7
Weaknesses of NB	7
Limitations of NB.....	7
Designing The Classifier	8
Evaluating Classifier Performance	9
Conclusion	11
Appendices	12
Appendix A - Training Dataset.....	13
Appendix B - Testing Dataset	14
References	15

Abbreviations

RF: Random Forest

NB: Naïve Bytes

Context

ScottishGlen is a small company within the energy sector whose employees have been receiving messages from a hacktivist group, who threaten to target the company following a recent blog post by the CEO. It's unclear the nature of the attack they are planning. The CEO is concerned and has asked the IT manager to improve the security posture to better protect the company.

The focus is on making the company resilient to attack therefore the IT manager has asked the technical staff to assess (evaluate) the company's resilience to attack from obfuscated malware. The technical staff managed to get something up and running and have some data for analysis. The system memory appears to be prone to Spyware, Ransomware and Trojan malware but the team is not quite sure how to make sense of it.

Using Machine Learning Algorithms to Categorise Data

The sample data collected by the technical staff for Scottish Glen has multiple attack categories (See Appendices A&B). They are:

- Benign (meaning not malware)
- Spyware
- Ransomware
- Trojan

The technical staff has used a train-test split to get the data therefore it is suitable to choose a supervised learning approach to this problem as a training data set has been created to train the models to predict and classify each memory dump file.

Furthermore, as the objective is to categorise the data according to the attack category a suitable machine learning algorithm to choose would be a multi-class classification algorithm. For example, whether a system memory dump file is Ransomware, a Trojan, Spyware or Benign. This is due to the categories having more than two which a binary classification would be more suitable.

This would also be a form of document classification, sorting the memory dump files to the corresponding category. By doing so, it allows for making sense of the system memory more clearly and faster.

Random Forest

What is this algorithm?

A very popular multi-class classification algorithm is Random Forest. It is a supervised learning algorithm that enables classification of tasks using multiple decision trees. Random Forests can be used for both classification and regression. Classification is predicting the probability of a given class (category). Whereas regression is predicting continuous outcomes. Although, Random Forest is primarily helpful for classification problems.

How it does this is by selecting random samples from the given training set for growing the tree. Then it constructs decision trees for the chosen data points from this training data. The best split out of the variables is used to split the node.

Finally, the leaf node in each decision tree is the binary result that the tree 'votes' which is the most likely class. The most voted prediction result out of all choices is outputted as the final prediction result. It outputs either yes or no, true, or false, or in this case Ransomware or not, Spyware or Not, Trojan or not, Benign or not. The final binary output will be what the memory dump file is categorised as.

Figure 1 on the next page is an example diagram that the Random Forest would look like. Multiple decision trees are utilised, and the highlighted nodes show the decision making the algorithm is making outputting the result. The result from each decision tree is then used to produce the final binary result of the input data assigning the most probable class to it.

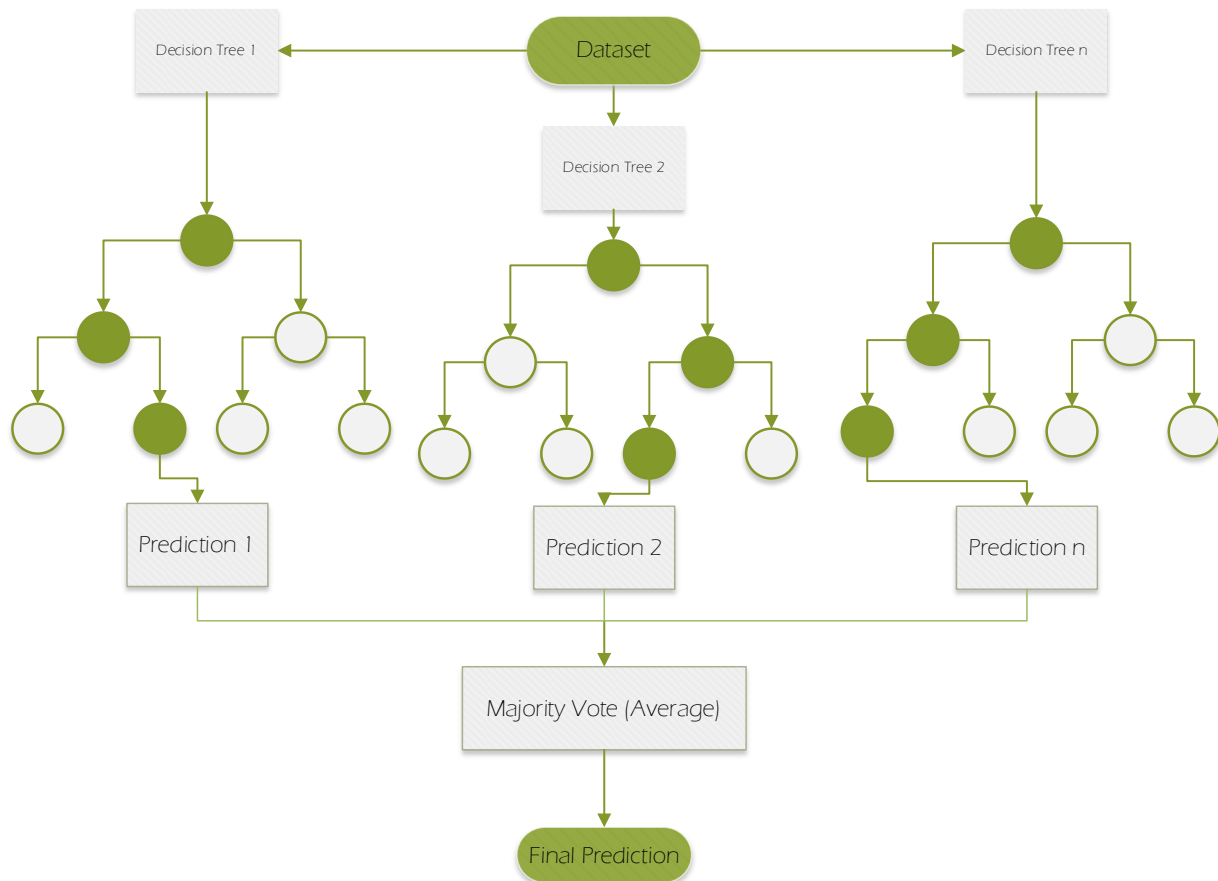


Figure 1: Random Forest

Strengths of RF

There are many strengths to choosing this algorithm to solve a problem. Random Forest fixes the overfitting problem present in standalone decision trees by using multiple trees and producing a result based on a majority vote or average.

Another benefit of choosing RF is that it is the best high-performance algorithm as it can run efficiently on very large datasets. The algorithm also provides flexibility as it can handle both regression and classification tasks making it popular choice for data scientists (IBM, 2023).

Weaknesses of RF

The downside to Random Forest is that it requires more resources as it processes a large dataset. Given this it then causes a time-consuming process as more time is required to process data for each individual decision tree in large data sets.

By providing the strength of more accurate predictions it needs more decision trees to do so, resulting in a slower model. This version of the bagging method is more complex as prediction of a single decision tree is easier to interpret than a random forest (IBM, 2023).

Limitations of RF

Random forest cannot be run when the data is very sparse as it will not produce good results. Some other algorithms like XGBoost perform better in these cases handling smaller data. Finally, if you need extrapolation of data random forest regression is not the best fit to solve this problem.

Naïve Bayes

What is this algorithm?

Another very popular multi-class classification algorithm is Naïve Bayes (NB) which is a supervised learning algorithm. This algorithm can be used for classification tasks, such as document classification making it suitable for categorising the files according to attack category.

The NB classifiers work differently in that they operate under a couple of key assumptions which is how it was given the name “Naïve” (IBM, 2023). It simplifies a classification problem by making it solvable in a smaller number of steps. Meaning that only a single probability is used for each variable making model computation easier (IBM, 2023). This algorithm works better with smaller data sets.

Class-conditional probabilities is the likelihood that a word is in a file. In an example where the file contains ransomware text the following formulae could be used to categorise it to appropriate class where y is ransomware (the class variable) and x is ransomware (feature) (IBM, 2023, Vadapalli, 2022):

$$P(y=[ransomware] | x=ransomware) = P(ransomware | ransomware)$$

Strengths of NB

Naïve Bayes is considered a simpler classifier since the parameters are easier to estimate (IBM, 2023). This makes it easy to understand and implement. The algorithm is considered a fast and efficient classifier that is fairly accurate unlike other models such as linear regression.

A benefit over Random Forest is this algorithm has low storage requirements as not a lot of resources is needed to process the data set. Depending on the classification problem to solve, like document classification, it can effectively manage a high number of dimensions (IBM, 2023).

Weaknesses of NB

However there are also some downsides to using this algorithm. One major one being it has zero frequency issues. If a category variable does not exist within the training set this will then cause an issue in the model (IBM, 2023). This could cause major problems therefore should only be chosen if the classification problem will be suitable for this. Another weakness is that the NB incorrect assumption can lead to incorrect classifications (IBM, 2023).

Limitations of NB

Finally, one limitation of Naïve Bayes is that it is not suitable for real world cases (Vadapalli, 2022). The algorithm is prone to assumptions that all features are independent which rarely happens in real world cases and should not be chosen for real-world applicability and classification problem solving.

Designing The Classifier

The IT manager recommends that ScottishGlen implement the Random Forest machine learning algorithm to make sense of the memory dump files. For the problem facing the business using a Random Forest is more suitable over Naïve Bayes as it will outperform it and run more efficiently.

A good rule of thumb to create a classifier is to complete all stages of a Data Pipeline. The technical team have already captured the data needed to create a classifier completing the Data Acquisition stage and then splitting it into training data and testing data (Appendix A & B).

Therefore, the first step the technical team should take is to perform data pre-processing on the training data. This is to reduce any variables that could disrupt the model process. The training set should be used for training the model and building the classifier which is the next stage of the pipeline.

To build the Random Forest classifier the team should focus on document classification. The goal of document classification is to create a classification model that can accurately assign documents to the right categories (Orza, 2022). As a reminder the categories are:

- Benign
- Spyware
- Ransomware
- Trojan

The classes in the forest should be set according to the attack category. The training data will first be fed to train multiple decision trees. Bootstrapping will then be performed with the algorithm selecting random samples. The algorithm will then construct a decision tree for every training data sample. As mentioned previously, the final prediction of all the decision trees will be the most likely category for each file.

Many programs may be utilised to complete this task though the IT manager recommends the team to use Python as the programming language and to use the “sklearn” python modules from Scikit-Learn. Its “RandomForestClassifier” is trained using bootstrap aggression making it very suitable for designing this classifier and for easy implementation (scikit, 2023).

Once the classifier model is ready the final stages which are to evaluate and test its accuracy using the testing data set which is discussed in the next section.

Evaluating Classifier Performance

The objective for this stage is to check the model's accuracy and performance at correctly identifying the system memory dump category (Benign, Spyware, Ransomware or Trojan).

To do this the team should test the classifier model on the testing data set previously generated. By testing the accuracy on the unseen testing set the accuracy of the model is obtained. For example, new data having a prediction accuracy of 78%. Small tweaks to the model can improve the accuracy even further.

However, it's important to choose a good evaluation metric that is appropriate to test a machine learning model. Two very popular metrics for testing a Random Forest classifier is a Confusion Matrix and Area Under Curve-Receiver Operating Characteristic (AUC-ROC).

A Confusion Matrix will show the results of how well the model correctly and incorrectly predicted the outcome. Scikit also provide a function to perform this evaluation metric called "confusion_matrix()" giving the team an easy implementation. When loading in the labels it will let the team know how accurate the classifier is performing and with the results of how many wrong predications are being made be able to know where the model is failing and improve the prediction outcome. Run the classifier through the metric again to test for lower incorrect predictions. A simple example can be seen in figure 2 to give the team an idea.

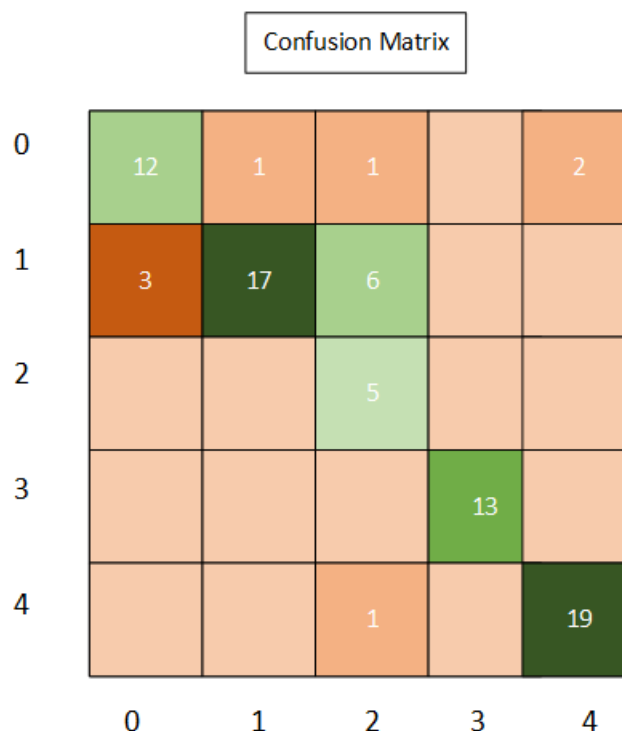


Figure 2: Confusion Matrix example

Another popular evaluation metric is AUC-ROC. Like Confusion Matrix Scikit also provide a function to perform this evaluation metric called "roc_curve()". The closer the AUC is to 1, the better. Area Under the Curve (AUC) of 1.00 is ideal. The classifier can be evaluated on how close it gets to the ideal. A simple example can be seen in figure 3 on the next page to give the team an idea.

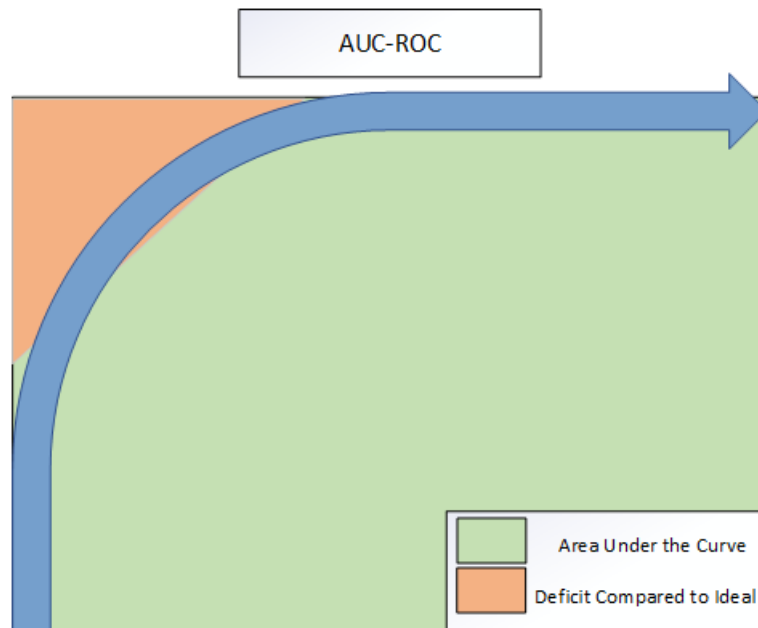


Figure 3: AUC-ROC example

Conclusion

Finally, now that the Random Forest classifier has been built and has been evaluated the final stage is to communicate the results. How effective and accurate is the model's prediction? Does the Random Forest classifier now help the technical team make sense of the memory dump files? Most importantly, is ScottishGlen reasonably safe from the hacktivist threat now that a classification model is in place to rapidly analyse suspicious memory files? If any doubt is at the stage, it may be necessary to redesign the classifier and possibly even choose a new machine learning algorithm (such as Naïve Bayes as discussed) to test its prediction outcomes.

Appendices

[illegible]

CMP417 Engineering Resilient Systems 1

Page 14 of 15

References

Carrier, T.; Victor, P.; Tekeoglu, A. and Lashkari, A. (2022). Detecting Obfuscated Malware using Memory Feature Engineering. In Proceedings of the 8th International Conference on Information Systems Security and Privacy - ICISSP, ISBN 978-989-758-553-1; ISSN 2184-4356, pages 177-188. DOI: <https://doi.org/10.5220/0010908200003120>

H. Lashkari, B. Li, T. L. Carrier and G. Kaur, "VolMemLyzer: Volatile Memory Analyzer for Malware Classification using Feature Engineering," 2021 Reconciling Data Analytics, Automation, Privacy, and Security: A Big Data Challenge (RDAAPS), Hamilton, ON, Canada, 2021, pp. 1-8, doi: <https://doi.org/10.1109/RDAAPS48126.2021.9452028>

IBM, 2023, What is naïve Bayes, IBM. Available at: <https://www.ibm.com/topics/naive-bayes> [Accessed: April 16, 2023].

IBM, 2023, What is Random Forest?, IBM. Available at: <https://www.ibm.com/topics/random-forest> [Accessed: April 16, 2023].

Orza, P, 2022, Document classification: A complete guide, Document Classification: A Complete Guide. Available at: <https://levity.ai/blog/document-classification-guide> [Accessed: April 23, 2023].

Scikit, 2023, Oob errors for random forests, OOB Errors for Random Forests. Available at: https://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html#sphx-glr-auto-examples-ensemble-plot-ensemble-oob-py [Accessed: April 23, 2023].

Vadapalli, P, 2022, Naive Bayes explained: Function, Advantages & disadvantages, applications in 2023, upGrad blog. Available at: <https://www.upgrad.com/blog/naive-bayes-explained/> [Accessed: April 17, 2023].