Aprendizagem Computacional em Biologia
Inteligência Geoespacial
Reconhecimento de Padrões

2024/2025

Project Assignment

Chasing Phishing URLs

1 Background

Every day, we face numerous attempts to obtain personal information without our permission. These attempts, known as phishing actions, typically arrive via SMS or email and often include URL links designed to capture information. Therefore, identifying such malicious URLs is crucial to maintaining our security.

2 Objective

Your task is to develop classifiers to identify if a given URL is or not related to *phishing*.

3 Practical Assignment

3.1 Dataset Description

To develop your classifiers, you must use the dataset available at https://archive.ics.uci.edu/dataset/967/phiusiil+phishing+url+dataset, called the PhiUSIIL Phishing URL Dataset. The dataset includes 134,850 legitimate URLs and 100,945 phishing URLs. Features were extracted from the webpage's source code and from the URL string. Additionally, other features were derived from these base features (CharContinuationRate, URLTitleMatchScore, URLCharProb, and TLDLegitimateProb), which can be referred to as hyper-features. In total, 54 features are available. To ease your work, consider only the non-categorical features. For more details about the dataset, please refer to https://doi.org/10.1016/j.cose.2023.103545.

3.2 Feature Selection and Reduction

Some of the supplied non-categorical features may be useless, redundant, or highly correlated with others. At this stage, consider using feature selection and dimensionality reduction techniques to evaluate their impact on the performance of the pattern recognition algorithms. Analyze the distribution of your feature values and calculate the correlation between them. Ensure you understand your features! Remember to include your findings in the final report.

3.3 Experimental Analysis

You should be able to design experiences that run the pattern recognition algorithms learned in the course on the provided data and evaluate their results. Keep in mind that this is an unbalanced binary data set. Aim to design the classifier while considering these challenges. Justify your assumptions and decisions. Define the performance metrics to evaluate your method (e.g., Sensitivity, Specificity, F-measure, and ROC Curves). To run the experiments multiple times and present average results and standard deviations (for the metrics used), you should use cross-validation. Remember that manually inspecting your algorithm's predictions can provide valuable insights into where it is failing and why, as well as how to improve it (e.g., what causes the algorithm to fail in specific cases? What unique characteristic makes it particularly challenging? How can I help the algorithm handle those cases more effectively?). Reassess the Pre-processing, Feature Reduction, and Feature Selection phases until you are satisfied with the results. It's wise to monitor the evolution of your algorithm's performance throughout this process. Aim to display these trends in your final report to support all related issues (parameter selection, model fit, etc.).

3.4 Pattern Recognition Methods

You can write your own code or utilize the available functions and methods. The methods employed in your work should be described, along with a discussion of the parameters used. Experiment with different pattern recognition algorithms. You should aim to understand how they perform differently on your data.

3.5 Results and Discussion

Present and discuss the results obtained from your project assignment. This issue has already been studied by the authors of the dataset in [1]. Compare your findings with the results they achieved.

3.6 Code

You should submit your software code in your chosen language. Remember to comment your code. Additionally, include a help section that describes the function's purpose, its usage, and an explanation of the parameters.

4 Documentation

Create documentation (in Portuguese or English) for your project. The documentation should feature a cover page that includes the course name, project title, date, and the names and student numbers.

Detail the methods used so that the reader can implement similar functionalities based solely on your documentation. Always justify your choices, even when they are based on intuition. Don't forget to verify your assumptions! Your documentation should include classification results with the given data. At the end, you should have a list of all references used.

4.1 Requirements

The practical assignment should be developed in pairs. However, students who prefer to work individually are also welcome. Larger groups are not permitted.

4.2 Project Submission & Deadlines

1. **Project First Milestone (Deadline: March 21, 2025!)**
   Deliverables:

- Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
- Minimum Distance classifier+Fisher LDA;
- Code + short report.

2. **Project Final Deadline (Deadline: May 11, 2025!)**
   Deliverables:
   - Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
   - Several classifiers;
   - Final Report
   - Code
3. **Presentation and Discussion (May 20 and 22, 2025!)**

Acknowledgments

Credits to Arvind Prasad (arvindbitm@gmail.com) and Shalini Chandra, Babashaheb Bhimrao Ambedkar University [1].

References

[1] Prasad, A., & Chandra, S. (2023). PhiUSIIL: A diverse security profile empowered phishing URL detection framework based on similarity index and incremental learning. Computers & Security, 103545. doi: https://doi.org/10.1016/j.cose.2023.103545