Fuzzy Expert System to Detect Phishing in Websites

Dániel MÁCSAI
Ismael RUIZ GARCIA
Mauro VÁZQUEZ CHAS

Master in Artificial Intelligence

UNIVERSITAT ROVIRA i VIRGILI

Planning and Approximate Reasoning
Delivery 3

15th December 2024

Contents

1 Introduction

For this work, we

2 Task 1

To design the fuzzy expert system to detect phishing websites, we consulted [3]. In this paper, they list 87 possible features (boolean, floats and integers) that could matter in the detection of phishing websites. The proposed features are divided into three categories: URL-based features, content features and external features. From this proposed variables, we selected 5 features that we consider relevant for the detection of phishing websites. For inspiration and further understanding of the problem, we consulted the following articles: [6] and [2].

2.1 Chosen Features

2.1.1 URL-based features:

Phish Hints

- **Description** Number of words in the URL that are typical of phishing websites
- **Integer** Number 51 in the paper

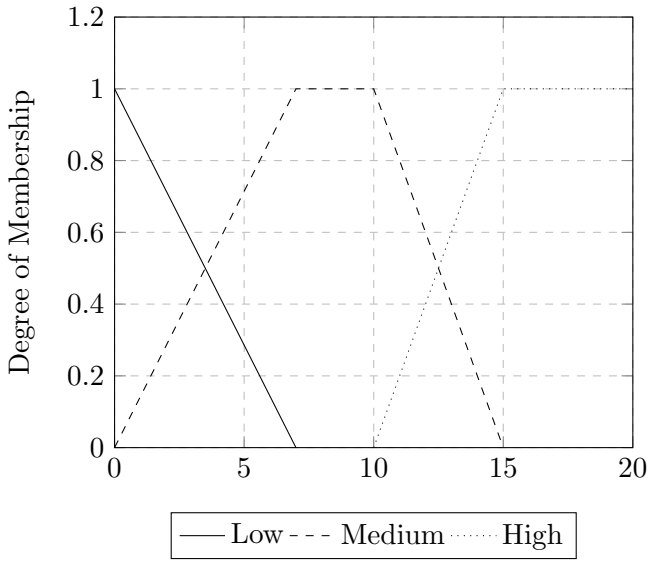

Figure 1: Membership Function Phish Hints

Domain Age

- **Description:** Age of the page in months
- **Integer** Number 83 in the paper

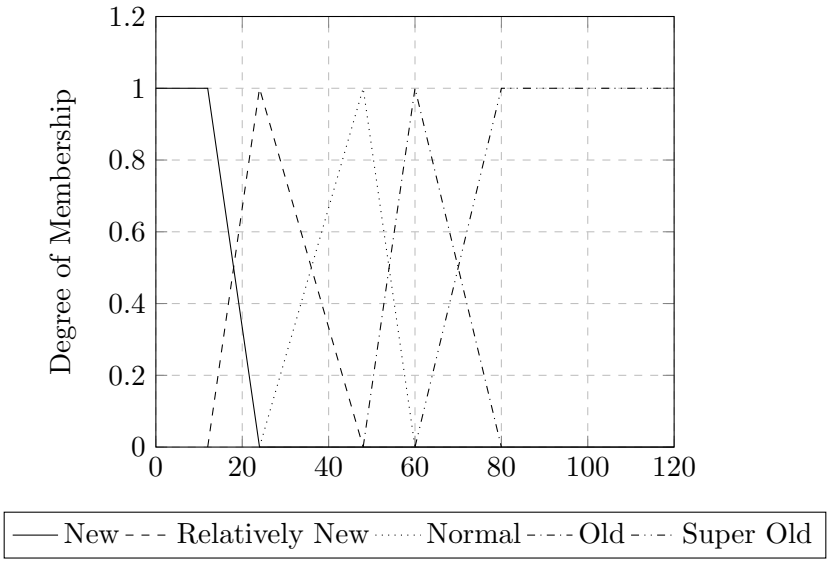

Figure 2: Membership Function Domain Age (in months)

2.1.2 Content features:

Ratio External Hyperlinks

- **Description:** The number of external hyperlinks in a web page divided by the total number of hyperlinks
- **Float** Number 59 in the paper

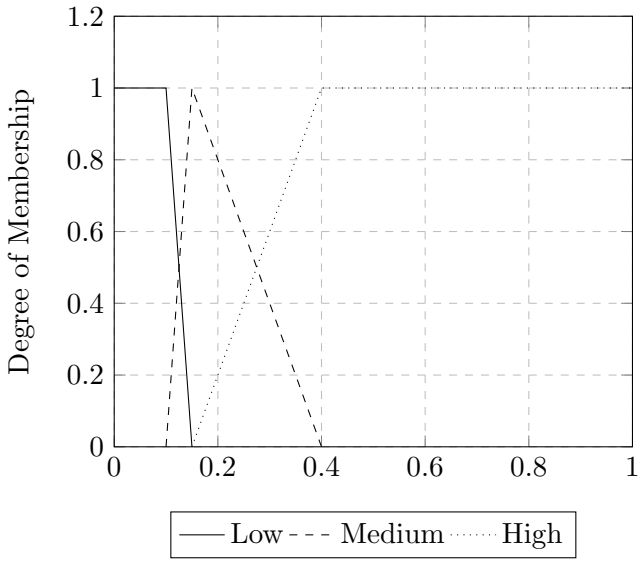

Figure 3: Membership Function Ratio External

2.1.3 External features:

Google Index

- **Description:** Whether a page is indexed in Google
- **Boolean** Number 86 in the paper

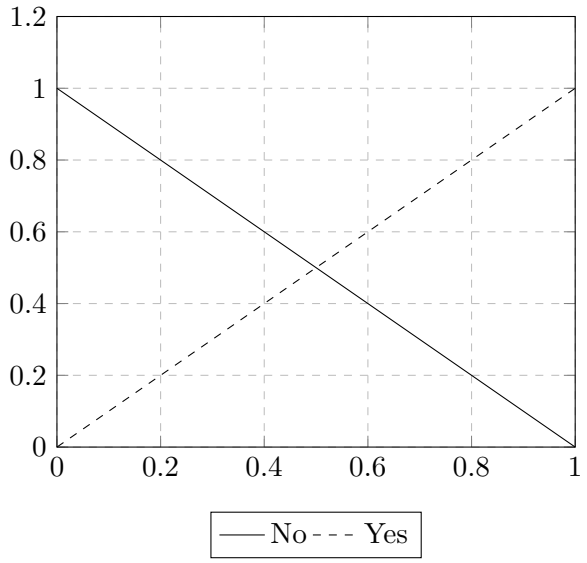

Figure 4: Membership Function Google Index

GTR

- **Description:** It is a slight modification of the usual GTR, created by Google. It means Google Toolbar Rank and takes values from 0 to 10.
- **Integer** Number 87 in the paper

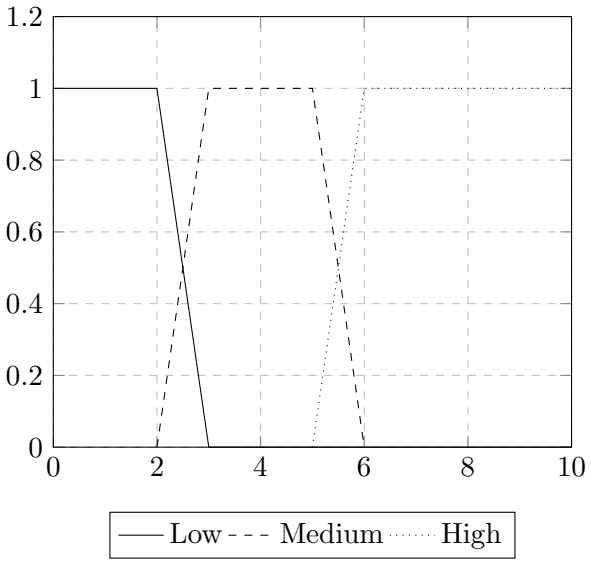

Figure 5: Membership Function Google GTR

3 Output Variable

Our output variable will be the phishing risk, where we will consider 5 different fuzzy sets, see 6.

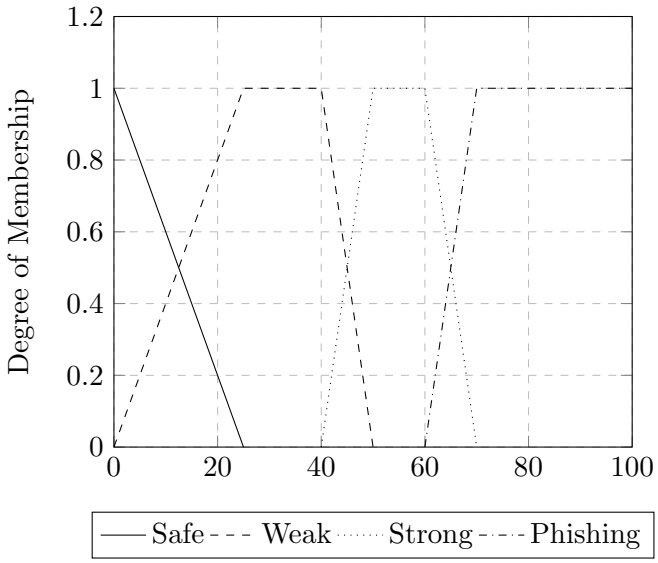

Figure 6: Membership Function Phishing Risk (Output Variable)

4 Rules

To design the rules involving Google Index, we consulted [3], where they state that pages not indexed by google are more likely to be phishing websites. For this reason we created the rules in 1.

Google Index	**Phishing Risk**	**Weight**
NO	PHISHY	1
YES	WEAK	1

Table 1: Rules for Google Index

For the GTR, we consulted [5], where they state the following:

> *GTR value is considered as a heuristic because PageRank value for legitimate site will be high and for phishing pages its value will be less*

In the case of the Phish Hints, it is already explained in [3] that more Phish Hints in the URL is an indicator of a phishing website. For this reason, we introduced the rules in Table 2.

GTR	**Phis Hints**	**Operator**	**Phishing Risk**	**Weight**
HIGH			SAFE	1
MEDIUM	LOW	AND	WEAK	0.5
MEDIUM	MEDIUM	AND	STRONG	1
MEDIUM	HIGH	AND	PHISHY	1
LOW	LOW	AND	STRONG	0.5
LOW	MEDIUM	AND	STRONG	1
LOW	HIGH	AND	PHISHY	1

Table 2: Rules for Google GTR and Phish Hints

To evaluate the effect of the domain age feature, we consulted [4], where they state the following:

> *The top feature in the list is domain age, which confirms our assumptions that long-running services are statistically more credible*

For this reason, we will consider that the older the domain, the less likely it is to be a phishing website. On the other hand, to evaluate the effect of the ratio of external hyperlinks, we consulted [1], where the following is affirmed:

> *Phishing websites often include numerous external hyperlinks pointing to target websites because cybercriminals frequently replicate the HTML code from legitimate websites to construct their phishing sites.*

All of this information, lead to the creation of the rules seen in Table 3.

Domain Age	**Ratio of Hyperlinks**	**Operator**	**Phishing Risk**	**Weight**
SUPER OLD			SAFE	1
OLD	LOW	OR	SAFE	0.5
NORMAL	LOW	AND	WEAK	0.5
NORMAL	MEDIUM	AND	STRONG	0.5
NORMAL	HIGH	AND	STRONG	0.5
RELATIVELY NEW	LOW	AND	WEAK	1
RELATIVELY NEW	MEDIUM	AND	STRONG	1
RELATIVELY NEW	HIGH	AND	PHISHY	0.5
NEW	LOW	AND	STRONG	0.5
NEW	MEDIUM	AND	STRONG	1
NEW	HIGH	AND	PHISHY	1

Table 3: Rules for Domain Age and Ratio of External Hyperlinks

5 Implementation

In our implementation, we utilized the MATLAB Fuzzy Logic Toolbox to develop the fuzzy expert system. The system was designed using the Mamdani fuzzy inference method, which is well-suited for decision-making processes that require a human-like reasoning approach. In this system, we employed the minimum (min) operation as the t-norm for the intersection of fuzzy sets, and the maximum (max) operation as the t-conorm for the union of fuzzy sets. For the defuzzification process, we selected the Center of Area (CoA) method, which calculates the centroid of the aggregated fuzzy set to produce a crisp output. This approach ensures that the output is a balanced representation of the input conditions, providing a reliable decision-making framework for detecting phishing websites.

To validate the fuzzy expert system, we employed the MATLAB 3D plot. As this plot only allows the representation of 2 inputs and the output at a time, we viewed the 10 possible combinations, all of which were consistent. At the same time, every possible combination was covered, as no combination was left without a rule being activated.

6 Testing

- 4 Test case that represent different situation (some must activate more than one lable) - Report the results of each testing case with screenshots and explanations that justify the output obtained (i.e. showing the activations of rules)

7 Complex Fuzzy Expert System

Design (just graphically, no implementation) a more complete fuzzy expert system that includes more features about the websites. Show in a figure the inputs, outputs, and rule blocks that you propose for such expert system. No specific definition of variables nor rules is required.

- We can use a hierarchical rule system

References

[1] Ali Aljofey et al. "An effective detection approach for phishing websites using URL and HTML features". In: *Scientific Reports* 12.1 (May 2022), p. 8842.

[2] Zuochao Dou et al. "Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection". In: *IEEE Communications Surveys Tutorials* 19.4 (2017), pp. 2797–2819. DOI: 10.1109/COMST.2017.2752087.

[3] Abdelhakim Hannousse and Salima Yahiouche. "Towards benchmark datasets for machine learning based website phishing detection: An experimental study". In: *Engineering Applications of Artificial Intelligence* 104 (2021), p. 104347. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2021.104347. URL: https://www.sciencedirect.com/science/article/pii/S0952197621001950.

[4] Radek Hranický et al. "Unmasking the Phishermen: Phishing Domain Detection with Machine Learning and Multi-Source Intelligence". In: (2024), pp. 1–5. DOI: 10.1109/NOMS59830.2024.10575573.

[5] A. Naga Venkata Sunil and Anjali Sardana. "A PageRank based detection technique for phishing web sites". In: (2012), pp. 58–63. DOI: 10.1109/ISCI.2012.6222667.

[6] Rasha Zieni, Luisa Massari, and Maria Carla Calzarossa. "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites". In: *IEEE Access* 11 (2023), pp. 18499–18519. DOI: 10.1109/ACCESS.2023.3247135.