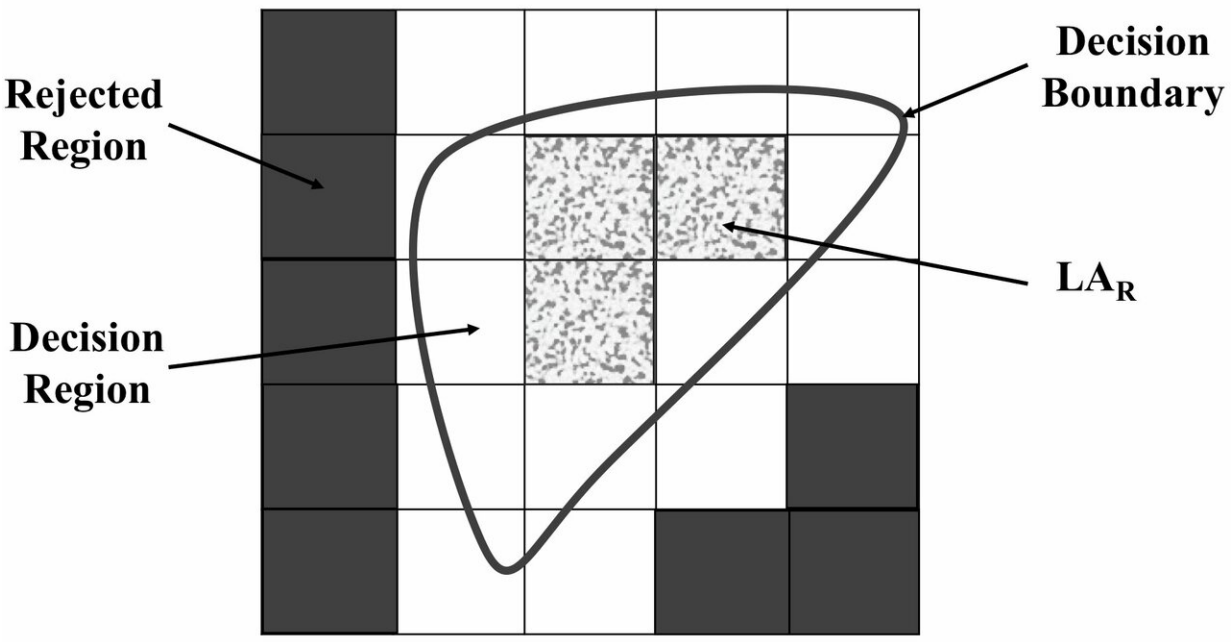
Selecting the best features for phishing attack detection algorithms

April 1 2019, by Ingrid Fadelli



The universe of discourse regions segregated by FRS. Credit: Zabihimayvan & Doran.

In recent decades, phishing attacks have become increasingly common. These attacks allow attackers to obtain sensitive user data, such as passwords, usernames, credit card details, etc., by tricking people into disclosing personal information. The most common type of phishing attack is email scams in which users are led to believe that they need to give their details to an established or trusted entity, while they are, in

fact, sharing this data with someone else.

IT professionals have developed a vast number of tools and strategies to detect and prevent phishing attacks, many of which are based on machine learning. The performance of such machine-learning algorithms often depends on the features they extract from websites.

Researchers at Wright State University have recently developed a new method to identify the best sets of features for phishing attack detection algorithms. Their approach, outlined in a paper pre-published on arXiv, could help to enhance the performance of individual machine-learning algorithms for uncovering phishing attacks.

"The performance of phishing detection algorithms that use machine learning strongly depends on the features of a website the algorithm considers, including the length of web page URL or if special characters like @ and dash exists in the URL," Mahdieh Zabihimayvan and Derek Doran, the two researchers who carried out the study, told TechXplore via e-mail. "In this work, we wanted to make it easier to build machine-learning algorithms for phishing detection by automatically recovering a 'best' set of features for any phishing detection algorithm, irrespective of the website under consideration."

While there are now several algorithms to identify phishing attacks, so far, very few studies have focused on determining the most effective features for detecting this particular type of attack. In their study, Zabihimayvan and Doran addressed this gap in the literature, by trying to uncover the most effective features for this particular task.

"We applied Fuzzy Rough Set (FRS) theory as a tool to select the most effective features from three benchmarked phishing website datasets," Zabihimayvan and Doran said. "The selected features are then used for three often used machine-learning algorithms for phishing detection."

To test the effectiveness and generalizability of their FRS feature selection approach, the researchers used it to train three commonly employed phishing detection classifiers on a dataset of 14,000 website samples and then evaluated their performance. Their evaluations yielded highly promising results, reaching a maximum F-measure of 95 percent when their feature selection method was applied to a random forest (RM) classifier.

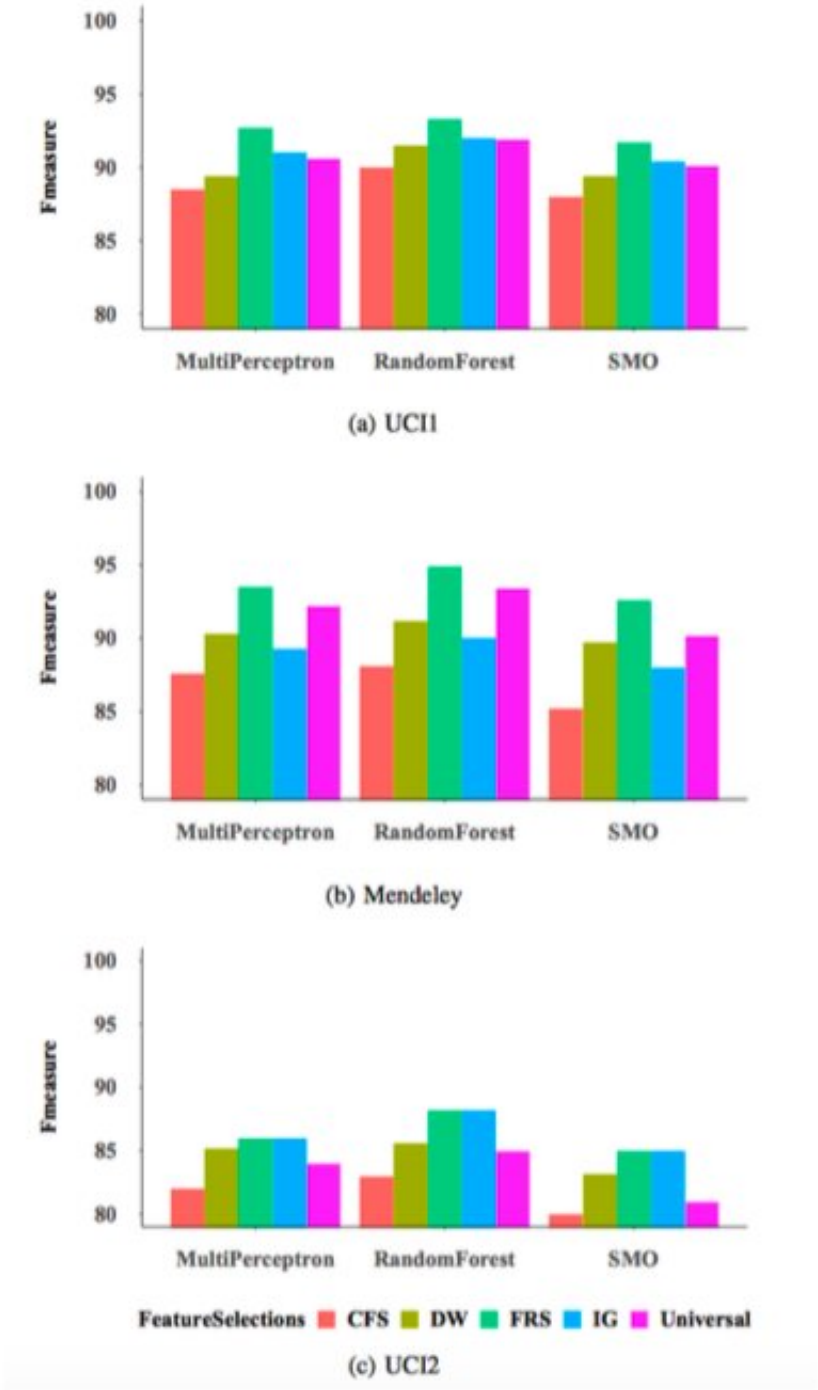
"FRS discovers feature dependencies based on the data," Zabihimayvan and Doran explained. "In other words, FRS decides how to separate a set of data based on their feature values and labels using a decision boundary and a similarity relation declared in the form of fuzzy membership functions. Features selected by FRS are the ones which can more distinguish between data samples that belong to different classes."

The FRS approach used by Zabihimayvan and Doran selected nine universal features across all datasets used in their study. Using this universal feature set, they attained an F-measure of approximately 93 percent, which is similar to that achieved by classifiers using their FRS approach. The universal feature set contains no features from third-party services, so this finding suggests that one could potentially detect phishing attacks faster with no inquiry from external sources.

"The features selected automatically by FRS yield the best detection performance across a number of classifiers," Zabihimayvan and Doran said. "We also find a set of 'universal features' – those aspects of a web page that FRS found to best predict if a page is attempting to fish information, no matter the type of website the page tries to mimic."

The study carried out by Zabihimayvan and Doran is one of the first to provide valuable insight about the most effective features for detecting phishing attacks. In the future, their work could pave the way for the development of more efficient and reliable phishing detection

techniques, which would uncover these attacks faster than current methods.



F-measure for different classifiers and features sets. Credit: Zabihimayvan &

Doran.

"We now hope to extend our study further by investigating feature selection for more sophisticated machine-learning algorithms, including deep learning architectures that automatically discover 'meta-features' to further enhance detection performance," Zabihimayvan and Doran said. "We also plan to extend our feature selection framework to detect phishing emails."

More information: Fuzzy rough set feature selection to enhance phishing attack detection. arXiv:1903.05675 [cs.LG]. arxiv.org/abs/1903.05675

Citation: Selecting the best features for phishing attack detection algorithms (2019, April 1) retrieved 25 April 2024 from https://techxplore.com/news/2019-03-features-phishing-algorithms.html