KANSAS STATE
UNIVERSITY
Computer Science

Android Malware Detection with Weak Ground Truth Data

Jordan DeLoach[1], Doina Caragea[1], Xinming Ou[2]

[1]Kansas State University [2]University of South Florida

Abstract

For Android malware detection, precise ground truth is a rare commodity. As knowledge evolves, what may be considered ground truth at one moment may change, and apps once considered benign may turn out to be malicious. The inevitable noise poses a challenge to crafting effective machine learning classifiers. Our work is focused on learning classifiers for Android malware detection in a manner that is sound with regard to the uncertain and changing ground truth. While you can be confident that an app is malicious, you can never be certain that a benign app is really benign, or just undetected malware. Based on this insight, we leverage a modified Logistic Regression classifier that allows us to learn from only positive and unlabeled data. We find Label Regularized Logistic Regression to perform well for noisy datasets, as well as datasets where there is a limited amount of labeled data, both of which are representative of real-world situations.

Background

- To generate a class label for a given Android app, the common approach is to use an ensemble of anti-virus (AV) solutions.
- Anti-virus solutions tend to bias initially towards false negatives as opposed to false positives.
- This creates a scenario with a comparative higher certainty in a single label.

Problem

- We seek to explore machine learning solutions that leverage the one-class insight.
- We seek to study the problem with consideration to how class labels may evolve over time as more knowledge is gathered.

Methods

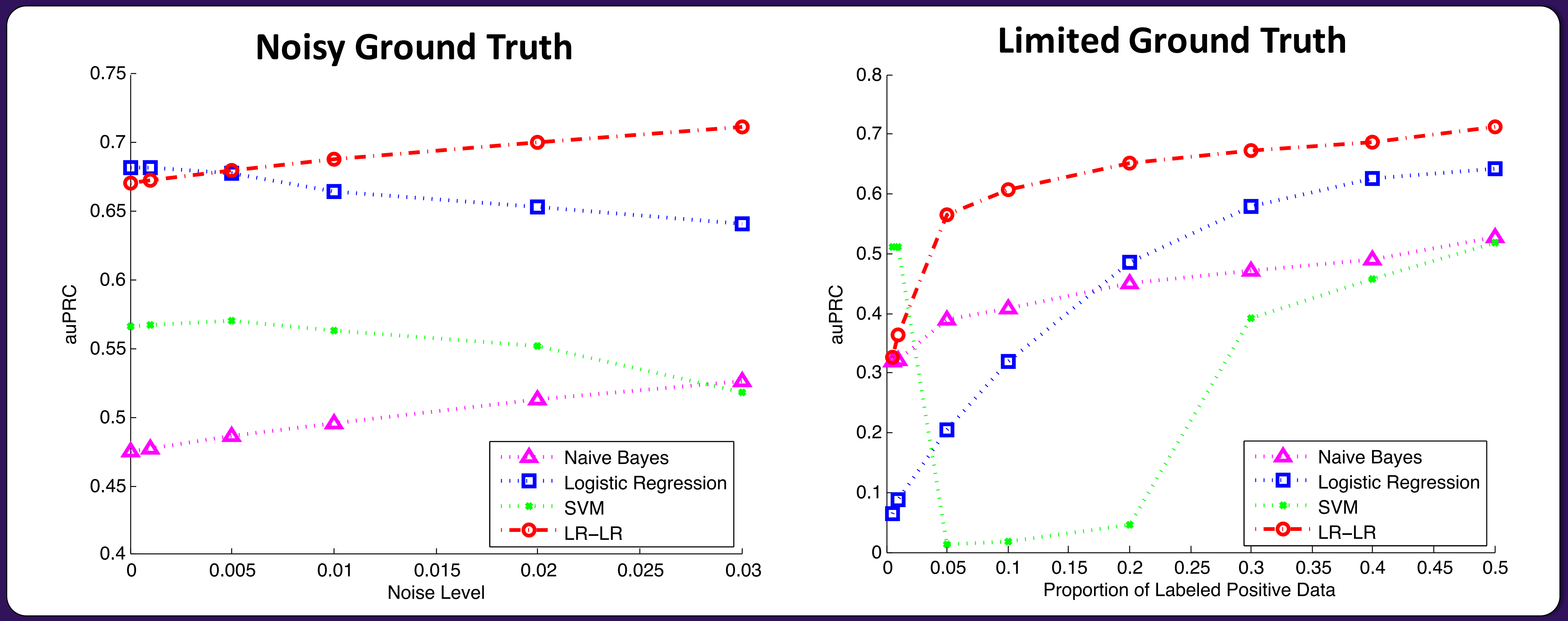
- Label Regularized Logistic Regression[1] (LR-LR) a one-class semi-supervised learning algorithm that is used to train on only the positive class, while regularizing expectations over the benign apps as unlabeled via the constant $\tilde{p}$.
- We implement LR-LR in Spark to run on over a million apps from the Google Play Store in addition to malicious apps from industry collaborators, with labels from VirusTotal[2].

$$O(\theta) = \sum_{i}^{N_p} \log p_\theta(y_i|x_i) - \lambda^U D(\tilde{p}||\hat{p}_\theta^{unlab}) - \lambda^{L^2} \sum_k \theta_k^2$$

$\tilde{p}$	**Expert prediction of distribution of unlabeled**
$\hat{p}_\theta^{unlab}$	**Current classifier expected distribution**
$\log p_\theta(y_i\|x_i)$	**Log likelihood over the data`**
$D(\|\|)$	**The Kullback-Leibler Divergence between two distributions**
λ^U	**Label Regularization Parameter**
λ^{L^2}	**L^2 Regularization Parameter**

Experiments and Results

- In the **Noisy Ground Truth** experiment, we aim to simulate the inherent noise that exists in class labels by systematically inducing noise in the training set.
- We vary the amount of noise in the training data and experiment with which approaches, both LR-LR and comparative supervised techniques, perform best in accounting for this noise.
- We tune the LR-LR $\tilde{p}$ parameter to account for this noise in the ground truth.
- In **Limited Ground Truth**, we see which approach is most effective with the least amount of labeled positive data.
- The aim is to see how using a small amount of high quality data outperforms solutions with more, lower quality data.
- In both experiments, we see LR-LR outperform comparative supervised approaches.



Discussion

- **In Noisy Ground Truth**, intuitively Logistic Regression outperforms LR-LR initially as it can leverage both classes.
- As the ground truth gets more noisy, LR-LR can better account for that and outperforms LR and other supervised approaches.
- **In Limited Ground Truth**, LR-LR, due to its semi-supervised nature, is able to perform better with a limited ground truth set.
- This is useful for restricting due to the cost of acquiring high confidence ground truth.

Conclusions

- We leveraged one-class semi-supervised learning for the first time in the Android malware detection field.
- We showed how LR-LR can remedy the weaknesses in current Android ground truth labeling approaches.
- Our results showed that the LR-LR approach works well in this domain, leading us to believe it is a promising area of further research.

Future Work

- We intend to continue to study how class labels for Android apps evolve over time and what insights can be extracted from that data.
- We intend to study how we can expand our feature sets used to learn classifiers from by incorporating data from social media sources.

References

[1] Ritter, A.; Wright, E.; Casey, W.; and Mitchell, T. 2015. Weakly supervised extraction of computer security events from Twitter. In Proc. of the 24th WWW, 896–905. ACM.

[2] Roy, S.; DeLoach, J.; Li, Y.; Herndon, N.; Caragea, D.; Ou, X.; Ranganath, V. P.; Li, H.; and Guevara, N. 2015. Exper- imental study with real-world data for Android app security analysis using machine learning. In Proc. of the 31st AC- SAC, 81–90. ACM.