WEBROOT
Smarter Cybersecurity™

The Webroot Approach to Machine Learning

Automating Threat Detection with Advanced Machine Learning at Scale

Introduction

The most dangerous cyber threats to organizations and individuals hide within everyday network traffic, cleverly disguised to avoid detection. Faced with a near constant stream of potential threat warnings, actual infections, and information on network activity, organizations of all sizes may struggle to successfully uncover threats. The heart of the issue here is that man is incapable of handling such an enormous amount of data and data analysis. That's where we have to rely on the strength of our machines, which have the processing power to comb through and analyze of all of the noise, then identify the items that truly need attention. Advanced machine learning can be used to classify the enormous volume of available data that is overwhelming threat researchers and traditional defenses; it can reduce the false positives/negatives, as well as the workload for human analysts, thereby enabling an organization's staff to focus exclusively on the actual threats themselves.

Webroot Machine Learning Models

By automating security intelligence using machine learning at scale, Webroot is solving significant problems in cybersecurity. We collect an ever increasing volume of data, and our networks, storage, and processing power make it possible to rapidly process massive data sets to detect previously unknown and never-before-seen threats in real time or near real time.

Webroot utilizes over 500 classifiers operating in parallel across URLs, IPs, files, multiple languages, etc. to recognize patterns, determine reputations, and accurately categorize internet objects. The number of threats that we can uncover daily is linearly related the number of classifiers we employ. For example, Webroot finds an average of 736,000 new malicious files every month. These include portable executables (PE files), JavaScript, Java files, VBScript, Android™ Application Packages (APKs), etc. which can all be used to deliver a malicious payload. Webroot has classified and re-classified over 27 billion URLs across 600 million domains, representing the well over 95% of the internet. We prioritize URLs and domains by how many people visit those websites, by the real-world frequency of traffic.

Size and Scale

Imagine that you are characterizing individual humans. You might consider attributes such as height and hair color. Height, as a quantitative characteristic, would be represented with numerical values or floating point numbers, while hair color would constitute a qualitative attribute or categorical value. There might also be various nuances between hair color or skin tone which must all be reflected in the data set. Additionally, we must consider sequential data, such as patterns of behavior or presentation. As applied to malware, considering so many different characteristics yields an exponentially high number of possibilities, which must be analyzed and categorized to determine whether something is a threat.

The number of characteristics (input vectors) that Webroot machine learning technology uses to evaluate an internet object is extremely large. When we encode the information about an object, we essentially create a dictionary of characteristics. Encoding the information contained in these characteristics in a form suitable for machine learning yields a massive quantity of different types of attributes for a given internet object, such as a file, IP address, URL, potential phishing site, or mobile application. These are called high dimensional input vectors. Numerical values may be one dimension, categorical values require additional dimensions, as do sequential values. For example, we capture all of the characteristics on a web page that help describe that particular page. These are then added to the dictionary of attributes. For some of Webroot's machine learning applications, we have the ability to capture up to 10 million characteristics pertaining to a single object, and our machine learning automates the research and classification of millions of objects daily.

By capturing up to 10 million characteristics, Webroot is able to collect and analyze practically any information pertaining to an internet object and determine if it poses a threat at the precise time of analysis. To make the machine learning results actionable, Webroot then assigns every internet object a reputation score ranging from one to one hundred. Objects receiving a score between one and twenty are considered malicious. Reputation scores are critical as they allow Webroot technology partners to consider the shades of gray in cybersecurity, rather than relying on a basic, binary good/bad determination. Partners can then fine-tune the scores at which their devices will block or tolerate IPs, URLs, files, etc.

Web analysts also provide direct feedback to the machine learning model through an "active learning" process. They evaluate behaviors like phishing redirects, then train and improve the algorithms and functions over time. Constant monitoring and feedback is crucial because the human feedback identifies new versions of files, phishing sites, etc. as well as very tiny modifications, and additional characteristics that can be added to the model.

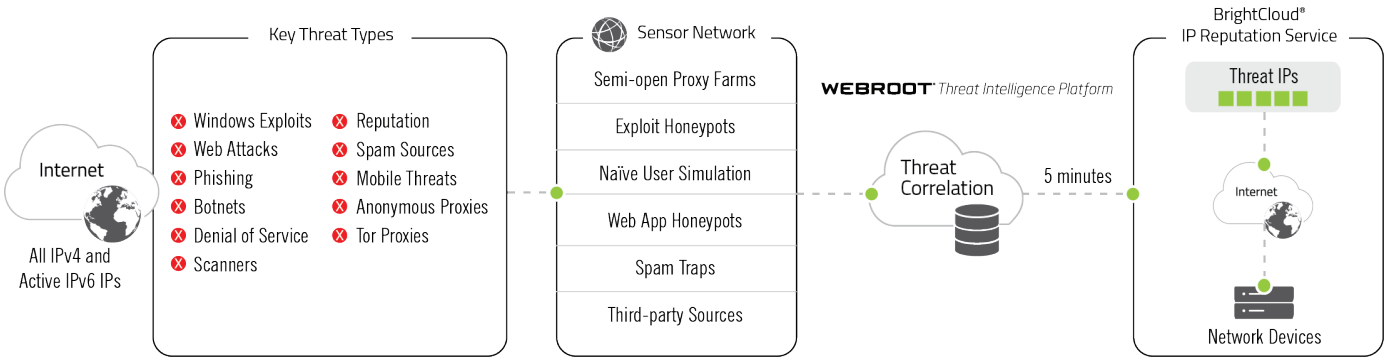
Due to the large scale of our machine learning and high level of automation, Webroot routinely runs up to 1,000 instances of these training models simultaneously. This order of magnitude and massive feature space is ideal for the effective characterization of brand new, zero-day objects to determine if they are threats. It also enables Webroot to provide fast and accurate notifications of existing and known threats.

Breadth and Depth of Data Sources

Supplying a substantial data set for the model is crucial for the machine to learn, adapt, and produce the desired output. To create an effective system of learning, the data set must be large enough for the algorithms to operate on to detect patterns. Pattern recognition is what helps make predictions and perform statistical analysis on the data set as a whole.

A machine learning system also needs access to broad and varied data sources for effective analysis. Webroot starts with sophisticated internet crawlers that catalog all URLs, IP addresses, files, and mobile applications. Thanks to the scalability of cloud infrastructures, our crawlers are now able to catalog the entire IPv4 space in a matter of minutes. Additionally, we also gather data by using large passive internet sensor networks, called "honeypots", that attract malicious connections such as exploitable spam relays.

Webroot also utilizes active scanning. We're able to send pings on different levels within the IPv4 space to get IPs to respond, adding incremental data on potential threats. We also incorporate a variety of third party lists from financial networks, partner networks, the Defense Information Systems Agency (DISA), the FBI, and various malware lists. While these lists are often peppered with false positives, we ensure the data is vetted and scrubbed prior to use. We also analyze actual web security traffic and



BrightCloud® IP Reputation Service

examine where people actually go on the internet, while other vendors offer little in the way of real-world user experience and therefore are not painting a complete picture.

However, the richest and most highly differentiated source of input for Webroot's machine learning-based security is our real-world endpoint and web sensor data. The endpoint data is two-fold; first we incorporate data from millions of endpoints around the world protected by Webroot SecureAnywhere® endpoint protection. The SecureAnywhere product family protects large and mid-sized businesses, as well as home and home office deployments. Then we add the data from the installed devices of our global technology partners. This enables us to incorporate real-world data from millions of endpoint sensors in near real time. This process provides our machine learning systems with an invaluable source of input and is yet another reason that Webroot technology can effectively identify never-before-seen and zero-day threats as soon as they are observed anywhere within our customer base, anywhere in the world.

Neural Networks and Complex Functions

Webroot applies extremely large and complex deep neural nets with 40 million nodes for its machine learning models. They are used to digest and analyze the massive number of characteristics we capture for each object. Neural nets represent a computational approach that is based on the way the human brain solves problems with large clusters of neurons connected by axons. Each node is connected with many others and these links can have varying impact on the activation state of the connected nodes. I.e. the nodes can be interpreted as a simple model of a neuron. The key to neural nets is that they are not explicitly programmed; they are self-learning, trained, and excel in areas such as cybersecurity where the solution or feature detection is difficult to express in a traditional software program.

While training the model, the machine learning selects and fine tunes the model parameters (i.e. its weights) thus determining the mapping from input vector to determination (in the simplest instance of benign or malicious file). When we allow the machine to establish the weights, we're essentially creating complex functions. These functions are exceptional for use as the activation function of artificial neurons in a neural net, or in explaining other natural processes such as those of complex system learning curves.

At a minimum, training and refining our models typically relies on millions of data points (a data point is a specific instance of an internet object). By using millions of data points, and adjusting the models iteratively, we can optimize our results to maximize discovery of malicious objects and minimize the false positive rate. For example, the Webroot BrightCloud® Real-Time Anti-Phishing Service identifies well over 98% of phishing sites in real time, as users request to go to a specific page or URL, with a false positive rate of just 0.001.

This combination of learning and mapping applies well to new threats. The large capacity of the model, in conjunction with the iterative training given new input, mean that our machine learning technology is very well suited to detect brand-new zero-day threats at the moment they appear. Because we capture such a large number of characteristics pertaining to an object, Webroot technology achieves this without having to change the algorithm or how we process input data.

Processing Power

Unlike competing intelligence providers, Webroot makes no assumptions or predeterminations as to which characteristics will be the most important in identifying an object. Every data point is considered thoroughly, and the machine determines how to weight the individual characteristics. In our training models, the machine can assign up to 400 million weights to the input vectors. The large number of characteristics that define each object, plus the weightings of each characteristic, comprise the massive feature space. This weight matrix must be kept in memory, which drives the enormous demand for computing resources.

The magnitude of our machine learning training models doesn't lend itself well to a distributed computing model. All of the input data and weights must be kept in memory, but not distributed over many nodes. Ten years ago, we used a 32-bit architecture with 4 GBs of RAM, which was sufficient for one million data points and one million weights. At the time, this technology was state of the art. Unfortunately, some threat intelligence competitors continue to operate at this level today.

Currently, training a Webroot model utilizes approximately 10 million data points (10 million instances of input vectors) to determine 400 million model parameters. For efficiency and speed, the model parameters are kept in memory while training the model. To accomplish this, we leverage Amazon

Web Services and the San Diego Supercomputer Center at the University of California, San Diego in La Jolla, CA. Our smaller training models use specially designed computer systems with 400–500 GB of RAM using multicore machines (around 20 nodes) for parallelization. Our larger training models will typically leverage instances with up to one terabyte of RAM and 64 nodes. A terabyte is equal to one trillion (10^{12}) bytes, which is a huge amount of memory. The smaller training models will run for 2–4 hours, while the larger models must complete their processing in less than 24 hours. Webroot acquires new information, runs a training model, and publishes the new models every day that incorporate this new knowledge. We improve and publish models daily, repeating the process for files, URLs, IPs, phishing sites, mobile apps, etc.

At run time, Webroot BrightCloud® Threat Intelligence Services, which have been integrated into our technology partners' devices, ask the resulting production model whether an object is malicious or not. This puts all of the heavy lifting in the cloud, keeping the local security appliances or other services efficient and effective.

Data Quality

The large amount of training data, coupled with the machine learning algorithms and computational power, makes it virtually impossible for threats to hide. Malicious files, phishing sites, etc. are built to avoid detection. The large input space is critical, however, in optimizing the model and thwarting malware or a phishing site's ability to hide. Webroot technology can detect the malicious code and enter it immediately in our training models, at which point all Webroot-secured systems and users then become protected against that threat.

Although machine learning does take on the more repetitive, tedious tasks that an organization's information security team doesn't have the time or resources to process, human analyst involvement must be highly leveraged to achieve a commensurate level of accuracy. At Webroot, we rely on a team of threat researchers who are closely and actively involved in the automated classification process that improves the model. It's a symbiotic relationship; humans train the machine to be more accurate, while the machines improve upon human classification effectiveness through speed and scalability.

The large number of characteristics we collect, in conjunction with the large scale of our models, ensures that any information pertaining to malware on any of our endpoints, in any location of the world will be incorporated into our training models. The same information that is important in improving our training models is also used at run time in the cloud to protect our users.

Real-World Results

The massive scale and order of magnitude of Webroot's machine learning pays practical dividends on a daily basis. For example, in the file space alone, Webroot witnessed 293 million unique instances of new files between January and October, 2016.

In this same timeframe, Webroot classified an average of 736,000 new files per month. 93% of the malicious files first seen in this period were witnessed on strictly one personal computer within Webroot's user base. This amounts to over 6 million unique, device-specific malicious files.

Webroot also classified 59% of malicious files within the first hour of first being seen, and the vast majority of them were in real time. The figure increases to 91% of zero-day malicious file variants after just 12 hours, again demonstrating the effectiveness of our machine learning models and scale to detect polymorphic variants. According to a third-party study conducted by MRG Effitas, Webroot achieved 100% detection within 24 hours.[1]

Summary

Today, machine learning technology has emerged as a critical component across all of the various domains of cybersecurity. Machine learning-based threat intelligence, such as Webroot's, can power numerous traditional security appliances, including next-generation firewalls, SIEM, UTM, IDS, IPS, and more. It enables threat activity across multiple security domains to be contextually associated to ensure optimal protection for users and organizations.

Webroot has been applying the latest machine learning technology to its advanced cybersecurity solutions for over ten years. This time in market and field experience, combined with our data scientists, statisticians, and the most advanced machine learning in the security industry, provides Webroot with a distinct advantage in combatting both known and unknown threats. Using fifth generation machine learning, Webroot systems can manage multiple and disparate data flows from sources such as endpoint security, next-generation firewalls, network behavioral analytics, and external IP, URL, and domain threat intelligence, among others. Additionally, big data constructs, massive scale, and the order of magnitude necessary to process massive machine learning training models enable Webroot's machine learning to operate in near real time.

Webroot also relies on machine learning to power its real-time endpoint, mobile, and network security offerings, the Webroot SecureAnywhere product family, along with its BrightCloud Threat Intelligence Services, and FlowScape® Network Anomaly Detection.

About Webroot

Webroot was the first to harness the cloud and artificial intelligence to protect businesses and individuals against cyber threats. We provide the number one security solution for managed service providers and small businesses, who rely on Webroot for endpoint protection, network protection, and security awareness training. Webroot BrightCloud® Threat Intelligence Services are used by market leading companies like Cisco, F5 Networks, Citrix, Aruba, Palo Alto Networks, A10 Networks, and more. Leveraging the power of machine learning to protect millions of businesses and individuals, Webroot secures the connected world. Headquartered in Colorado, Webroot operates globally across North America, Europe, and Asia. Discover Smarter Cybersecurity® solutions at webroot.com.

World Headquarters
385 Interlocken Crescent
Suite 800
Broomfield, Colorado 80021 USA
+1 800 772 9383

Webroot EMEA
6th floor, Block A
1 George's Quay Plaza
George's Quay, Dublin 2, Ireland
+44 (0) 870 1417 070

Webroot APAC
Suite 1402, Level 14, Tower A
821 Pacific Highway
Chatswood, NSW 2067, Australia
+61 (0) 2 8071 1900

[1] MRG Effitas. www.mrg-effitas.com/recent-projects/comissioned-tests (2016)