symantec.

Neural Network-based Antispam Heuristics

by Chris Miller

Group Product Manager
Enterprise Email Security

INSIDE

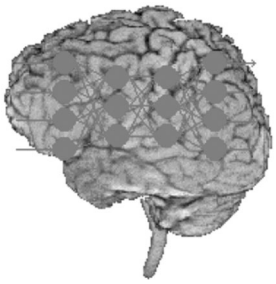
INSIDE

Contents

Introduction

The purpose of this paper is to explain how and why Neural Network technology is used for detecting spam.

What are neural networks?



By definition, a "neural network" is a collection of interconnected nodes or neurons. The best-known example of one is the human brain, the most complex and sophisticated neural network. Thanks to this cranial-based neural network, we are able to make very rapid and reliable decisions in fractions of a second.

In spite of the apparent simplicity of regular thought, decisions are generally not black-and-white or "binary", but rather involve a wide variety of conscious and subconscious inputs, each reacting in tandem with the other. As we grow and learn about the world around us, we have an incredible ability to recognize familiar patterns as well as anomalous patterns almost instantaneously, without a great deal of active thought. This can be seen in the following example:

When we see leopard fur, we don't simply run for our lives, even though all man-eating leopards wear it. If we did, this would be a very "binary" and probably inappropriate response. Other information inputs like - you're standing in central Manhattan, you're in a department store, your wife is standing beside you and she's wearing a fur coat – are other variables that have to be taken in context in making a decision not to flee. If the additional inputs were - you're alone, you're standing in the jungle, the thing wearing the leopard fur is growling and snarling, there are no tour group members between you and the leopard - you should of course react very differently.

Although this example is exaggerated, the key is that seldom does any single piece of information determine the outcome or decision. However, in making a decision, you or your brain don't put together logical arguments on the fly, but you rely subconsciously on your brain to quickly recognize patterns and draw conclusions. In the example above, the brain automatically drew the connections between relevant "neurons" of information and made a "fight (the crowds) or flight" decision on your behalf.

Although an artificial approximation, computer-based neural network technology attempts to mimic the human brain. Hence, we typically refer to this computer-simulated way of thinking as "artificial neural network" (or A.N.N., for short).

Why neural networks for spam

Spam presents a unique challenge for traditional filtering technologies: both in terms of the sheer number of messages (millions of messages daily) and in the breadth of content (from pornographic to products and services, to finance). Add to that the fact that today's economic fabric depends on email communication – which is equally broad and plentiful and whose subject matter contextually overlaps with that of many spam messages – and you've got a serious challenge.

The basic principal used in any spam filtering technique, whether heuristic or keyword-based, is identical: spam messages generally look different than good messages and detecting these differences is a good way to identify and stop spam. The difference between these technologies really comes down to the problem of distinguishing between these two classes of email. The neural networks approach is more refined, more mathematical and potentially far more accurate and reliable in accomplishing this task.

Put simply, the neural network approach attempts to mimic the way humans visually recognize spam from non-spam mail. Without being exposed to every spam message created, the average user quickly learns to recognize spam from legitimate communications, even from other solicited bulk communications like newsletters. The reason for this is generally because we expose our brains both consciously and subconsciously to a wide variety of message content, both good and bad, on a daily basis, and the brain learns to make lightning-fast, highly accurate guesses as to what spam is and what it is not.

How it works

Since neural networks is based on pattern recognition, the underlying premise is that each message can be quantified according to a pattern. This is represented below in Figure 1. Each plot on the graph (also known as a "vector") represents an email message. Although this 2-D example is an over-simplification, it helps to visualize the principle used behind neural networks.

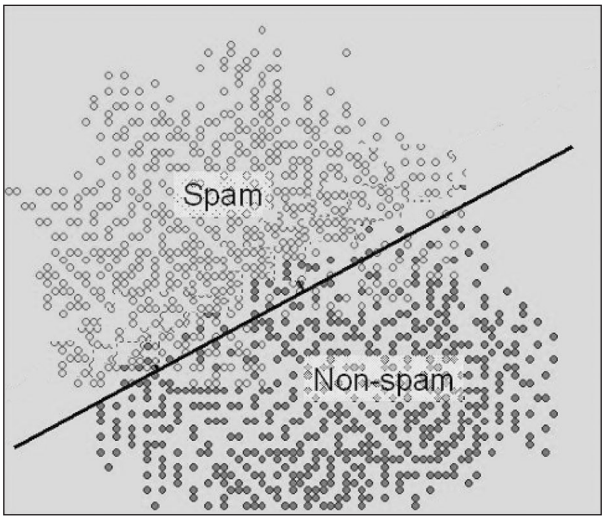


Figure 1: Distinctive patterns of good and spam messages cluster into relatively distinct groups.

To identify these patterns, the neural network must first be "trained". This training involves a computational analysis of message content using large representative samples of both spam and non-spam messages. Essentially the network will "learn" to recognize what we humans mean by "spam" and "non-spam". To aid in this process, we first need to have a clear, concise definition of "spam":

> *Spam, n., email sent in bulk where there is no direct agreement in place between the recipient and the sender to receive email solicitation.*

U.B.E. (Unsolicited Bulk Email) is another acronym for spam that effectively encapsulates this definition.

To create training sets of spam and non-spam emails, each email is carefully reviewed according to this simple, yet restrictive definition of spam. Although the average user often considers all unwanted emails as "spam", emails that border on "solicited" (it was likely requested at some point by the user) should be rejected outright. Examples of these might include email sent from easily recognizable domains, such as Amazon.com or Yahoo.com. A good motto to follow here is: "when in doubt, throw it out". Similarly, non-spam email should be restricted to personal email communications between individuals or groups, and avoid any forms of bulk mailings, regardless of whether they were solicited or not. Once these sets have been gathered and approved, the neural network is ready for training.

Using statistical methodologies, all words that are unique to the respective classes of spam and non-spam email are identified computationally. For example, words such as "free", "best" and "deal", are more closely associated with spam mail, and will therefore be considered significant indicators of spam. Similarly, words like "meeting" or "review" would be more closely tied to regular business email, and could be considered significant indicators of "good" email. Common words, for example, articles and conjunctions like "the", "and" etc., which are typically found in both classes of email, are simply ignored. Thousands of these significantly relevant words are then selected for the next step in the training.

The ANN system now preprocesses each email in the respective training sets to determine exactly which of these relevant words are found in each spam email, and which of these words are found in the non-spam email. Next, the ANN is trained to recognize certain combinations or patterns of interesting or relevant words to identify spam, or if it sees other combinations, to identify non-spam. The artificial neural network uses a set of sophisticated mathematical equations to perform this type of computation.

As some spam and non-spam messages will often "share" characteristics, a clear distinction cannot always be made. This is represented in Figure 1 by the "non-spam" plots or vectors that find themselves in the "spam" cluster and vice versa. In this "grey area" lies the potential for false positives.

After the training is complete, the ANN can now be used to scan live-stream email. Each message is scanned to identify relevant words, which are then processed by the ANN. If the ANN again sees certain types of combinations of word usage indicating a probability of spam, it will report spam, along with a probability value. Following the example in Figure 1, if the vector or plot computed for the message landed above the dividing line, it would be considered "spam". Its probability or confidence level would depend on the relative distance away from the line.

To maximize detections while avoiding false positives, a well-designed heuristics engine will accommodate different sensitivity thresholds, or levels of aggressiveness, in identifying spam. What this means is that the cut-off or dividing point between spam and non-spam can be adjusted so that the likelihood of a false positive match will be greatly reduced. This can be seen in Figure 2 below.

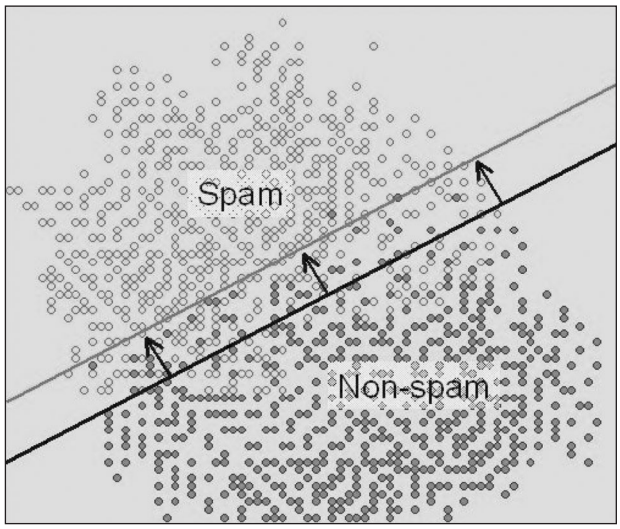


Figure 2: The sensitivity threshold can be adjusted to avoid the "grey" area.

In other words, the further away from the central dividing line between ham and spam email clusters, the lower the chance of false positive detections. Note in Figure 2 that there are far fewer non-spam vectors or patterns above the new cut-off or dividing line.

However, the trade-off for greater sensitivity to false positives is a corresponding reduction in detections. By moving the line away from the original cut-off point, you will no longer catch the spam that now resides below the line (see Figure 2). Essentially, more spam will be let through, but fewer good mails will be tagged in error. This "trade-off" is true of all heuristic approaches where the prevention of the greater evil takes precedence.

Conclusion

Although no single technology can achieve one hundred percent spam detection with zero false positives (despite vendor claims), machine-learned heuristics in general and neural networks in particular have proven extremely effective and reliable at accurately identifying spam and minimizing errors to an acceptable minimum.

About the Author

Chris Miller is group product manager at Symantec, responsible for the strategic direction of its Enterprise Email Security solutions, encompassing virus protection, spam prevention and content filtering to address the security needs of corporate messaging environments. Miller has 10 years of product management experience in enterprise software development

SYMANTEC, THE WORLD LEADER IN INTERNET SECURITY TECHNOLOGY, PROVIDES A BROAD RANGE OF CONTENT AND NETWORK SECURITY SOFTWARE AND APPLIANCE SOLUTIONS TO INDIVIDUALS, ENTERPRISES AND SERVICE PROVIDERS. THE COMPANY IS A LEADING PROVIDER OF VIRUS PROTECTION, FIREWALL AND VIRTUAL PRIVATE NETWORK, VULNERABILITY ASSESSMENT, INTRUSION PREVENTION, INTERNET CONTENT AND EMAIL FILTERING, AND REMOTE MANAGEMENT TECHNOLOGIES AND SECURITY SERVICES TO ENTERPRISES AND SERVICE PROVIDERS AROUND THE WORLD. SYMANTEC'S NORTON BRAND OF CONSUMER SECURITY PRODUCTS IS A LEADER IN WORLDWIDE RETAIL SALES AND INDUSTRY AWARDS. HEADQUARTERED IN CUPERTINO, CALIF., SYMANTEC HAS WORLDWIDE OPERATIONS IN 38 COUNTRIES

FOR MORE INFORMATION, PLEASE VISIT WWW.SYMANTEC.COM

WORLD HEADQUARTERS

20330 Stevens Creek Blvd.
Cupertino, CA 95014 U.S.A.
408.517.8000
800.721.3934

www.symantec.com

For Product Information
In the U.S., call toll-free
800.745.6054

Symantec has worldwide operations in 38 countries. For specific country offices and contact numbers please visit our Web site.