Forcepoint
DLP

Machine Learning

Contents

Introduction to Forcepoint DLP Machine Learning

Machine learning is a branch of artificial intelligence, comprising algorithms and techniques that allow computers to learn from examples instead of predefined rules.

Administrators can provide examples that train the Forcepoint DLP machine learning system to help protect sensitive, proprietary, and confidential information. After training, the system creates a classifier to identify documents based on how similar they are to the positive examples provided during the learning process.

There are two main types of machine learning algorithms:

- Supervised learning algorithms

  The algorithms are given labeled examples for the various types of data that need to be learned.

- Unsupervised learning algorithms

  Data is unlabeled and the algorithms attempt to find patterns within the data or to cluster the data into groups or sets.

Forcepoint DLP machine learning uses both types of algorithms.

This article offers a general introduction to Forcepoint DLP machine learning and explores the types of data that can be effectively protected using machine learning. See:

- *Knowing when to use machine learning*
- *How Forcepoint DLP machine learning works*
- *Selecting examples for training*
- *What happens during training*
- *Accuracy of machine learning*
- *Using the classifier*
- *Tuning the classifiers*
- *Comparison with other types of classifiers*

Knowing when to use machine learning

Machine learning offers advantages and disadvantages compared with other Forcepoint DLP classification methods. It is important to assess whether machine learning is the best solution for a particular deployment.

Like any other decision systems that handle complicated data, Forcepoint DLP machine learning may generate false positives (unintended matches) and false negatives (undetected matches). The total fraction of false positives and false negatives is sometimes referred to as the accuracy of the system.

Accuracy of machine learning is derived from the properties of the data, and finding the best data sets can sometimes be challenging. Because of this, before considering machine learning, administrators may want to determine if other types of classifiers, such as fingerprinting or pre-defined policies, are sufficient to classify and protect their data.

An example of when machine learning could be most effective is in differentiating between proprietary and non-proprietary data found in source code. It can be hard to fingerprint source code that is under constant development and continually changing, and predefined policies cannot distinguish between proprietary and non-proprietary source code.

Forcepoint DLP provides several predefined content types that address common use cases, including source code (in C, C++, Java, Perl, and F#), patents, software design documents, and documents related to financial investments. To protect content that belongs to these content types, consider using machine learning, and ensure that you select the appropriate predefined content type.

Machine learning can also be used to complement and enhance fingerprinting and predefined policies and other Forcepoint DLP detection and classification methods.

How Forcepoint DLP machine learning works

Supervised machine learning for data protection requires, in general, two types of examples:

- Content that needs to be protected ("positive" examples)
- Counterexamples ("negative" examples)

  Counterexamples are documents that are thematically related to the positive set, yet are not meant to be protected. Examples might be public patents versus drafts of patent applications, or non-proprietary source code versus proprietary source code.

Because it can be difficult and quite labor-intensive to find a sufficient number of documents for the negative set (while ensuring that no positive examples are in the set), Forcepoint has developed methods that allow the system to use a generic ensemble of documents as counterexamples. (See *Negative examples consisting of "All documents"* and *Positive examples*.)

For text-based data, some of the algorithms automatically create an optimal "weighted dictionary" that assigns positive weights to terms and phrases that are more likely to be included in the positive set and negative weights to terms and phrases that are more likely to be included in the negative set. The algorithms also find an optimal threshold. When the weighted sum of the terms that are found in a given document is greater than that threshold, the algorithm decides that the document belongs to the positive set. The assumption is that positive examples are more likely to have common themes.

Most machine learning algorithms are designed to be used with several hundred or several thousand positive and negative examples and require "clean" data, or data that is correctly labeled. Forcepoint DLP machine learning, however, utilizes different algorithms for different data sizes and attempts to automatically match the type of algorithm to the size of the data.

In addition, Forcepoint DLP machine learning algorithms can detect "outliers" among a set of positive examples. These are examples that should probably not be labeled "positive." Forcepoint algorithms also allow learning to take place even when negative examples are not provided.

Related concepts
Positive examples on page 4
Negative examples consisting of "All documents" on page 5

Selecting examples for training

Which examples you are selecting is important for machine learning training.

Positive examples

For effective machine learning to occur, it is most important to select the best positive examples.

- These are textual examples of the data to protect.
- The documents in this set should be related to the same theme or share other commonalities.

  Without the commonalities, the learning algorithm will not be able to find a way to categorize the data.

The required number of examples depends on the level of commonality. If the positive examples share many common terms that are very rare, in general, a small number suffices. On the other hand, if the differences between the positive and the negative set are more subtle, more examples will be required. A positive set typically consists of 100–200 text documents.

Negative examples

Negative examples are samples of data that are semantically or thematically similar to the set of positive samples, but that should not be protected.

The size of this set of negative examples can be similar to the size of the positive set, although a larger set is preferable.

Negative examples consisting of "All documents"

To create a generic ensemble of documents that Forcepoint DLP machine learning can use as negative examples, select the path to a large folder with a representative sample of documents from the organization. This folder can contain both positive and negative examples, but substantially more negative examples should exist.
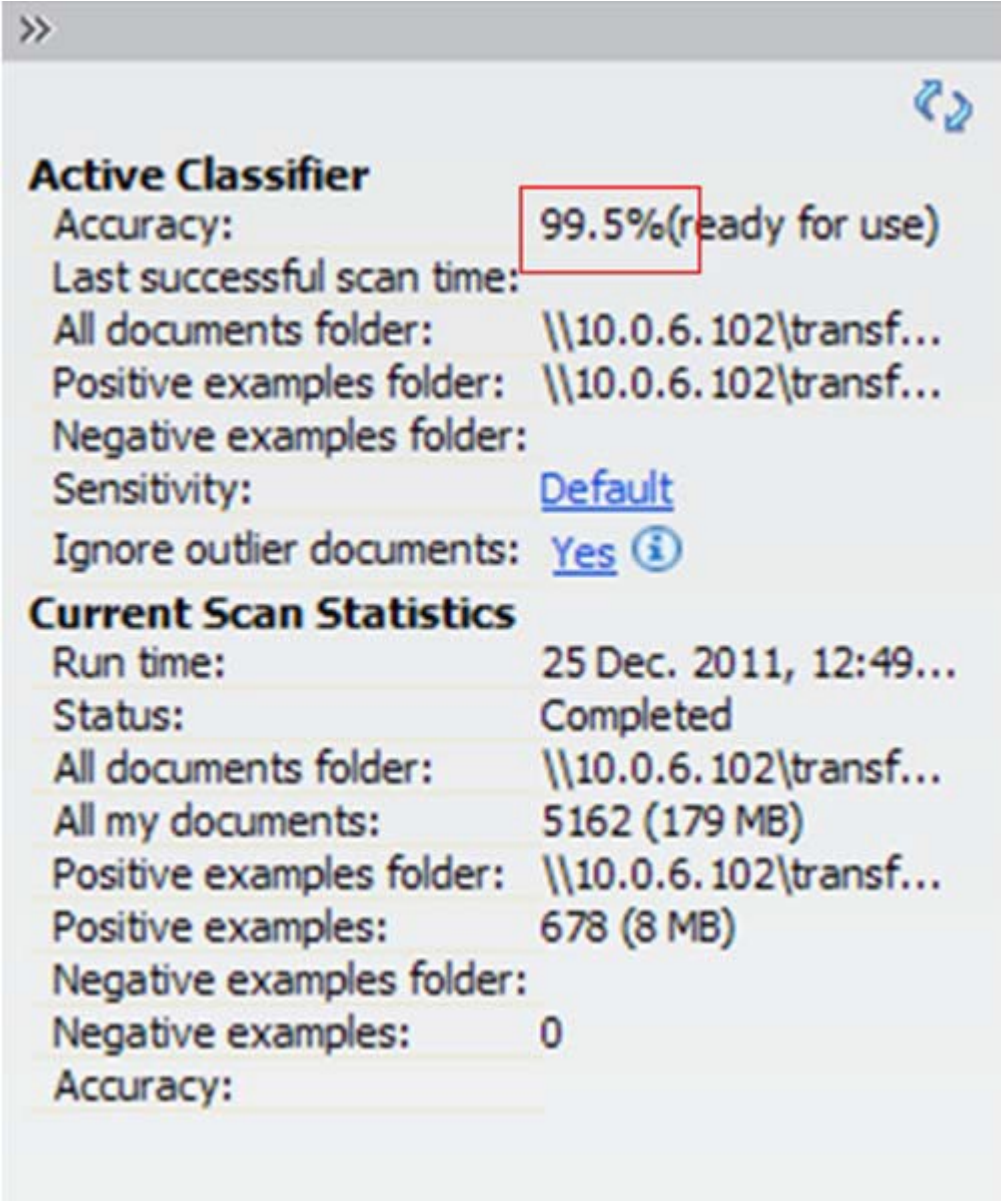
The size of this set of counterexamples can be similar to the size of the positive set, although a larger set is recommended.

What happens during training

After the examples are submitted, the crawler starts examining the files and providing them to the learning algorithms. If the number of files in a folder is very large, a sampling algorithm samples the folder several times and checks for convergence.
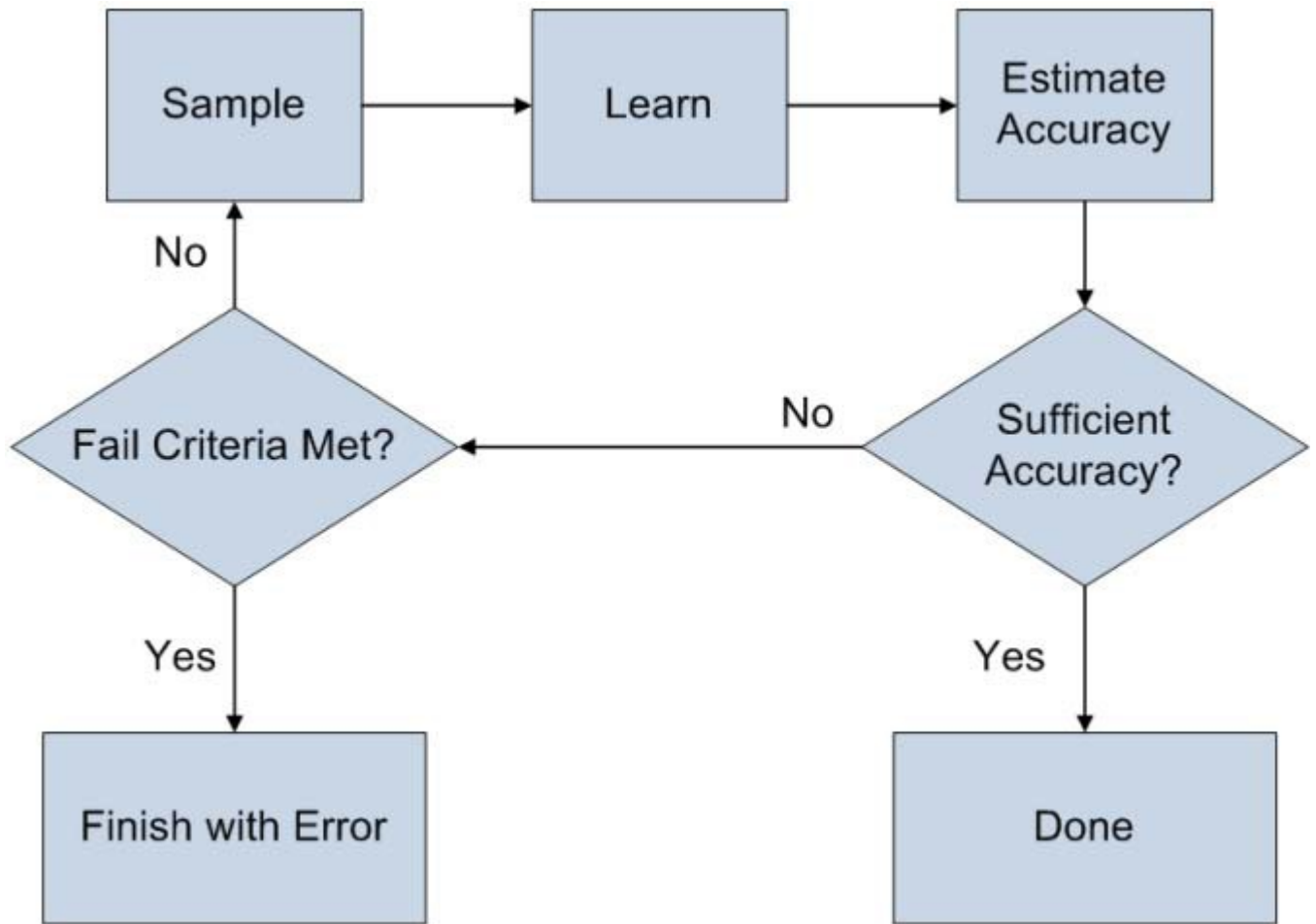
If learning is successful (meaning that the data is "learnable"), the following window appears:
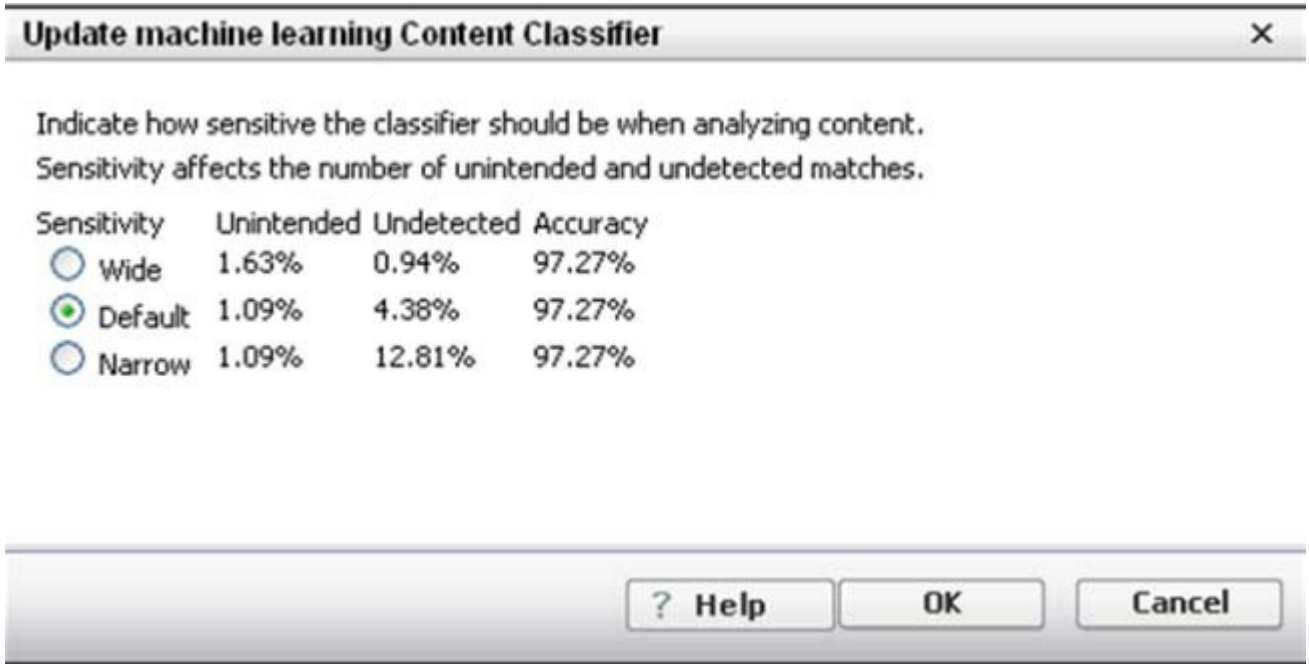
Active Classifier

Accuracy: 99.5%(ready for use)
Last successful scan time:
All documents folder: \\10.0.6.102\transf...
Positive examples folder: \\10.0.6.102\transf...
Negative examples folder:
Sensitivity: Default
Ignore outlier documents: Yes ⓘ

Current Scan Statistics

Run time: 25 Dec. 2011, 12:49...
Status: Completed
All documents folder: \\10.0.6.102\transf...
All my documents: 5162 (179 MB)
Positive examples folder: \\10.0.6.102\transf...
Positive examples: 678 (8 MB)
Negative examples folder:
Negative examples: 0
Accuracy:

By default:

- The sensitivity level is set to "Default," an optimal trade-off between false positives (unintended matches) and false negatives (undetected matches). To change the sensitivity level, click **Default**, which opens the Update machine learning Content Classifier window:

Update machine learning Content Classifier   ✕

Indicate how sensitive the classifier should be when analyzing content.
Sensitivity affects the number of unintended and undetected matches.

Sensitivity	Unintended	Undetected	Accuracy
○ Wide	1.63%	0.94%	97.27%
⦿ Default	1.09%	4.38%	97.27%
○ Narrow	1.09%	12.81%	97.27%

? Help   OK   Cancel

It is important to consider the percentage of unintended and undetected matches before changing the sensitivity level. For example, selecting "Narrow" increases the expected number of undetected matches without reducing the expected number of unintended matches. It is, therefore, highly undesirable.

■ The training is performed ignoring outliers, or examples that could be labeled "positive," but that do not seem to belong to the positive set.

To avoid ignoring outliers, administrators can click **Yes** next to "Ignore outlier documents" and change it to **No**.

Accuracy of machine learning

The ability of the system to accurately classify data depends to a large extent on the examples provided. If Forcepoint DLP machine learning fails to find enough common elements, its results may not be accurate. Should this happen, the system performs another stage of validation to assess the level of false positives (unintended matches) and false negatives (undetected matches) on new data that is not used during the training phase, sometimes referred to as "zero-day documents."

If the "recall" level of the classifier (the total number of true positives divided by the sum of false positives and false negatives in the new data) is below 70 percent, the system returns a FAIL message that includes the likely reason the attempt to accurately classify data failed.

Error messages include:

Error Code	Error Message
DSCV_ERR_-420_CODE	There are not enough examples in your positive examples folder. X were provided and at least Y are required. Please add more examples then restart the machine learning process.
DSCV_ERR_-421_CODE	There are not enough examples in your negative examples folder. X were provided and at least Y are required. Please add more examples then restart the machine learning process.

DSCV_ERR_-422_CODE	The files in your positive examples folder do not contain enough text. Of X files provided, only Y have enough text. At least Z are required. Please update the files or point to another folder, then restart the machine learning process.
DSCV_ERR_-423_CODE	The files in your negative examples folder do not contain enough text. Of X files provided, only Y have enough text. At least Z are required. Please update the files or point to another folder, then restart the machine learning process.
DSCV_ERR_-424_CODE	Your positive and negative examples are too similar. No significant difference in words distribution was found. Please provide new examples.
DSCV_ERR_-425_CODE	Your positive and negative examples are too similar, or your positive examples may not be consistent enough to draw conclusions. There were bad error rates on both training X and validation Y. Use different example folders in the classifier.
DSCV_ERR_-426_CODE	The examples you provided were not sufficient for accurate training. Though the accuracy of the training set is good X, the machine learning process cannot make accurate conclusions on unseen data X. Your positive examples may not be homogeneous enough. Please provide more consistent examples then restart the machine learning process.
DSCV_ERR_-427_CODE	Your examples do not fit the content type you specified. You provided X positive examples, but only {2} of them fit the type.
DSCV_ERR_-428_CODE	The files in your example folders don't contain enough meaningful text (only X words). Please add files with more meaningful content or point to other folders, then restart the machine learning process.
DSCV_ERR_-429_CODE	More than one file in your examples folders doesn't contain enough text (only X words). Please update the files or point to other folders, then restart the machine learning process.

By adjusting the sensitivity level of the classifier, administrators can reduce the number of false negatives (unintended matches) while accepting a higher level of false positives (undetected matches) or accept some false negatives to reduce the rate of false positives (or find an acceptable balance in between).

Factors influencing the choice include:

- The level of commonality in the positive set of examples (a low level tends to decrease accuracy)
- The business implications of false positives
- The resources that available to deal with false positives

Using the classifier

After successful training, the machine learning classifier can be used to create rules and policies. An incident that resulted from a match with a classifier might look like this:

Tuning the classifiers

In some cases, administrators may want to tune the classifiers. For example, if too many false positives occur, start by setting the sensitivity level to "Narrow."

It is also possible to combine the classifier with other classifiers, such as looking at certain file types, like both Microsoft Office files and PDF files.

If the overall accuracy level is too low, check to see if all of the positive examples are related to the same subject. If there is a small number of subjects and enough samples for each of them, optionally create a different classifier for each subject:

Steps

1)   Assign a folder to each subject.

2)   Place documents related to the subject in the corresponding folder.

3)   Train the system separately on each folder.

Next steps

In many cases, several small specific classifiers can provide better accuracy than one general classifier.

Comparison with other types of classifiers

The following table summarizes the advantages and disadvantages of the various classifier types:

	Machine Learning	**Fingerprint- ing**	**Pre-Defined Policies**	**User-Defined Dictionaries and Regular Expressions**
Coverage	High: Covers any document with semantic similarities to the learned data	Medium: Detects only derivatives of fingerprinted documents	Limited to the existing pre- defined types	Unlimited, providing that the user has properly defined the dictionaries and the regular expressions
Accuracy	Depends on the data	Very High	High for data types that are common enough	Medium
"Zero-Day" Protection	High	Very Low	High	High
Size/Footprint	Medium	High	Low	Low
Deployment and Config Effort	Medium (may require some tuning)	Medium	Low	High - requires careful setting and tuning

For more information on how to use machine learning, see:

- Forcepoint DLP Administrator Help
- Using Machine Learning for Optimal Data Loss Prevention