Cross-Validation of Filesystem Layers for Computer Forensics

By

Joe Sremack

Presented At

The Digital Forensic Research Conference

DFRWS 2003 USA Cleveland, OH (Aug 6th - 8th)

Cross-Validation of File System Layers for Computer Forensics

Joe Sremack

North Carolina State University

Overview

- What Cross-Validation(CV) is and why it is used
- File system layers and how they are cross-validated
- Current research results and the problems being faced
- The direction of future CV research

Motivation

- Major problem for investigators is complexity
- Working with abstract data is easier and faster than low-level data
- Focusing on abstract data sacrifices completeness of an investigation
- Need an effective means for balancing quickness with completeness

Cross-Validation

- Mathematical technique for evaluating the performance of data sets
- Process

1. A set of data is split into subsets
2. Most of the subsets are used to model the remaining subsets
3. An error ratio is formed by comparing the model to the actual remaining subsets

CV Example

i_1	k_1
i_2	k_2
i_3	k_3
i_4	k_4

- First Iteration

  1. Use i1, i2, and i3 to derive i4'

  2. Compare i4' to i4 and calculate the error ratio.

  3. Do the same for the second column
- Iterate three more times
- Take the average error ratio.

Abstract File System Layers

- File Systems have layers, abstract and low-level
- Focusing on abstract could overlook low-level evidence
- Need method for quick, easy, and complete analysis of file systems

File System CV Technique

- Start with file system image and obtain Layer 1
- Predict the Layer 2 using Layer 1
- Compare the model to Layer 2 and record the discrepancies
- Perform the same steps for each remaining layer
- Sum the results to arrive at the error ratio

FAT12

- Boot Sector: contains all of the information needed for the OS to call disk driver functions to access sectors, e.g. number of FATs
- File Allocation Table (FAT): a linked list structure that stores information about each cluster; it specifies where the next cluster is and if a cluster is unallocated, reserved, or bad.
- Root Directory: the topmost directory, e.g. c:\ or a:\
- Data Area: the location of every sector, i.e. every file and directory

FAT12 Layers

Layer	Input	Output
1	Raw file system image	Boot Sector values
2	File system image and Boot Sector values	FAT and Data Areas
3	FAT Area, FAT Entry size	FAT Entries
4	Data Area and Cluster size	Clusters
5	Raw cluster content and content type	Formatted cluster content
6	Starting cluster and FAT Entries	Linked list of clusters
7	List of clusters, clusters, formatted cluster content and type	All Directory Entries in a directory or raw content of a file

FAT12 Layer 7

A file, a.txt, has a filesize of 10K in its Directory Entry (Layer 7)

a.txt's starting cluster and FAT listing = 8K (Layer 6)

- Cross-validating Layers 6 and 7 will show the discrepancy.
- This discrepancy will be reflected in the cross-validation error ratio.

File System CV Assumption

- Know how to transform information from lower layers into higher layers
- Focusing on the file system, not application layer information
- Looking for anomalies, not syntactical correctness of the file system
- The more accurate the model, the less likely either layer has been tampered with

CV Formula

- Cross-validating file systems requires a formula specific to each file system
- The formula must be derived in advanced in an ad hoc manner
- It is a summation of every layer, with each being weighted
- Each layer is also formed by summing its information in a weighted manner
- $E_r = 0.05(L_1) + 0.1(L_2) + 0.05(L_3) + 0.1(L_4) + 0.2(L_5) + 0.2(L_6) + 0.3(L_7)$

File System CV

- Similarities between mathematical CV and file system CV

1. Subsets vs. File System Layers
2. Parallels of One-to-one and one-to-many mappings in mathematics and file systems
    - lossy vs. lossless layers
    - Benefits of losslessness

File System CV

- Differences Between Mathematical CV and File System CV

  1. Mathematical CV sets are very large -- number of file system layers are not

  2. Many iterations of CV for mathematics; small number of layers that can be cross-validated against each other

Implementation

- Requirements
  1. Access to each layer
  2. CV Formula
  3. Output of discrepancies
- Brian Carrier’s Sleuth Kit and a Perl wrapper were used
- Wrapper calls SK to gather information from each layer
- Information from every layer stored

Implementation Cont.

- Layer 1 data used to form a model of Layer 2
- Model is compared to Layer 2 (stored)
- Discrepancies and error ratio are stored
- Model of Layer 3 is formed and then compared to Layer 3
- Again, discrepancies and error ratio are stored
- …
- Model of Layer 7 is formed and then compared to Layer 7
- Discrepancies and error ratio are stored
- Stored layer error ratios are placed into the weighted CV formula to arrive at the file system image’s ER

Difficulties

- File system layers are lossless, but that does not make predicting perfect
- Suppose information at Layer 4 requires information from Layers 2 and 3
- A discrepancy was discovered between 2 and 3, so which information do you use?
- The predicted Layer 3 or the actual Layer 3?
- If Layer 2 were incorrect, then the prediction would be too, but if Layer 2 were correct, then Layer 3 would be correct

Difficulties Cont.

- The heuristic used is to do a truer cross-validation by using both the model and the actual value, resulting in separate error ratios
- This becomes messy if there are hundreds of discrepancies, in which case a heuristic of believing the lower layer can be used

Future Research

- Including Low-level (e.g. partition table) and application level layers for increased accuracy
- CV formulas for all modern file systems
- Methods for cross-validating multiple partitioned media, along with virtual systems

Conclusion

- Human analysis of CV still necessary
- CV is only a semi-automated process
- The results can sometimes be borderline ones requiring the investigator to analyze the discrepancies
- This method saves time and money, gives extra credibility to an investigation, and provides a means for staying on top of the ever-increasing complexity of computer systems

Thanks for Listening. Now, Questions?

I know you’ve got them!