Triage in-Lab: case backlog reduction with forensic digital profiling

Leopoldo Sebastián M. GÓMEZ

sebastian.gomez@jusneuquen.gov.ar

Abstract. Since it exist a huge backlog of cases and few digital forensic specialists in the Justice System, usually there is not possible to move them to contribute directly into the digital crime scene. On the other side, the law enforcement has a lack of skilled forensic staff available to perform forensic triage. Moreover, the reviews on the fly are taking significant time delays, under pressure, technical restrictions and time framed. At this point, when a suspect target system and data are found, it leads to be seized and moved to a dedicated forensic laboratory where the expert can perform the analysis of their content. Under some circumstances, all that may be required is to quickly and efficiently review a number of target systems to establish if they are likely to contain material of interest to an investigation. However, when the digital evidence comes to the specialist, he has a little knowledge of the previous stage, and it is difficult to make decisions about the priorities or activities on the sized devices. Such reviews are often referred to as "forensic triage" reviews and must be performed using forensically acceptable methods in order that any evidence that is identified during the forensic triage process is not damaged, modified or contaminated, literally or from a legal perspective, by the process of acquiring and reviewing the evidence. We have developed a novel triage tool, which tries to catch a criminal profile with an automated predictive classifier focused on child pornography and intellectual property theft. This software detects few critical attributes into the digital evidence and they are compared with other vectors of characteristics extracted from a digital data corpus based on devices of past cases. As a result of this automated process, a criminal profile prediction is done. This tool will assist to computer forensic experts, in order to make decisions about priorities to make full analysis of suspect devices or discard them with low probabilities of losing digital evidence. Our approach should be useful to mitigate the backlog of computer forensics laboratories.

Keywords. Triage, digital profiling, prioritization, case backlog reduction.

1. Introduction

The most commonly used strategies to collect data of potential evidential interest fall into two categories: a) use staff with limited forensic training and seize everything or b) skilled-expert with selective acquisition.

In the first approach, unskilled staff can potentially damage digital evidence during the process of collection and/or acquisition. However, these risks can be mitigated to an acceptable level via appropriate training in the use of suitable techniques, principles of evidence protection and deployment processes. Every item brought into the lab has to be examined, even if it's just so that it can be excluded from an investigation. Naturally, in serious cases this is the most appropriate course of action, but for the vast majority of cases this can lead to items of no relevance being passed for forensic examination which extends the case processing time and delays the eventual result.

The second one certainly has merit in serious crime or major incidents, but it removes valuable core skills from the forensic staff reducing caseload throughput and capacity to respond in the laboratory. In such circumstances, those attending may be required to make critical decisions about which items to select. Making such decisions in the absence of a scientific method for selection is both unsatisfactory and dangerous in that items containing potential evidence may be overlooked. Forensic analysts tend to seize items rather image on site because imaging on site is already taking extended times and with continued grown in device capacities, these times can only continue to grow.

Some researchers (Sheldon, A., 2005) believe that the idea of having one forensic expert for a case is no longer suitable. Digital forensics has moved on from when a single examiner's understanding was adequate to complete an investigation, and is suggested to digital forensics must unite to develop and reuse information.

As digital forensics becomes better known the number of digital device analysis requests will grow. This added with the problem of the increasing sizes of typical digital storage devices. It is not unusual for a forensic laboratory to have a 9 to 12 month backlog. One possible way to reduce this backlog is to use digital triage in case prioritization and intelligence gathering for use in the actual examination phase (Cantrell, G., Dampier, D., Dandass, Y., Niu, N. and Bogen, C., 2012).

Today's forensic tools were developed to find all the evidence, but we are looking for tools that work on time-constrained environments. There are some triage tools that help to non-technical operators like FieldSearch, DriveProphet and Spektor. They are provided with an automated system to rapidly extract data and analyze information in-field in real-time, but it requires human visualization and filter strategies. Academic efforts are leading to build new evidence collecting tools as Windows Forensic Analyzer (Dashora, K., Tomar, D., and Rana, J., 2010), but they only automatize the gathering process, and they appoint that in future work, to make effective investigation, knowledge discovery techniques may be applied to analyze captured evidence in Windows environment. A few commercial forensic tools, as ADF Triage Tools, are going to our approach in order to reduce the backlog at lab.

Digital profiling takes computer forensics to the next step, based upon the experiences gained from processing hundreds of drives. It should only be attempted by experienced examiners, preferably those who have had some education or training

in the behavioral sciences. It blends aspects of technology, investigation, psychology, and sociology to provide a larger picture than may be documented. It is important to remind the high value that this approach brings to the criminal investigation. By mean of a fast analysis, the forensic examiner will get the profile of a software pirate, for example, looking installed applications like P2P or CD/DVD burning software. If they are found, it can be considered as potential digital evidence in the crime. Furthermore, in cases involving child abuse, the criminal may communicate with the child using MSN Messenger or Skype (Alghafli, Jones and Martin, 2010). Obtaining digital evidence of this software can be useful to lead a deep inspection on criminal´s account name and contact list, in order to match the child user name and establish the relationship.

A triage approach technique (Grillo, Lentini and Ottoni, 2008) has appeared but it doesn't have an integrated forensic tool available. Instead the researchers have used WEKA to the classification process. Other people (Liu, Lin, and Guo, 2008) are focused to automate and speed up the process of locating potential evidence in the network forensics, using a forensic framework based a one-class SVM algorithm with a modified Gaussian (RBF) kernel as outlier detector. They have appointed that feature extraction and selection form the available data is important to the effectiveness of the methods employed because the great capability in selecting the suitable features of a classifier can lead directly to faster training and more accurate results. Usually the selection of what kinds of features depends on the target objects defined or constructed by the forensic investigator.

2. The digital forensic corpus

Researchers (Garfinkel, S., Farrell, P., Roussev, V. and Dinolt, G., 2009) have argued that the development of representative standardized corpora for digital forensic research is essential for the long term scientific health and legal standing of the field. It is a notable effort to perform controlled and repeatable experiments that produce reproducible results. If a standard for digital investigation is set, building a forensic corpus makes sense. There is available a real data corpus of a collection of raw data extracted from data-carrying devices that were purchased on the secondary market around the world. This public collection doesn't have any association with criminal cases, and we observed that we needed prior digital data devices labeled to be in agreement with our local crime taxonomy.

According to our methodology and technical resources in the laboratory we preserve temporarily the forensic images of complex cases. This selective archiving criterion allows in the near future to carry out new forensic tasks whenever it is needed, eliminating needs to re-acquisition of digital evidence. It has been a little improvement to reduce response times in critical cases, but also it has enabled us to maintain a real data corpus for research and training. Using this one is possible to build predictive models whereby pattern recognition and re-occurrences of similar crimes are identified.

Each investigation can produce a set of unique features regarding the specific crime. These are used to describe a criminal profile which can then be used to cross examine further digital devices for similarities in both investigations. It is understood that adapting current forensic methods and practices to incorporate profiling a means to automate forensics is a way to combat the demands placed on digital forensics.

3. Extraction phase

We tagged devices of relevant past cases to use them as the training set to test several classifier methods, in order to conduct future investigations applying predictive methods and triage techniques. From our experience, we were able to build a digital DNA of many devices which we worked on. We have seen that in the past years the most of the investigations were on Windows operative system in all kind of versions, so we focused our efforts to catch digital information from this one.

Our main goals were to select digital devices in leading cases of child pornography and intellectual property theft, which were labeled with tag called "Notable". These two categories are useful for triage techniques in lab because usually these crime cases have many devices to examine. Also, we included other digital devices to build a generic profile called "Other", which was used to discard deep analysis of the device in these investigations. We worked on 21 forensic images of hard disks to make the training dataset and we built a script in Perl to extract the digital DNA of these "leading cases".

3.1. Vector modelling

The vector structure of our digital DNA has values extracted of critical artifacts from the Windows registry, based on statistics and other metadata from the filesystem. Originally our primary structure had 22 characteristics, and it was:

X	Description
1	Had the user activated the "see hidden files" option?
2	How many suspect burning, compression, edition, copy or p2p software are installed?
3	What is the usual working time of the user?
4	What is the percentage of image files located into the filesystem?
5	What is the average size of image files located into the filesystem?
6	What is the percentage of audio files located into the filesystem?
7	What is the average size of audio files located into the filesystem?
8	What is the percentage of document files located into the filesystem?
9	What is the average size of document files located into the filesystem?
10	What is the percentage of video files located into the filesystem?
11	What is the average size of video files located into the filesystem?
12	How many applications are installed?
13	How many USB devices had the user connected?

14 Had the user activated the “hidden extension” option for known files?
15 Had the user activated the “see system files” option?
16 Had the “multiuser profile” created into the operative system?
17 Had the user communicated with Skype?
18 Had the user communicated with Messenger?
19 Had the user worked with MS-Word files recently?
20 Had the user worked with MS-Excel files recently?
21 Had the user worked with compressed files recently?
22 Had the user worked with image files recently?

3.2. Engineering the input

Successful data mining involves far more than selecting learning algorithm and running it over our data. Because there is no substitute for getting to know our data, we had to see which data were filled into the characteristics vector defined to detect outliers and other filtering tasks with the dataset.

We have considered a numeric attribute discretization to make improvements over experiments. We applied the entropy-based method with MDL (minimum description principle) stopping method, which is one of the best general techniques for supervised discretization. Dimensionality reduction yields a more compact, more easily interpretable representation of the target concept. In fact, we considered to prune some unsuitable attributes with automatic methods. After applying this data mining technique, we got a new smaller characteristics vector with sixteen attributes, some of them discretized. It could be redefined as <X1, X2, X3, DX4, DX6, DX7, DX13, X14, X15, X16, X17, X18, X19, X20, X21, X22>. Note that the prefix D involves a discretized variable.

4. Experimentation

We used an open source datamining toolkit to choose a machine learning algorithm for our triage software. Several algorithms like C4.5, Naive Bayes, kNN, SVM and Classification Tree were tested and compared. Further experiments could be conducted with different learning algorithms and paradigms to allow performance comparisons with the proposed approach.

Because we got limited data for training and testing, we used a 10-fold cross-validation technique. Extensive tests on numerous datasets, with different learning techniques have shown that 10 is about the right number of folds to get the best estimate of error, and there is also some theoretical evidence that backs this up. Although these arguments are by no means conclusive, 10-fold cross validation has become the standard method in practical terms.

In order to measure the performance of the proposed methods, the ROC curve is used. The ROC curve is a plot of detection accuracy against the false positive rate (Wilson, D. and Martinez, T., 1997). The acronym stands for Receiver Operating Characteristics, a term used in signal detection to characterize the tradeoff between hit

rate and false alarm rate over a noisy channel. ROC curves depict the performance of a classifier without regard to class distribution or error costs. It can be obtained by varying the detection threshold. They plot the number of positives included in the sample on the vertical axis, expressed as a percentage of the total number of positives, against the number of negatives, on the horizontal axis. True positive rate and false positive rate may be defined as follows: True Positive Rate = TP/(TP+FN), False Positive Rate = FP/(FP+TN), where TP denotes true positives, FN denotes false negatives, FP is false positives and TN is true negatives.

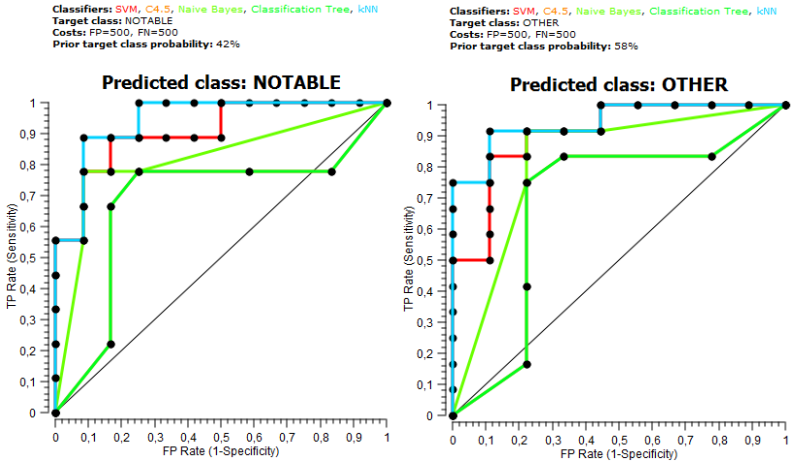

Fig. 1. ROC curve

Algorithm	**Classification Accuracy**	**Area Under ROC Curve**
C4.5	0.75	0.6806
Naive Bayes	0.85	0.8194
kNN	**0.9**	**0.9444**
SVM	0.8	0.9074
Classification Tree	0.75	0.6806

Table 1. Comparative metrics between classifiers

5. Classifier selection to implement the triage tool

Nearest neighbor instance-based learning is simple and often works very well. Let us assume that we have a training dataset D made up of (*xi*), i=1..n, training samples. The examples are described by a set of features F and any numeric features have been normalized to the range [0,1]. When we want to classify un unknown example *q*, for each *xi* in the D, we can calculate the distance between *q* and *xi* as follows:

$$d(\mathbf{q}, \mathbf{x}_i) = \sum_{f \in F} w_f \delta(\mathbf{q}_f, \mathbf{x}_{if}) \quad (1)$$

The *k* nearest neighbors are selected based on this distance metric. It will often make sense to assign more weight to the nearer neighbors in deciding the class of the query. A fairly technique to achieve this is distance weighted voting where the neighbors get to vote on the class of the query with votes weighted by the inverse of their distance to the query.

The general formula for the Minkowski distance is:

$$MD_p(\mathbf{q}, \mathbf{x}_i) = \left(\sum_{f \in F} |\mathbf{q}_f - \mathbf{x}_{if}|^p \right)^{\frac{1}{p}} \quad (2)$$

There are a large range of possibilities for this distance metric. When p=1, Minkowski distance is the Manhattan distance and the for p=2 Minkowsky distance is called the Euclidean distance. It is unusual but not unheard of to use p values greater than 2. Larger values of p have the effect of giving greater weight to the attributes on which the objects differ most. We have selected the Euclidean distance to implement our triage software tool.

Basic kNN classifiers that use a simple Minkowski distance will have a time behaviour that is O(|D||F|) where D is the training set and F is the set of features that describe the data, i.e. the distance metric is linear in the number of features and the comparison process increases linearly with the amount of data.

We considered kNN algorithm for our triage software tool because it is amongst the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, typically small). In binary (two class) classification problems, it is helpful to choose *k* to be an odd number as this avoids tied votes. Benchmarks showed us that the best algorithm were kNN (*k*=5).

6. Conclusion

Is possible to establish a basis of interaction between automation and profiling, this creating a stepping stone for early time-saving when a typical investigation is conducted. At present, computer forensics laboratories do not have the resources and time to keep up with the growing demand for digital forensic examinations with the current methodologies. One solution to this problem is the use of pre-examinations techniques commonly referred to as digital triage. These techniques can assist the examiner with intelligence that can be used to prioritize and lead the examination process.

We have developed a novel triage tool for specific crime cases of child pornography and intellectual property theft. Connecting this one with standard operative procedures (SOPs) we will get a fast device prioritization for in-depth analysis. We expect to reduce the lab backlog for these specific crime cases when they have many digital devices. Our triage tool is delivered as a standalone executable and if is considered necessary it could be translated as a smart script for usual commercial forensic software. Of course, we suggest changing mindset from "automated software" to "automated analysis".

University researchers are not aiming for automation because they do not have enough domain knowledge or a large corpora of forensically interesting data to develop reliable automated algorithms and tools. By the other side, it seems that only a few computer forensics examiners follow an academic approach by researching to the new frontiers of digital forensic science. It could involve reading and writing scientific papers, testing new forensic tools, coding specific scripts and looking for scientific challenges, not just the daily operative duties using familiar forensic tools.

There exists an underlying tension between digital forensic experts with an academic approach and merely technicians that are only interested in solving cases with push-button forensic applications. The first group adopts an investigative, predictive or comparative analysis, but the second one just completes their investigations at result level without deeper analysis. Despite this, forensic automation software is already becoming a problem by giving untrained examiners a false sense of security when in reality they are not conducting an examination at all.

Aspiring investigators and technicians with basic to moderate technical skills but lacking in deep forensics, are taking short tool-oriented courses, often a day or a week long that would to potential misinterpretations of complex digital forensic cases. Clearly push-button forensics will not become them in real digital forensic examiners but such automated tools have improved the response time on daily activities.

The real forensic circumstances show that we should take the best of both worlds. We agree to distribute the everyday workload among technical staff using SOPs and push-button forensics, enhancing their ability to contribute and decreasing the bottleneck on senior resources to free up them for tasks that truly require more experience and knowledge.

New computer forensics certification and degree programs will not solve the problem any time soon. Enforcing a grading level that allows only experienced investigators to conduct investigations at a specific level presents a degree of professionalism in the digital forensics discipline. This means that representation in court should only be allowed for those that are skilled enough. In the meantime, a wise combination of novel triage tools, joint teams of expert examiners and technicians, and standard operative procedures can contribute to mitigate the case backlog of many computer forensics laboratories.

6. Future work

It is feasible to expand the training set to include new crime classes or digital DNA examples. Furthermore, other forensic examiners can refine our vector model by

adding other digital characteristics, including some components that are suitable on a specific crime class. Some researchers have proposed a set of crime templates (Nance, K., Hay, B. and Bishop M., 2009). For example, is useful to apply flesh tone filtering to images during child pornography cases, as done by the first responder tool File Hound (Choudhury, A., Rogers, M., Gilliam, B. and Watson, K., 2008), and include a new characteristic dimension from prior analysis of images, like a skin rate.

Working on the lab procedures for child porn investigation and theft intellectual property cases, it is valuable to make refinements on the standard operative procedures to incorporate this novel approach.

References

Alghafli, K., Jones, A., and Martin, T., “Forensic analysis of the Windows 7 registry”, Proceedings of the 8th Australian Digital Forensics Conference, Edith Cowan University, Perth Western Australia, November 30th 2010.

Cantrell, G., Dampier, D., Dandass, Y., Niu, N. and Bogen, C., "Research toward a Partially-Automated, and Crime Specific Digital Triage Process Model", Computer and Information Science, Vol. 5, No. 2, March 2012.

Choudhury, A., Rogers, M., Gilliam, B. and Watson, K., "A novel skin tone detection algorithm for contraband image analysis", Third International Workshop on Systematic approaches to digital forensic engineering, May 2008.

Cunningham, P. and Delany S., "k-Nearest Neighbour Classifiers", Technical Report UCD-CSI-2007-4, 2007.

Dashora, K., Tomar, D., and Rana, J., “A practical approach for evidence gathering in Windows environment”, International Journal of Computer Applications (0975-8887), Vol. 5, No. 10, pp.21-27, August 2010.

Garfinkel, S., Farrell, P., Roussev, V. and Dinolt, G., "Bringing science to digital forensics with standardized forensic corpora", Digital Investigation, N°6, 2009.

Grillo, A., Lentini, G. and Ottoni, M., “Fast user classifying to establish forensic analysis priorities”, Proceedings of 5th International Conference on IT Security Incident Management & IT Forensics, IMF 2009.

Liu, Z., Lin, D., and Guo, F., “A method for locating digital evidences with outlier detection using support vector machine”, International Journal of Network Security, pp.301-308, May 2008.

Nance, K., Hay, B. and Bishop M., "Digital Forensics: Defining a Research Agenda", System Sciences 42nd Hawaii International Conference on, pp 1-6, 2009.

Sheldon, A., "The future of Forensic Computing", Digital Investigation, 2005, 2, pp. 31-35.

Wilson, D., and Martinez, T., “Improved heterogeneous distance functions”, Journal of Artificial Intelligence Research, n° 6, pp. 1-34, 1997.