Digital Investigation 10 (2013) 193–204

Contents lists available at SciVerse ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

A Machine Learning-based Triage methodology for automated categorization of digital media

Fabio Marturana [a,*], Simone Tacconi [b]

[a] *University of Rome "Tor Vergata", Civil Engineering and Computer Science, Via del Politecnico, 1, 00133 Rome, Italy*
[b] *Postal and Communications Police, Ministry of the Interior, Italy*

ARTICLE INFO

Article history:
Received 7 May 2012
Received in revised form 17 October 2012
Accepted 8 January 2013

Keywords:
Digital Forensics Triage
Machine Learning
Automated digital device categorization
Crime-related feature extraction
Data mining algorithms
Child pornography exchange
Copyright infringement

ABSTRACT

The global diffusion of smartphones and tablets, exceeding traditional desktops and laptops market share, presents investigative opportunities and poses serious challenges to law enforcement agencies and forensic professionals. Traditional Digital Forensics techniques, indeed, may be no longer appropriate for timely analysis of digital devices found at the crime scene. Nevertheless, dealing with specific crimes such as murder, child abductions, missing persons, death threats, such activity may be crucial to speed up investigations. Motivated by this, the paper explores the field of Triage, a relatively new branch of Digital Forensics intended to provide investigators with actionable intelligence through digital media inspection, and describes a new interdisciplinary approach that merges Digital Forensics techniques and Machine Learning principles. The proposed Triage methodology aims at automating the categorization of digital media on the basis of plausible connections between traces retrieved (i.e. digital evidence) and crimes under investigation. As an application of the proposed method, two case studies about copyright infringement and child pornography exchange are then presented to actually prove that the idea is viable. The term "feature" will be regarded in the paper as a quantitative measure of a "plausible digital evidence", according to the Machine Learning terminology. In this regard, we (a) define a list of crime-related features, (b) identify and extract them from available devices and forensic copies, (c) populate an input matrix and (d) process it with different Machine Learning mining schemes to come up with a device classification. We perform a benchmark study about the most popular mining algorithms (i.e. Bayes Networks, Decision Trees, Locally Weighted Learning and Support Vector Machines) to find the ones that best fit the case in question. Obtained results are encouraging as we will show that, triaging a dataset of 13 digital media and 45 copyright infringement-related features, it is possible to obtain more than 93% of correctly classified digital media using Bayes Networks or Support Vector Machines while, concerning child pornography exchange, with a dataset of 23 cell phones and 23 crime-related features it is possible to classify correctly 100% of the phones. In this regards, methods to reduce the number of linearly independent features are explored and classification results presented.

1. Introduction

Over the past few years, the diffusion of computer crimes caused by the large availability of low-cost and powerful digital devices (Internet Crime Compliance Center, 2009), has contributed to increased complexity and proportional expansion of digital forensic examinations. Traditional manually intensive and time-consuming forensic procedures have become unfit for large dataset that require huge analysis effort to extract crime-related evidence from digital media.

* Corresponding author. Tel.: +39 331 3683611.
E-mail address: marturana@libero.it (F. Marturana).

1742-2876/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.diin.2013.01.001

In 1955, in an article in The Economist, Cyril Northcote Parkinson first wrote that "work expands so as to fill the time available for its completion" (Parkinson's Law). The article was referring to public administration but today's corollary to this law might be that "computer forensic examinations expand in proportion to the increase in size of forensic units thus maintaining a significant backlog" (Parsonage, 2009).

We can thus argue, for the above considerations, that Parkinson's Law corollary is particularly suitable to describe today's issues faced by Digital Forensics.

As a consequence, researches aimed at proposing alternative methods to extract in short time crime-related features from digital media and find plausible crime-device connections are welcome as they could reverse this negative trend. For the purpose, various methods have been already developed to deal with the growing amount of digital evidence being encountered in criminal cases, enterprise and military investigations (Pearson and Watson, 2010). Among these a recently emerged one, called Triage, allows to rank digital media by probative content (i.e. crime-related features) and quickly identify the most relevant ones (Rogers et al., 2006). In accordance with this trend, our research in the field of Machine Learning-based device categorization aims at adding a contribution to the field of Digital Forensics Triage.

1.1. Methodology principles in brief

The paper describes a Triage model for crime-related and content-based classification of digital media, based on the work of Marturana et al. (2012a) and presents the results of two case studies in which the methodology was tested against forensic data from court cases of copyright infringement (Marturana et al., 2012b) and child pornography exchange (Marturana et al., 2011a,b). The proposed methodology, intended for both "live" and "dead" digital forensics investigations, includes a few steps aimed at processing digital media and identifying, at an early stage, the most interesting ones from the investigative point of view. Performing a Triage session of each device in the following area of interest: a) timeline of events, b) crime's specific features and c) suspect's private sphere (habits, skills and interests), it is indeed possible to rank it by likelihood that it may contain (or not) actionable evidence about the crime.

Hereafter the term "feature" will be regarded as a quantitative measure of a "crime-related indicator" or "plausible digital evidence", according to the Machine Learning terminology. As a consequence, an abstract concept (i.e. plausible digital evidence) is translated into a measurable one (i.e. feature) suitable to be processed by the proposed statistical model. An example of plausible copyright infringement evidence is the presence of file sharing URLs in the browser history of a PC found at the crime scene; the exact number of such visited websites represents the corresponding feature.

The proposed Triage model consists of the following four steps: *forensic acquisition*, *feature extraction and normalization*, *context and priority definition*, *data classification*.

The first one is an optional step that may be skipped due to time constraints and consists of creating a forensic copy of each device.

The second step is tasked of analyzing the input device or its forensic copy, and extracting a set of features from system configuration files, installed software, file statistics, browser history, system event log, mobile phone's call records, phonebook, sms, mms list etc. Occurrences of each of the aforementioned features are counted and normalized in a two-dimensional matrix, called *complete matrix*.

The third step is in charge of selecting a subset of them from the complete dataset. It is also possible to introduce in the model the timeline of interest, if needed. As a consequence, only a subset of the *complete matrix*, called *reduced matrix*, will be processed.

The final step is tasked of processing the *reduced matrix* and providing as output a Machine Learning-based categorization of each input device.

1.2. Paper outline

The outline of the remainder of the paper is as follows: we describe the proposed Triage-based model for automated categorization of digital media and define the operational workflow (Section 2). We define the methodology guidelines (Section 3) and discuss the results of two case studies conducted on the crimes of copyright infringement (Section 4) and child pornography exchange (Section 5). A published literature survey on the topic is presented (Section 6), and we finally draw the conclusion (Section 7) and suggest possible future work (Section 8).

2. Triage process model

This paragraph describes the proposed model for crime-related automated analysis and categorization of digital devices, that may be adopted in investigations about child pornography exchange, copyright infringement, hacking, murder and terrorism etc.

The process may be carried out in a forensic lab (*post-mortem Triage*), in parallel with the traditional forensic procedures (acquisition, retrieval and analysis) or on the crime scene (*live Triage*), to buy time in critical situations and is summarized in Fig. 1.

2.1. Model stages description

The *first stage* of the process, called *forensic acquisition*, is aimed at preserving digital evidence integrity and guarantee analysis repeatability. As mentioned early, when *live* analysis is performed, this stage could be skipped, due to time constraints.

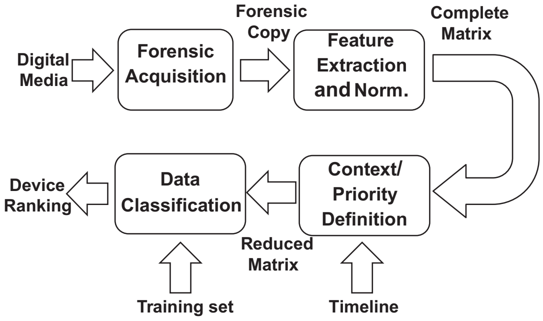


Fig. 1. Triage-based model.

The *second stage* of our workflow, called *feature extraction and normalization*, is tasked of extracting relevant from, alternatively, a forensic copy or a digital device using appropriate tools, scripts and procedures. Although entailing a real risk of exhibit tampering, on-scene *live* analysis may be justified by the need to provide investigative leads in time-critical situations.

This stage of the process is inspired, on the one hand, by Kent et al. (2006) who suggests to:

- look at the last date of change on critical files,
- examine configuration and start-up files,
- look for hacking tools (password crackers, copies of passwords, etc.),
- examine the password file for unauthorized accounts,
- search for keywords appropriate to the incident,
- search for hidden areas, slack space, and cache,
- look for changes to files, critical file deletions, and unknown new files,
- collect a list of all email addresses as well as visited and bookmarked URLs,

and, on the other hand, by Jansen and Ayers (2007) who identify the following potential evidence on cell phones:

- Subscriber and equipment identifiers,
- Date/time, language, and other settings,
- Phonebook information,
- Appointment calendar information,
- Text messages and multimedia messages,
- Dialed, incoming, and missed call logs,
- Electronic mail,
- Photos, audio and video recordings,
- Instant messaging and Web browsing activities,
- Electronic documents,
- Location information.

The aforementioned features, concerning user's habits, skills and interests, may be extracted from *system configuration files*, *installed software*, *file statistics*, *browser history*, *system event log* (i.e. device usage parameters), *mobile phone's call records*, *phonebook*, *sms*, *mms list* etc. and represent our model's independent variables. In this regards, we assumed to associate user's habits with the following set of parameters:

- percentage of modified files, ordered by time slot (morning, afternoon, evening, night),
- internet connections,
- monthly login frequency,
- system utilization,
- mobile phone's call records,
- phonebook,
- sms/mms list,
- location information.

With regards to user's technical skills, we included the analysis of:

- system configuration files,
- system log settings.

With regards to user's interests we focused on:

- stored and deleted files (audio and video recordings, photos, electronic documents, executable files),
- type of installed applications,
- instant messaging and web browsing activity.

The dataset collected so far is then normalized in a two-dimensional matrix called *complete matrix*, an example of which is summarized in Fig. 2, as output of the second stage.

An identifier of each digital device is indicated in the first row of the matrix whereas features nickname are shown in the leftmost column; By reading the matrix "by row", one can find values assumed by feature #X in each of the searched devices. As one can see, features are discrete variables assuming values from countable sets that may be both finite, in case a test or a multiple selection is made (true/false, yes/no/maybe etc.) or infinite (i.e. integer), in case occurrences of a variable are accounted for. Example of a multiple selection is, for instance, the presence or absence of a particular feature (i.e. are there installed hacking tools on the digital device? Yes/No) whereas examples of counting based features are, for instance, the number of stored or deleted jpg and video files, visited and bookmarked URLs. In our model, variables assuming values from finite countable sets are also called *nominal* whereas those assuming values from infinite countable sets are also called *numeric*.

The *third stage* of our workflow is called *context and priority definition* as it may introduce in the model the timeline of interest (i.e. during the crime or immediately before/after, when we suppose to find more evidence) and select the crime-specific features (i.e. such as the presence of child photos, in child pornography exchange or illegally downloaded software or movies in copyright infringement etc.). The output is a two-dimensional matrix called *reduced matrix* representing a subset of the aforementioned *complete matrix*.

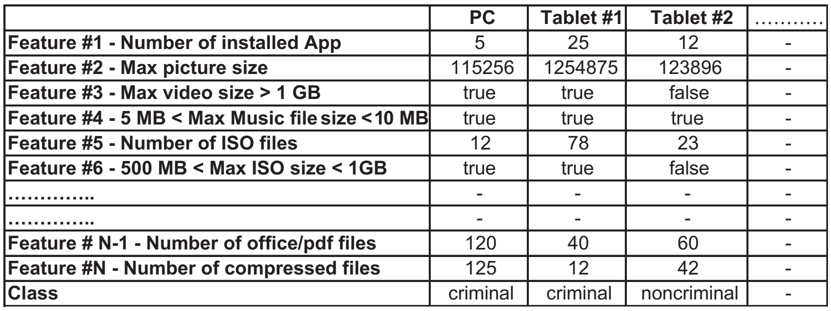
The *fourth stage* of the workflow is called *data classification* and is tasked of mining the *reduced matrix* in order to provide a categorization of each device, on the basis of the crime-related relevance of its content. In Fig. 3 a detailed workflow of the proposed model is depicted.

This phase is based on a collection of Machine Learning algorithms for categorization tasks taken from the Waikato Environment for Knowledge Analysis (WEKA) (Witten et al., 2011). WEKA[1] includes a large collection of categorization and clustering algorithms and tools under a common framework with an intuitive graphical user interface (Fig. 4).

3. Methodology guidelines

The preliminary step of the categorization process is to train a Machine Learning classifier to analyze digital

[1] WEKA is an open-source Java-based Machine Learning workbench, created by the Machine Learning Group at the University of Waikato, New Zealand and downloadable at http://www.cs.waikato.ac.nz/ml/weka.



	PC	Tablet #1	Tablet #2	………..
Feature #1 - Number of installed App	5	25	12	-
Feature #2 - Max picture size	115256	1254875	123896	-
Feature #3 - Max video size > 1 GB	true	true	false	-
Feature #4 - 5 MB < Max Music file size <10 MB	true	true	true	-
Feature #5 - Number of ISO files	12	78	23	-
Feature #6 - 500 MB < Max ISO size < 1GB	true	true	false	-
…………..	-	-	-	-
…………..	-	-	-	-
Feature # N-1 - Number of office/pdf files	120	40	60	-
Feature #N - Number of compressed files	125	12	42	-
Class	criminal	criminal	noncriminal	-

Fig. 2. Complete matrix.

devices and predict a dependent variable called *class*, according to the aforementioned crime-related features.

To do so it is necessary to create first a training-set consisting of representative samples of the crimes to be pursued (i.e. devices already classified) where each crime-sample relation is already known.

Once the training is over, the classifier is tasked of elaborating unclassified devices (i.e. included in the so-called test-set) and predicting their classes by determining plausible connections among devices and target crimes. Two categorization methods are available for the purpose: the Single Multiclass Categorization, where each processed device is associated, in one shot, to the most likely crime in the target set (i.e. child pornography exchange, copyright infringement, hacking, murder and terrorism) and the Multiple Binary Categorization, where different classifiers (one for each target crime) are derived by independent training processes.

In a Single Multiclass Categorization, the device under Triage is associated with the most likely crime among the aforementioned targets and the class variable is given the correspondent value.

In a Multiple Binary Categorization the class is given a binary value (i.e. yes or no, true or false etc.), depending on the likelihood that a plausible connection between the digital media and the target crime exists.

It is important to mention that the training step is crucial for the whole classification process and that crime representative samples must be chosen accurately to avoid systematic distortion and training bias. Adopting the iterative and predictive method called *N-folds cross-validation*, it is possible to evaluate classifier's learning effectiveness (Witten et al., 2011). Such procedure splits the training-set into N approximately equal partitions, each in turn used for testing and the remainder for training and it is repeated N times so that every sample is used exactly once for testing.

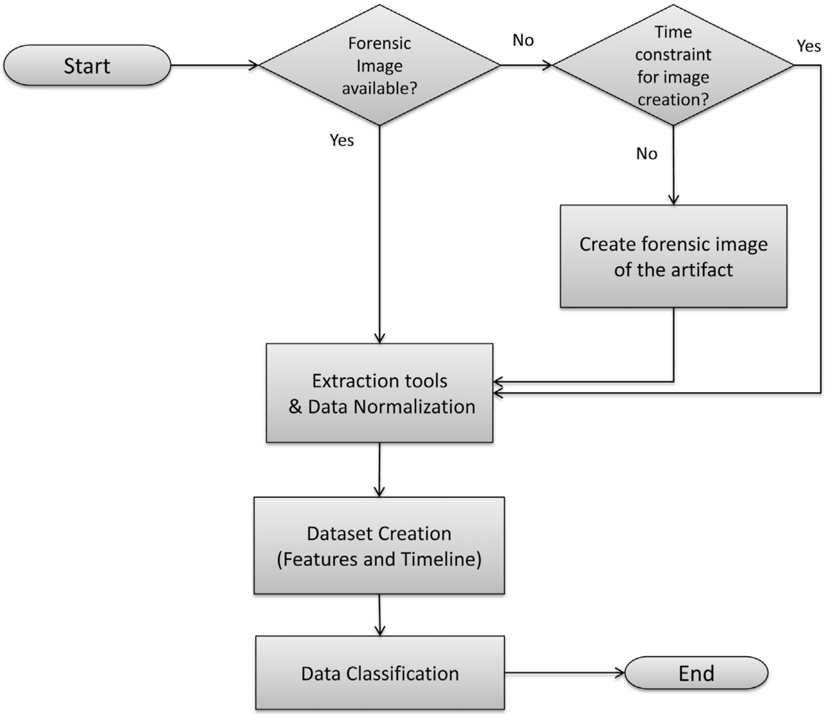
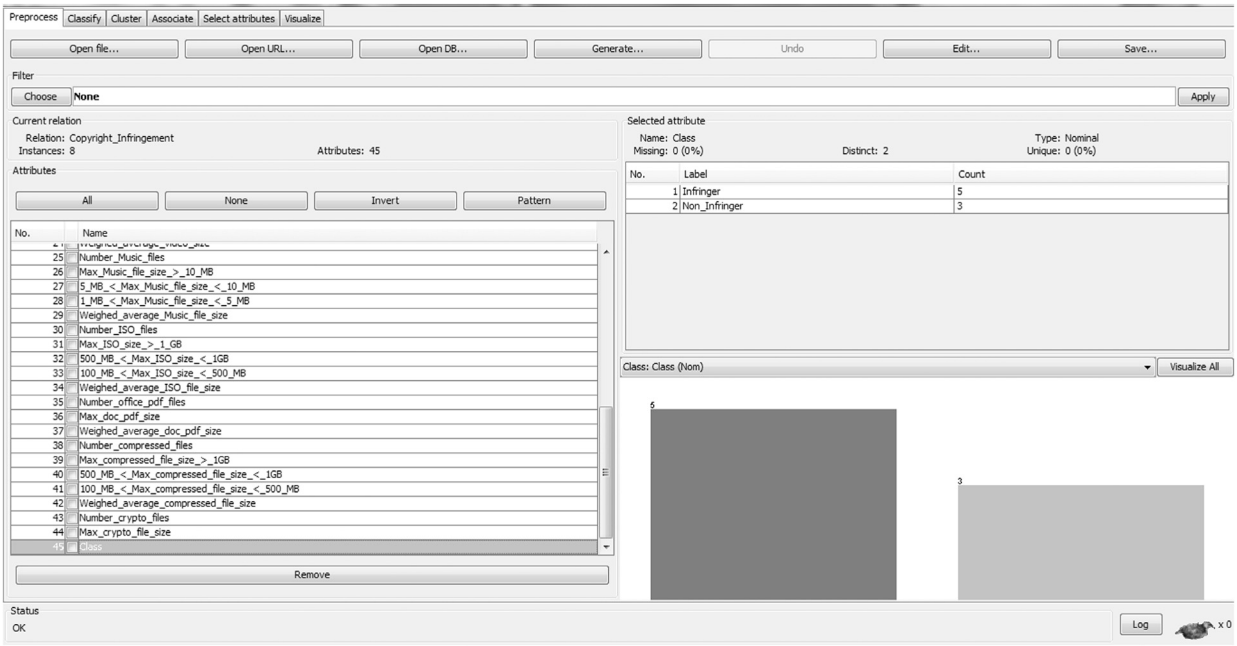

Start
Forensic Image available?
No
Time constraint for image creation?
Yes
Yes
No
Create forensic image of the artifact
Extraction tools & Data Normalization
Dataset Creation (Features and Timeline)
Data Classification
End

Fig. 3. Triage-based workflow.



Filter: None

Current relation
Relation: Copyright_Infringement
Instances: 8
Attributes: 45

Selected attribute
Name: Class
Type: Nominal
Missing: 0 (0%)
Distinct: 2
Unique: 0 (0%)

No.	Label	Count
1	Infringer	5
2	Non_Infringer	3

Class: Class (Nom)

Attributes

No.		Name
24	□	Weighed_average_video_size
25	□	Number_Music_files
26	□	Max_Music_file_size_>_10_MB
27	□	5_MB_<_Max_Music_file_size_<_10_MB
28	□	1_MB_<_Max_Music_file_size_<_5_MB
29	□	Weighed_average_Music_file_size
30	□	Number_ISO_files
31	□	Max_ISO_size_>_1_GB
32	□	500_MB_<_Max_ISO_size_<_1GB
33	□	100_MB_<_Max_ISO_size_<_500_MB
34	□	Weighed_average_ISO_file_size
35	□	Number_office_pdf_files
36	□	Max_doc_pdf_size
37	□	Weighed_average_doc_pdf_size
38	□	Number_compressed_files
39	□	Max_compressed_file_size_>_1GB
40	□	500_MB_<_Max_compressed_file_size_<_1GB
41	□	100_MB_<_Max_compressed_file_size_<_500_MB
42	□	Weighed_average_compressed_file_size
43	□	Number_crypto_files
44	□	Max_crypto_file_size
45	□	Class

Status
OK

Fig. 4. WEKA data-mining tool (GUI).

Classifier's learning effectiveness is evaluated according to the following performance indicators:

- *precision* = *TP*/(*TP* + *FP*),
- *recall* = *TP*/(*TP* + *FN*),
- *f-measure* = 2**recall***precision*,

where *TP* = *True Positive*, *FP* = *False Positive* and *FN* = *False Negative*.

4. Case study on copyright infringement

This paragraph describes a case study about digital media categorization in court cases of copyright infringement, courtesy of the Italian Postal and Communications Police, as an application of the model described in Section 2. We provide a definition of copyright infringement in Section 4.1, describe the crime-related features and plausible traces retrievable from devices in Section 4.2 and finally summarize experiment details and discuss results in Section 4.3.

4.1. Copyright infringement definition

Copyright infringement can be defined as the criminal offense of unauthorized or prohibited use of works under copyright, infringing the copyright holder's exclusive rights. It occurs when a copyrighted work is reproduced, distributed, performed, publicly displayed, or made into a derivative work without the permission of the copyright owner.

This criminal offence can be equated with both *piracy* and *theft* since the first is the practice of predating statutory copyright law, intentionally committed for financial gain whereas the latter is considered a misuse of the exclusive rights of the copyright holder for personal gain and without authorization.

4.2. Copyright infringement-related features

The following is a list of copyright infringement-specific features retrievable from digital media with a description of their relevance to such criminal behavior:

- *audio/video files*. Files are classified by extension (.wav, .mp3, .mp4, .avi, .dvx, .mpeg, .mpg etc.) and recognized by means of the relative file header. A different weight is assigned to compressed audio and video files (e.g. MP3 and MPEG) since these may be illegal copies of copyrighted works.
- *ISO files or compressed archives*. An illegal reproduction of a copyright protected software, indeed, is usually stored within an ISO image or a compressed file (.zip, .rar, .tar, .gz) with an associated key generator to unlock the installation process.
- *Hacking tools* (key generators, password crackers etc.).
- *Peer-to-peer clients* (e.g. Emule, Kazaa, uTorrent etc.). P2P networks indeed are the most common way to share copyrighted material illegally.
- *Specific multimedia players* such as Winamp or VLC, commonly used by infringers to play copyrighted music or movies. Copyright infringers, indeed, are used to dislike multimedia players like Windows Media Player as it doesn't include most divx audio/video codec.
- *Web URL history*. The URL history may uncover illegal traces since some P2P clients, such as uTorrent, to work

properly, need first to collect the illegal file URL from a search engine such as www.isohunt.com for example.

The aforementioned features extracted from installed software list, file statistics and browser history identify with our model's independent variables.

The following is the complete list of features included in the case study:

- Number of installed App
- Number of chat/communication/IM App
- Number of illegal download/P2P App
- Number of crypto App
- Number of browser App
- Number of entertainment/utility App
- Number of downloader App
- Number of MP3 converter App
- Number of disk/image mount utilities App
- Number of visited URLs
- Number of hacking URLs
- Number of warez/illegal download URLs
- Number of picture files
- Number of produced picture files
- Number of downloaded picture files
- Max picture size
- Weighted average picture size
- Number of video files
- Number of produced video files
- Number of downloaded video files
- Max video size $>$ 1 GB
- 500 MB $<$ Max video size $<$ 1 GB
- 100 MB $<$ Max video size $<$ 500 MB
- Weighted average video size
- Number of Music files
- Max Music file size $>$ 10 MB
- 5 MB $<$ Max Music file size $<$ 10 MB
- 1 MB $<$ Max Music file size $<$ 5 MB
- Weighted average Music file size

–Number of ISO files

- Max ISO size $>$ 1 GB
- 500 MB $<$ Max ISO size $<$ 1 GB
- 100 MB $<$ Max ISO size $<$ 500 MB
- Weighted average ISO file size
- Number of office/pdf files
- Max doc/pdf size
- Weighted average doc/pdf size
- Number of compressed files
- Max compressed file size $>$ 1 GB
- 500 MB $<$ Max compressed file size $<$ 1 GB
- 100 MB $<$ Max compressed file size $<$ 500 MB
- Weighted average compressed file size
- Number of crypto files
- Max crypto file size

4.3. *Experiment details and results*

In the case study, we processed a dataset consisting of 9 personal computers related to copyright infringement investigations, courtesy of Italian Cybercrime Police Unit (1,3 Terabytes) and 4 forensic copies from Garfinkel's M57-Patents corpus[2] (24,1 Gigabytes). A scheme of the adopted post-mortem Triage workflow is shown in Fig. 5.

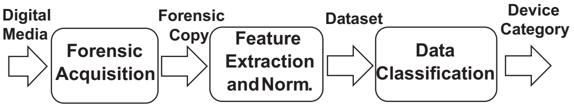


Fig. 5. Copyright infringement post-mortem triaging model.

In this case study we considered a simplified three-steps model where steps two and three described in Section 2 have been merged. We also skipped the timeline of interest being it considered a non-critical factor in copyright infringement.

During *forensic acquisition*, a forensic copy of each device was created, if not available. The next stage, called *feature extraction and normalization* was in charge of extracting the crime-related features from each forensic copy and creating the *dataset*, resulting from the combination of *training-set* (i.e. a set of samples with known classification) and *test-set* (i.e. a set of samples to classify).

In the case study the *complete matrix* defined in Section 2 resulted to be a rectangular $N \times M$, an example of which is depicted in Fig. 2; Excluding the header row and line, $N = 45$ were the crime-related features and $M = 13$ was the cardinality of the dataset, i.e. number of available digital media or forensic copies.

Finally, the *data classification* stage provided a categorization of the dataset by processing it with the Binary Categorization method described in Section 3.

We used WEKA Experimenter to perform two kind of analysis: the first, described in Section 4.3.1, called *complete dataset analysis* where the dataset was processed at once considering all the 45 features, and the second, described in Section 4.3.2, called *incremental features selection* where three different iterations, with a stepwise increasing number of features (respectively 15, 30, and 45), have been executed and compared. The benchmark study was performed on the basis of the following performance indicators: *percentage of correctly classified samples*, *mean absolute error*, *root mean square error*, *weighted average precision*, *weighted average recall*, *weighted average F_measure* (Witten et al., 2011).

In the case study we have also made a comparison of the four most popular categorizers, i.e. *Bayesian Networks*, *Decision Trees*, *Locally Weighted Learning and Support Vector Machines*, described in brief as follows, to find the ones that best fit the case in question:

Bayesian Networks (BN) estimate the conditional probability distribution of the values of the "class" feature (i.e. the *usage class*), given the values of the others. BN are drawn as a network of nodes, one for each feature, connected by directed edges in a directed acyclic graph.
Decision Trees (DT) apply a "divide-and-conquer" approach to the problem of learning from a set of independent

[2] M57-Patents corpus is downloadable at http://digitalcorpora.org/corpora/scenarios/m57-patents-scenario.

samples where the binary tree representation consists of nodes implying a test on one or more features and leafs giving a classification to all the samples that reach it. For example, to be classified, an unknown sample is routed down the tree according to the values of the features tested in successive nodes and, when a leaf is reached, the sample is assigned the same class.

Locally Weighted Learning (LWL) is a general algorithm associated with any learning technique that can handle weighted samples. It assigns weights using a sample-based method and builds a categorizer from the weighted samples, assuming independence within a neighborhood, not globally in the whole sample space.

Support Vector Machines (SVM) are algorithms used for classification, regression, or other tasks that constructs a hyper-plane or set of hyper-planes in a high-dimensional space. An SVM model is a representation of sample data as points in space, mapped so that samples of separate categories are divided by a clear gap. New samples are mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

4.3.1. *Complete dataset analysis*

In this experiment we adopted the iterative and predictive method called *10-folds cross-validation*, described in Section 3, to compare Bayesian Networks, Decision Trees, Locally Weighted Learning and Support Vector Machines performance. The method analyzed the complete set of 45 features producing the experimental results summarized in Table 1.

As one can see, in this scenario Bayesian Networks was the best performing categorizer with a 99% of correctly classified samples and weighted average precision of 0.99 and Support Vector Machines has ranked second with 93.5% of Percentage correct and weighted_average_precision of 0.93.

4.3.2. *Incremental feature selection*

In the second experiment we adopted an incremental approach, based on the aforementioned 10-folds cross-validation, where the dataset was processed three times, each with a stepwise increasing number of features, respectively 15, 30, and 45. The goal of the experiment was to make a benchmark study about the aforementioned categorizers with regards to the number of available features. No particular feature selection algorithm was adopted. In this regards Hall and Holmes (2003) and Yan and Shuqiang (2008) have proposed and compared different *feature reduction* techniques to reduce the feature-space dimension showing that it is possible to generally improve categorizers' performance in problems with considerable feature-space complexity. Our experimental results are summarized in Tables 2–5, where each row indicates a different sized feature set.

Comparing the results, it is possible to note that half of the categorizers (i.e. Bayes Networks and Support Vector Machines) have shown an increasing precision with regards to the number of available features and half (i.e. Decision Trees and Locally Weighted Learning) have behaved contrary.

Hughes (1968) has described this behavior, concerning the mean accuracy of statistical pattern recognizers (i.e. categorizers), showing that, to the contrary of expectations, an excessive increase in available features is often associated to a significant performance degradation and an increasing error rate. In particular, with a fixed number of training samples ($M = 13$), the overall accuracy (OA) initially grows up to a max_OA corresponding to $N*$ features and then decreases with the further increase of N. This is due to the limited number of training samples that may be inadequate for the classification process.

Table 1
Complete dataset comparative results.

Performance parameter	Machine Learning schemes			
	BN	*DT*	*LWL*	*SVM*
Percentage correct (%)	99	89.5	78.5	93.5
Mean_absolute_error	0.03	0.11	0.22	0.07
Root_mean_square_error	0.04	0.12	0.23	0.07
Weighted_avg_Precision	0.99	0.88	0.77	0.93
Weighted_avg_Recall	0.99	0.9	0.79	0.94
Weighted_avg_F_Measure	0.99	0.89	0.77	0.93

5. Case study on child pornography exchange

This paragraph describes a case study about cell phone categorization in court cases of child pornography exchange as an application of the model described in Section 2. We provide a definition of the crime in Section 5.1, define the crime-related features list in Section 5.2 and finally summarize experiment details and discuss results in Section 5.3.

5.1. *Child pornography exchange definition*

One of the primary ways in which pedophiles are misusing the Internet is to exchange child pornography. Traditionally pedophiles would traffic this material via clandestine exchange networks. Now pornographic pictures can be transmitted through computer networks on the Internet (Durkin, 2002). In this regards, the crime of child pornography exchange may be defined as the diffusion, under any form (e.g. download, upload, peer-to-peer, file sharing etc.), of sexual images and videos of young minors and babies for personal or commercial purposes through networks of sexual exploitation of children.

5.2. *Child pornography exchange related features*

Searching cell phones for child pornography possession or exchanged material is a forensic task in which phones' memory is analyzed to find specific files such as downloaded

Table 2
Percentage correct comparative results.

Dataset	Percentage correct (%)			
	BN	*DT*	*LWL*	*SVM*
Copyright_Infringement_15	83	100	100	76.5
Copyright_Infringement_30	90	97	98.5	88
Copyright_Infringement_45	99	89.5	78.5	93.5

Table 3
Weighted_avg_Precision comparative results.

Dataset	Weighted_avg_Precision			
	BN	*DT*	*LWL*	*SVM*
Copyright_Infringement_15	0.8	1	1	0.73
Copyright_Infringement_30	0.89	0.97	0.98	0.87
Copyright_Infringement_45	0.99	0.88	0.77	0.93

images or videos recordings and pictures taken with the embedded camera. Such procedure implies the use of extraction tools whose output must be analyzed and interpreted by a forensic practitioners. The proposed methodology, on the contrary, permits to identify the most relevant cell phones without a detailed preliminary inspection.

In this regards, the following is a list of mobile phone usage parameters that have been considered during the search of cell phones for child pornography possessed or exchanged material:

- *Phone model* (Smartphone, GSM);
- *Number of phonebook contacts* (stored both on SIM and phone);
- *Number of dialed/received/missed calls*;
- *Percentage of dialed/received/missed calls* (with regards to the specific time slot: Morning, Afternoon, Evening and if generated or received from phonebook contacts or not);
- *Average duration of dialed/received calls* (with regards to the specific time slot: Morning, Afternoon, Evening and if generated or received from phonebook contacts or not);
- *Number of received/sent SMS/MMS*;
- *Percentage of received/sent SMS/MMS* (with regards to the specific time slot: Morning, Afternoon, Evening and if they are sent or received from phonebook contacts or not);
- *Number and percentage of visited URLs* (with regards to the specific time slot: Morning, Afternoon, Evening and if they are bookmarked or not;
- *Number and percentage of downloaded images and videos files* (downloaded or created by the embedded camera);
- *Number of sent/received email*;
- *Number of stored notes.*

The following is an example of *complete matrix* analyzed in the case study of child pornography exchange (Fig. 6).

5.3. Experiment details and results

The case study is based on a collection of data extracted from 23 seized cell phones and smartphones, courtesy of the Italian Postal and Communications Police and concerning different crimes such as child pornography exchange, non-disclosure agreement violation, human trafficking and extortion.

Table 4
Weighted_avg_Recall comparative results.

Dataset	Weighted_avg_Recall			
	BN	*DT*	*LWL*	*SVM*
Copyright_Infringement_15	0.83	1	1	0.77
Copyright_Infringement_30	0.9	0.97	0.99	0.88
Copyright_Infringement_45	0.99	0.9	0.79	0.94

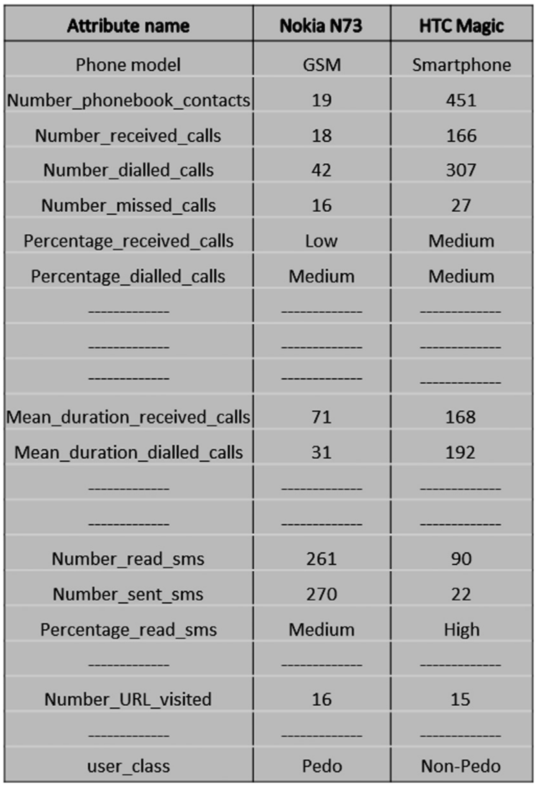
Table 5
Weighted_F_Measure comparative results.

Dataset	Weighted_F_Measure			
	BN	*DT*	*LWL*	*SVM*
Copyright_Infringement_15	0.81	1	1	0.74
Copyright_Infringement_30	0.89	0.97	0.98	0.87
Copyright_Infringement_45	0.99	0.89	0.77	0.93

In the case study we considered the simplified three-steps model described in Section 4.3 and skipped the timeline of interest as it is a non-critical factor in child pornography exchange.

The *complete matrix* defined in Section 2 resulted to be a rectangular $N \times M$ where $N = 114$ were the initial crime-related features and $M = 23$ was the cardinality of the dataset (i.e. the number of available phones).

The case study compared the following three classifiers: Bayesian Networks, Decision Trees and Locally Weighted Learning, described in Section 4.3, to find the ones that best fit the case in question. Classifiers' learning effectiveness was evaluated on the basis of *Precision*, *Recall* and *F-measure* described in Section 3. We used WEKA Explorer to perform different kind of analysis: the first, described in Section



Attribute name	Nokia N73	HTC Magic
Phone model	GSM	Smartphone
Number_phonebook_contacts	19	451
Number_received_calls	18	166
Number_dialled_calls	42	307
Number_missed_calls	16	27
Percentage_received_calls	Low	Medium
Percentage_dialled_calls	Medium	Medium
-------------	-------------	-------------
-------------	-------------	-------------
-------------	-------------	-------------
Mean_duration_received_calls	71	168
Mean_duration_dialled_calls	31	192
-------------	-------------	-------------
-------------	-------------	-------------
Number_read_sms	261	90
Number_sent_sms	270	22
Percentage_read_sms	Medium	High
-------------	-------------	-------------
Number_URL_visited	16	15
-------------	-------------	-------------
user_class	Pedo	Non-Pedo

Fig. 6. Complete matrix in child pornography exchange.

5.3.1, called *complete dataset*, the second, described in Section 5.3.2, called *reduced training-set (numeric)* and the third, described in Section 5.3.3, called *reduced training-set (numeric-nominal)*.

5.3.1. *Complete dataset*

The available dataset (i.e. 114 features and 23 phones) was processed at once with the Binary Categorization method described in Section 3. The 10-folds cross-validation was adopted to train Bayesian Networks, Decision Tree and Locally Weighted Learning classifiers. The output of the classification process is summarized in Table 6 and shows that Decision Tree algorithm performed better than others.

On average, all three classifiers were able to classify correctly more than half of the 23 phones.

To reduce average error rates and improve classifiers performance, according to Huges' theory described in Section 4.3.2, we applied some linear algebra to calculate the number of linearly independent features and reduced it accordingly. Two reduced feature sets, respectively with numeric features only (Section 5.3.2) and with both numeric and nominal features (Section 5.3.3) were created. Finally, the dataset was split up in two parts: a training-set consisting of 21 phones out of 23, and a test-set with the remaining two. Corresponding results are summarized in the following two sections.

5.3.2. *Reduced training-set (numeric)*

Given that that the rank or maximum number of linearly independent rows and columns of a generic $m \times n$ matrix (M) cannot be greater than m nor n (i.e. rank (M) $<=$ min (m,n)), in our categorization problem there were no more than 23 linearly independent features. Adopting therefore a manual feature selection algorithm, we first selected the numeric features, obtaining an $m' \times n$ matrix with $m' = 63$ and $n = 23$ and then applied the Gauss Elimination algorithm to reduce the $m' \times n$ matrix to a row-echelon form, finding out that the value of rank (M) is exactly 23, that confirms the initial hypothesis. We selected accordingly the set of 23 plausibly linearly independent numeric features summarized as follows (Table 7).

Classification results are summarized in Table 8.

In this case all the categorizers classified correctly all the testing instances (i.e. cell phones) with the implication that reducing the feature space to a set of non-redundant numeric ones implies that classifiers perform better than in the case of the complete dataset.

5.3.3. *Reduced numeric-nominal training-set*

In this case being not possible to apply the Gauss Elimination algorithm, we manually selected the set of 23 plausibly linearly independent numeric and nominal features summarized as follows (Table 9).

Classification results are summarized in Table 10.

In this case two categorizers out of three classified correctly all the testing instances (i.e. cell phones) with the implication that reducing the feature space to a set of non-redundant numeric-nominal ones implies that classifiers perform better than in the case of the complete dataset.

Table 6
Complete dataset comparative results.

Performance parameter	Machine Learning schemes		
	BN	*DT*	*LWL*
Weighted_avg_Precision	0.553	0.68	0.644
Weighted_avg_Recall	0.579	0.684	0.632
Weighted_avg_F_Measure	0.56	0.644	0.636

Table 7
Reduced features list (numeric).

Feature name	Feature type
Phone model	{GSM, smartphone}
Number_phonebook_contacts	Numeric
Number_received_calls	Numeric
Number_dialled_calls	Numeric
Number_missed_calls	Numeric
Mean_duration_received_calls	Numeric
Mean_duration_dialled_calls	Numeric
Number_read_sms	Numeric
Number_sent_sms	Numeric
Number_read_mms	Numeric
Number_sent_mms	Numeric
Number_downloaded_picture_files	Numeric
Number_downloaded_video_files	Numeric
Number_downloaded_audio_files	Numeric
Number_produced_picture_files	Numeric
Number_produced_video_files	Numeric
Number_produced_audio_files	Numeric
Number_URL_visited	Numeric
Number_URL_bookmarks	Numeric
Number_sent_email	Numeric
Number_received_email	Numeric
Number_notes_memo	Numeric
user_class	{Criminal, non-criminal}

6. Related work

Dealing with crimes such as murder, child abductions, missing persons, death threats, just to name a few, the need for a timely identification, analysis and interpretation of digital evidence found at the crime scene is crucial since it could be the difference between life and death for the victim. In those cases the inspection of each powered-on computer and digital media found on-scene with a live Triage tool could provide, indeed, investigators with actionable intelligence to proceed with the search. On the other hand, when the aim is to search lots of digital media or to dispose of a consistent backlog of data at lab, emerging post-mortem Triage techniques could provide a viable way to prioritize and rank digital media in order to make the subsequent evidence analysis easier.

Rogers et al. (2006) have proposed a live forensics model called Cyber Forensic Field Triage Process Model (CFFTPM),

Table 8
Reduced training-set (numeric) comparative results.

Performance parameter	Machine Learning schemes		
	BN	*DT*	*LWL*
Weighted_avg_Precision	1	1	1
Weighted_avg_Recall	1	1	1
Weighted_avg_F_Measure	1	1	1

Table 9
Reduced numeric-nominal features list.

Feature name	Feature type
Telephone type	{GSM, smartphone}
Number_phonebook_contacts	Numeric
Percentage_received_calls	{Low, medium, high}
Percentage_dialled_calls	{Low, medium, high}
Percentage_missed_calls	{Low, medium, high}
Mean_duration_received_calls	Numeric
Mean_duration_dialled_calls	Numeric
Percentage_read_sms	{Low, medium, high}
Percentage_sent_sms	{Low, medium, high}
Percentage_read_mms	{Low, medium, high}
Percentage_sent_mms	{Low, medium, high}
Percentage_downloaded_picture_files	{Low, medium, high}
Percentage_downloaded_video_files	{Low, medium, high}
Percentage_downloaded_audio_files	{Low, medium, high}
Percentage_produced_picture_files	{Low, medium, high}
Percentage_produced_video_files	{Low, medium, high}
Percentage_produced_audio_files	{Low, medium, high}
Number_URL_visited	Numeric
Number_URL_bookmarks	Numeric
Number_sent_email	Numeric
Number_received_email	Numeric
user_class	{Criminal, non-criminal}

that deals with gathering actionable intelligence on the crime scene. The model, aimed at time-critical investigations, defines a workflow for on-scene identification, analysis and interpretation of digital evidence, without the requirement of acquiring a complete forensic copy or taking the system back to the lab for an in-depth examination. The proposed methodology, although entailing a real risk of exhibit tampering and justified by the need to provide investigative leads in time-critical situations, involves however protecting the integrity of the evidence and/or potential evidence for further examination and analysis.

Drawing its inspiration from CFFTPM and considering the challenges posed by the battlefield crime scene, Pearson and Watson (2010) have recently defined the Digital Triage Forensics (DTF) process, a methodology designed to let military investigators collect actionable intelligence as well as evidence of a crime. The DTF model differs from the CFFTPM in that the Triage processing is done at the Forward Operating Base (FOB), before lab analysis, instead of the crime scene. The reason is simple: the forensic team, who is tasked of collecting evidence on the battlefield, must always operate in a limited time frame and under safe conditions. Safety and time are defined by Pearson et al. as follow:

– *Safety* for a combat investigator is the amount of exposure to be able to gather evidence from the battlefield crime scene. While the investigator is exposed, he/she risks sniper attacks, indirect or direct fire, secondary IED, or mortar attacks.
– *Time* is defined as the time at disposal to conduct the battlefield investigation. Typically, the investigator will have from 10 to 60 min to collect and process the battlefield crime scene.

Recently a new trend combining Computer Forensics and Machine Learning principles is taking hold in the research community. Veena et al. (2010) have recently proposed a methodology for file type analysis across flash drive based on clustering and supervised classification techniques.

Table 10
Reduced training-set (numeric-nominal) comparative results.

Performance parameter	Machine Learning schemes		
	BN	*DT*	*LWL*
Weighted_avg_Precision	1	1	0.687
Weighted_avg_Recall	1	1	0.5
Weighted_avg_F_Measure	1	1	0.667

As far as Mobile Forensics is concerned, moreover, Marturana et al. (2011a,b) have recently conducted a research on mobile handsets' content classification and they have proposed two possible applications of post-mortem Triage-based device categorization. In Marturana et al. (2011a), authors have created a corpus of data extracted from handsets and smartphones in court cases of child pornography exchange, corporate espionage, murder and human trafficking and they have defined a procedure to rank devices by owner's usage profile with a Data-Mining-based Single Multiclass Categorization. In Marturana et al. (2011b) authors have used the same data corpus to further implement a Single Binary Categorization to classify mobile handsets and smartphones allegedly used in child pornography exchange. In both cases, adopting the proposed methodology, authors were able to identify devices requiring additional lab processing by analyzing the data corpus.

Further, Decherchi et al. (2009, 2010) have conducted an important study on the application of clustering text mining techniques for text analysis purposes during digital investigations. The work has addressed forensic text clustering based on an adaptive model that arranged unstructured documents into content-based homogeneous groups.

We finally mention a valuable research conducted by Garfinkel et al. (2009) in the field of digital forensic experiments reproducibility in which the authors have made an important effort to explain the need to create standardized data corpora that may be adopted to test new forensic tools and techniques. They also contributed to create some large-scale standardized forensic corpora, available for research and educational purposes.[3]

7. Conclusion

We have proposed a new approach to Digital Forensics Triage and described a methodology for automated device categorization that may be adopted during *live* and *post-mortem* digital investigations. The proposed methodology, potentially applicable to a variety of crimes such as child pornography exchange, copyright infringement, hacking, murder and terrorism, just to name a few, was implemented and tested in two case studies of copyright infringement and child pornography exchange to prove its viability and effectiveness.

[3] The data corpora are downloadable at http://digitalcorpora.org/corpora/disk-images.

It is important to mention that Triage in Digital Forensics is still in its infancy as further experiments are underway, and every approach to it has potential limitations due to the lack of experimental data that must be considered on a case-by-case basis. Every investigation is indeed unique and in many cases it is impossible to identify specific kinds of data that can be related to a crime. For example the only evidence of a fraud on a computer may be the presence of certain documents, spreadsheets, and coded communications. Such evidence is highly case specific and not very different from normal emails, documents and spreadsheets that the person creates. Furthermore, the fraudster may be creating these artifacts during normal working hours as part of his normal activities.

Nevertheless, the statistical approach presented in the paper has shown to be valid in identifying evidence related to copyright infringement and child pornography exchange. However, for this approach to work it is necessary (a) to own a deep understanding of plausible connections between investigated crime and potential digital evidence and (b) to create large dataset of historical data to use for classification purpose.

Moreover the proposed methodology is intended to be of support to Traditional Digital Forensics techniques and not to replace them, speeding-up the identification of plausible connections among criminal conducts and device usage on a commonality basis that, if done manually, would be a time-consuming activity.

The two research questions that we have addressed in this paper were the following:

Which are the potential benefits of applying Machine Learning principles to Digital Forensics?
Is it possible to adopt Triage-based techniques to (a) automate the analysis of digital devices, (b) reduce analyst's backlog, (c) increase investigation efficiency?

To provide a concrete answer to the questions above, we have analyzed the model proposed by Marturana et al. (2012a), whose working principles are detailed in Section 2 and then we have extended it to both *live* and *post-mortem* analysis of digital media.

The proposed methodology is based on a workflow consisting of four phases: *forensic acquisition*, *feature extraction and normalization*, *context and priority definition*, *data classification*, is aimed at extracting and analyzing crime-related features concerning user's habits, skills and interests from digital devices, and categorizing them accordingly.

Such features, representing our model's independent variables, may be extracted from *system configuration files*, *installed software*, *file statistics*, *browser history*, *system event log* (i.e. device usage parameters), *mobile phone's call records*, *phonebook*, *sms*, *mms list* etc. and processed by Machine Learning categorizers. The dependent variable of our model, called *class*, is the output of the categorization process and represents the classification given to each device.

The methodology goal is to identify as soon, at the crime scene if possible, which device requires further analysis and which may be postponed or skipped.

8. Future work

Hopefully our work will serve as an inspiration to forensic professionals and researchers given that the model described can be virtually extended to a number of possible implementations concerning a variety of crimes. In this regard, interested readers who want to try their own implementations should spend some time to identify the crime-related features (i.e. the model's independent variables) and collect a consistent set of classified samples related to the crimes under investigation.

It is important to mention that success of digital device categorization will depend on accuracy adopted in the creation of the training-set that may be considered the core activity of the whole process. The higher the number of digital media that we have used to train the categorizer (e.g. hard disk drives, smartphones, handsets, tablets, PDAs etc.) is, indeed, the better the model will classify new devices.

It is also important to mention the importance of making a benchmark study to compare different categorizers to find the ones that best fit the case under investigation given that, under certain circumstances, a classifier may outperform others.

Acknowledgments

The paper is based on a research conducted in collaboration with the Italian Carabinieri Corps and the Computer Forensics Unit of the Italian Postal and Communications Police, Ministry of the Interior.

References

Decherchi Sergio, Tacconi Simone, Redi Judith, Sangiacomo Fabio, Leoncini Alessio, Zunino Rodolfo. Text clustering for digital forensic analysis. In: Proceedings of 2nd international workshop on computational intelligence in security for information systems (CISIS). Advances in intelligent and soft computing (AISC), vol. 63. Springer; 2009.

Decherchi Sergio, Tacconi Simone, Redi Judith, Sangiacomo Fabio, Leoncini Alessio, Zunino Rodolfo. Text clustering for digital forensic analysis. Journal of Information Assurance and Security (JIAS) 2010; 5(4):384–91. Dynamic Publishers.

Durkin Keith. Misuse of the internet by pedophiles. Current perspectives on sex crimes. Sage Publications; 2002.

Garfinkel Simson, Farrell Paul, Roussev Vassil, Dinolt George. Bringing science to digital forensics with standardized forensic corpora. Digital Investigation 2009;6(Suppl. 1):S2–11. Elsevier.

Hall Mark A, Holmes Geoffrey. Benchmarking attribute selection techniques for discrete class data mining. IEEE Transactions on Knowledge and Data Engineering 2003;15(3).

Hughes Gordon P. On the mean accuracy of statistical pattern recognizers. IEEE Transactions on Information Theory 1968.

IC3 (Internet Crime Compliance Center). Internet crime report; 2009.

Jansen Wayne, Ayers Rick. Guidelines on cell phone forensics. Recommendations of the National Institute for Standard and Technology (NIST); 2007.

Kent Karen, Chevalier Suzanne, Grance Tim, Dang Hung. Guide to integrating forensic techniques into incident response. Recommendations of the National Institute for Standard and Technology (NIST); 2006.

Marturana Fabio, Bertè Rosamaria, Me Gianluigi, Tacconi Simone. Mobile forensics "triaging": new directions for methodology. In: Proceedings of VIII conference of the Italian chapter of the Association for Information Systems (ITAIS). Springer; 2011a.

Marturana Fabio, Bertè Rosamaria, Me Gianluigi, Tacconi Simone. A quantitative approach to triaging in mobile forensics. In: Proceedings of international joint conference of IEEE TrustCom-11/IEEE ICESS-11/FCST-11 (TRUSTCOM 2011); 2011b. p. 582–8.

Marturana Fabio, Bertè Rosamaria, Me Gianluigi, Tacconi Simone. Data mining based crime-dependent triage in digital forensics analysis. In: Proceedings of international conference on affective computing and intelligent interaction (ICACII), IERI lecture notes in information technology; 2012a.

Marturana Fabio, Bertè Rosamaria, Me Gianluigi, Tacconi Simone. Triage-based automated analysis of evidence in court cases of copyright infringement. In: Proceedings of first IEEE international workshop on security and forensics in communication systems (SFCS 2012), in conjunction with IEEE international conference on communications (ICC); 2012b. p. 8249–53.

Parsonage Harry. Computer forensics case assessment and triage – some ideas for discussion; 2009.

Pearson Stephen, Watson Richard. Digital triage forensics-processing the digital crime scene. Syngress; 2010.

Rogers Marcus K., Goldman James, Mislan Rick, Wedge Timothy. Computer forensics field triage process model. In: Conference on digital forensics, security and law; 2006.

Veena H Bhat, Prasanth G Rao, Abhilash RV, Deepa Shenoy P, Venugopal KR, Patnaik LM. A data mining approach for data generation and analysis for digital forensic application. IACSIT International Journal of Engineering and Technology 2010;2(3).

Witten Ian H, Frank Eibe, Hall Mark A. Data mining practical machine learning tools and techniques. 3rd ed. Elsevier; 2011.

Yan Jia, Shuqiang Yang. Forward semi-supervised feature selection based on relevant set correlation. In: Proceedings of international conference of IEEE computer science and software engineering; 2008. p. 210–3.

Fabio Marturana is a Ph.D. student at Computer Science Engineering Faculty of University of Rome “Tor Vergata”. He holds an M.S. in Electronic Engineering from Polytechnics of Turin. His research interests include Computer and Mobile Forensics, information security and Cloud Computing. In these fields he is co-author of the book “Cybercrime and Cloud Forensics: Applications for Investigation Processes”, IGI Global Publisher and of several scientific papers in international conferences.

Simone Tacconi holds a Ph.D. in Artificial Intelligent Systems from the Polytechnic University of Marche and an M.S. in Electronic Engineering from University of Ancona. He is currently Technical Director of the Computer Forensics Unit at the Italian Postal and Communications Police. His main scientific interests include computer security, cryptography, Cloud Computing and Digital Forensics. In these fields he is co-author of the book “Cybercrime and Cloud Forensics: Applications for Investigation Processes”, IGI Global Publisher and of several refereed scientific papers in international journals and conferences.