Fuzzy System-Based Suspicious Pattern Detection in Mobile Forensic Evidence

Konstantia Barmpatsalou(✉), Tiago Cruz, Edmundo Monteiro, and Paulo Simoes

Pólo II-Pinhal de Marrocos, CISUC/DEI, University of Coimbra, 3030-290 Coimbra, Portugal
{konstantia,tjcruz,edmundo,psimoes}@dei.uc.pt

Abstract. Advances in Soft Computing have increased the probabilities of implementing mechanisms that are able to predict human behaviour. One of the fields that benefits more from the particular improvements are Digital Forensics. Criminal activity involving smartphones shows interesting behavioural variations that led the authors to create a technique that analyzes smartphone users' activity and recognizes potentially suspicious patterns according to predefined expert knowledge in actual use case scenarios by the use of fuzzy systems with different configurations.

Keywords: Mobile forensics · Fuzzy systems · Membership functions

1 Introduction

In the recent years, new Digital Forensic (DF) techniques emerged with the aid of Hard Computing (HC) [1]. However, activity driven by human behaviour is characterized by uncertainty [2] and renders them inefficient. Actions performed by individuals that are depicted in the digital fingerprint of a mobile device cannot be strictly characterized as innocent or guilty, but as entities that provoke different degrees of suspiciousness concerning specific criminal actions. This paper is the first part of a two-step approach aiming to create a semi-automated decision-making methodology for Mobile Forensic (MF) investigation purposes. Firstly, expert knowledge is used in order to create the ground truth and generate suspicious patterns concerning the outcome of user actions in data types retrieved during a forensic acquisition. Afterwards, the knowledge is diffused to the creation of fuzzy systems and their equivalent rules. Finally, the fuzzy system outputs are evaluated against the ground truth. However, the schema will be complete in the second part, which consists of the use and performance evaluation [3] of a Neuro-Fuzzy System (NFS) or a back-propagation neural network (NN) in comparison to the fuzzy systems and is the authors' future work.

The rest of the paper is presented in the following manner. Section 2 contains the related work in the field, while Sect. 3 presents the respective methodology the authors followed. Section 4 performs the results evaluation and Sect. 5 concludes the paper.

P. Matoušek and M. Schmiedecker (Eds.): ICDF2C 2017, LNICST 216, pp. 106–114, 2018.
https://doi.org/10.1007/978-3-319-73697-6_8

2 Related Work

To the best of the authors' knowledge, noteworthy research has been conducted in the area of fuzzy and Neuro-Fuzzy data analysis for MF and similar disciplines, such as Intrusion Detection. Stoffel et al. [4] applied the fuzzy sets theory to evidence deriving from criminal activity in Switzerland and proved that their methodology is appropriate for "inferring expert-system-like rules from a forensic database" [4]. In order to detect *Denial of Service (DoS)* attacks in a computer network infrastructure, Kumar and Selvakumar [5] profited from the combination of the precise rule definition of fuzzy systems and the automatic rule acquisition of NNs. Automatic rule definition by a Neuro-Fuzzy system was also successful in cases of Android malware detection [6]. The next section describes the methodology the authors followed in order to develop the fuzzy systems for detecting suspicious patterns in mobile data.

3 Methodology

This section presents the proposed methodology concerning suspicious pattern detection from mobile datasets. The procedure consists of the construction of a use case scenario, the rule inference and the ground truth generation. Further details concerning the used datasets are provided and the fuzzy systems for the use case are configured.

3.1 Use Case Scenario

The authors used the FP 7 Project SALUS D2.3 publicly available deliverable [7] so as to determine a use case scenario with potential criminal activity occurrences. One of the use cases of the deliverable, public order demonstration or riot, was considered as the most suitable for the research purposes, due to the high probability of occurrence of unfortunate events involving mobile devices belonging to Protection and Disaster Relief (PPDR) officers. The case under examination concerns PPDR officers infiltrating the rioting forces and how this can be proved by their device seizure. The investigation authorities capture an image of the device at a given moment after the rioting incident, which is used as the base for further investigation. However, no assumptions can be made without the presence of expert knowledge, which is elaborated in detail below.

3.2 Expert Knowledge

The knowledge base encountered in the current paper is a hybrid compilation of incidents the use cases provided in the SALUS FP7 Project deliverables [7] and of on-field investigation practices provided by an officer of the *Greek Police Escort Teams Department (GPETD)*. The authors structured the rules of each fuzzy system present in the research. Due to space limitations, only the example of SMS data deriving from three devices will be presented. Another challenge

that the authors faced was the lack or unavailability of actual evidence retrieved from devices involved in criminal activities. As a result, delinquent actions had to be simulated and injected in the datasets as standalone patterns. The a-priori expert knowledge served as a solid background for the rule generation, which is analyzed in the following subsection.

3.3 Rule Inference

Using the aforementioned expert knowledge, the authors created the respective rules from a combination of the available data and the investigation directives for the use case. For the scenario of the rioting infiltration by PPDR officers, the following setup was created. Sent SMS texts retrieved from a device of a potential infiltrator may have the following attributes. If officers are infiltrators, they will use their devices to communicate with their accomplices only in cases of extreme necessity. As a result, the rate with which a sent message will appear is going to be very low. Most of the accomplices may use one-time payphones, which are equipped with SIM modules from the same country the incidents occur. Thus, recipients with local numbers are considered more suspicious. Finally, messages exchanged right before or during rioting are very short in length. Consequently, the sent SMS pattern (very low appearance frequency–very short length–local country code source) is considered the most suspicious. Nonetheless, the rule inference procedure needs a functioning dataset that is able to fulfil the research requirements in size and content. The following subsection covers in detail the challenges the authors faced in the quest of a suitable data source.

3.4 Datasets and Ground Truth Generation

Due to the increased sensitivity of mobile device data, there are not many available sources of mobile device images. A more appropriate alternative was the "Device Analyzer Dataset" [8], a collection of real-time usage data from Android devices. Each dataset is a compilation of snapshots belonging to a certain device and contains lists of attributes such as call logs, SMS texts, network usage statistics, location data, etc., retrieved during a considerable period of time. All the information is stored in a Comma Separated Value (.csv) file and each row consists of the data type header, alongside with the existing data. Pre-processing is essential in order to separate the data types and adjust the information to the research needs. Adapted information from three different mobile devices, namely (Dev. 1, Dev. 2 and Dev. 3) is used for SMS data. The data are formatted in a three-column .csv file and each column represents one attribute; message length, receivers' appearance frequency and receivers' localization. Each row is a SMS text with its equivalent characteristics, which will from now on be referred to as a pattern. The SMS data type can be represented as follows:

$$\texttt{SMS}(\texttt{Appearance\_Frequency}, \texttt{Length}, \texttt{Country\_Source}) \tag{1}$$

The next step is the generation of ground truth data, which included manual labelling for all the SMS patterns. Every tuple of attributes (see Eq. 1) corresponds to a suspiciousness numerical value in a scale from zero to one, where

zero is the lowest and one is the highest value. Since the datasets were not originally created for DF analysis purposes and the existence of potentially suspicious patterns is unlikely, the authors injected the datasets with suspicious attribute combinations so as to have a complete view of the future system performance.

3.5 Fuzzy System Configuration

In order to proceed to the creation of the fuzzy systems, the authors followed the guidelines provided by Fuller [9]. One of the first factors to be taken into consideration is that all input and output variables should be described approximately or heuristically. Their fuzzy approximation is depicted in Table 1.

Table 1. Fuzzy variable ranges

Input variable	Fuzzy approximation	Numerical range
Length	VERY SHORT, SHORT, MEDIUM, LONG, VERY LONG	1–600 characters
Appearance frequency	VERY LOW, LOW, MEDIUM, HIGH, VERY HIGH	1–1100 appearances
Country source	FOREIGN, UNDEFINED, LOCAL	0, 1 and 2
Output variable	Fuzzy approximation	Numerical range
Suspiciousness	VERY LOW, LOW, MEDIUM, HIGH, VERY HIGH	0.15, 0.25, 0.50, 0.75, 1

The first column represents the variable, whereas the second shows the linguistic ranges attributed to it. The third column presents their numerical range. The rules in Subsect. 3.3 have to be represented in a formal manner and be placed in the appropriate system section so as to become structural elements of the rule base. An example of a rule concerning suspicious patterns is presented below. The rest of the rules are formed in a similar manner, with different variable values.

```
IF (Appearance == Very_Low) && (Length == Low) && (Country ==
        Local) THEN (Suspiciousness == Very_High)
```

Afterwards, the authors reviewed and verified the criteria for "readability and interpretability of the variables and the rules that are deriving from them" [10], as they were presented by Guillaume and Charnomordic [11]. While aiming to maintain a high degree of semantic cohesion, every fuzzy set should represent a well-defined and non-vague concept. The fuzzy sets and the value range of each variable have specific meanings (See Table 1). Additionally, each fuzzy variable should not exceed the 7 ± 2 range fields, which is defined as the threshold for human perception capabilities [10]. In the current paper, the maximum number of different value ranges is 5. There is no point within the system's universe of

discourse that does not belong to at least one fuzzy set. Furthermore, a fuzzy set should be normal; in a fuzzy system $\overline{F}$, there should always exist at least one χ, the membership degree (height) of which should be equal to 1. Lastly, it is obligatory that "all fuzzy sets should overlap in a certain degree" [10]. After concluding the fuzzy system configuration phase, the system evaluation takes place.

4 Evaluation

The authors followed an evaluation methodology based on the comparison of the fuzzy systems' output and the ground truth values. With the ground truth considered the target and the fuzzy output being the feature variable, the fuzzy

Table 2. Evaluation metrics per membership function for the SMS Dev. 1 dataset

M.F	Algorithm	AUC	Accuracy	Precision	Recall	FPR
Triangular	kNN	0.583	0.267	0.811	0.267	0.175
	SVM	0.578	0.809	0.800	0.809	0.169
	Naive Bayes	0.567	0.805	0.649	0.805	0.174
	AdaBoost	**0.592**	**0.815**	**0.842**	**0.815**	**0.164**
	Random Forest	0.592	0.814	0.840	0.814	0.164
Trapezoidal	kNN	0.573	0.808	0.799	0.808	0.172
	SVM	0.573	0.808	0.799	0.806	0.172
	Naive Bayes	0.561	0.802	0.648	0.802	0.176
	AdaBoost	**0.574**	**0.808**	**0.846**	**0.808**	**0.171**
	Random Forest	**0.574**	**0.808**	**0.846**	**0.808**	**0.171**
Bell	kNN	0.923	0.951	0.951	0.9512	0.029
	SVM	0.748	0.824	0.825	0.824	0.102
	Naive Bayes	0.904	0.872	0.910	0.872	0.035
	AdaBoost	**0.974**	**0.981**	**0.981**	**0.981**	**0.009**
	Random Forest	0.945	0.963	0.964	0.963	0.021
Gauss	kNN	0.908	0.952	0.952	0.952	0.037
	SVM	0.858	0.864	0.889	0.864	0.058
	Naive Bayes	0.858	0.852	0.880	0.852	0.055
	AdaBoost	**0.925**	**0.960**	**0.961**	**0.960**	**0.030**
	Random Forest	0.915	0.956	0.956	0.956	0.032
Gauss2	kNN	0.924	0.961	0.961	0.961	0.0299
	SVM	0.884	0.871	0.903	0.871	0.0481
	Naive Bayes	0.882	0.865	0.893	0.865	0.0450
	AdaBoost	0.926	0.963	0.963	0.963	0.0305
	Random Forest	**0.931**	**0.963**	**0.963**	**0.963**	**0.0276**

output values of five systems configured with different membership functions (Triangular, Trapezoidal, Bell, Gauss and Gauss2) were classified into five different groups of suspiciousness using the *Nearest Neighbour*, *SVM*, *Naive Bayes*, *AdaBoost* and *Random Forest* classification techniques.

The confusion matrices were created and the following metrics were calculated in average for all the groups of suspiciousness; Area Under Curve (AUC) (higher positive-over-negative value ranking capability of a classifier), Accuracy (amount of correctly classified patterns over the total amount of patterns), Precision (ratio of True Positive (TP) values over the sum of TP and False Positives (FP)), Recall (TP rate or sensitivity, ratio of TP over the sum of TP and False Negative (FN) values) and False Positive Rate (FPR) (ratio of FP values over the sum of FP and True Negative (TN) values).

Table 2 contains the cumulative results for all the candidate membership functions and their respective metrics. After evaluating all the datasets (See Appendix A), the authors concluded that the Triangular and Trapezoidal membership functions perform worse than the rest of the other candidates under every classification algorithm. Moreover, the Bell membership function shows the best performance rates in every dataset. In the majority of the tests, AdaBoost showed the best performance rates. On the contrary, kNN, SVM and Naive Bayes performed poorly. Finally, the performance difference among the Bell, Gauss and Gauss2 membership function is very low and they can be considered as efficient alternatives. Figure 1 depicts the Receiver Operating Characteristic (ROC) Curves for two out of the five suspiciousness values of Table 1 (S = 0.75, S = 1) for the Dev. 3 dataset and the Bell membership function.

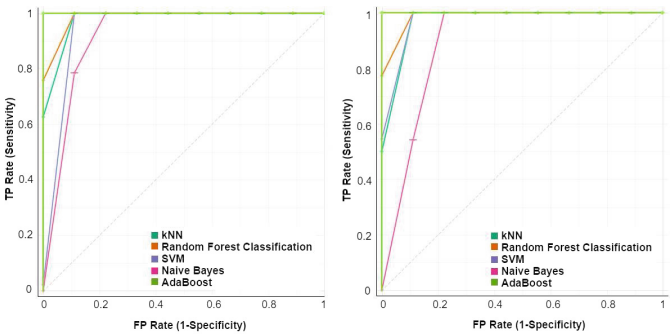

Fig. 1. ROC curves for the Dev. 3 dataset

5 Conclusions

The evaluation procedure was concluded successfully. The most appropriate parameters for the fuzzy systems were selected and the detection of potentially suspicious patterns was rather successful. Despite the satisfactory results, the aforementioned procedure revealed the need for a mechanism that will be able

to optimize the parameters of a fuzzy system, so as to achieve the replacement of trial and error methods by automatic approaches. Moreover, accessing real data concerning the use case circumstances would be the best approach for evaluating the fuzzy systems' efficiency. The upcoming stage of the authors' work comprises the experimentation with different data types and the development of an appropriate NFS or back-propagation NN that will co-operate with the fuzzy systems and complete the current contribution.

Acknowledgments. This work was partially funded by the ATENA H2020 EU Project (H2020-DS-2015-1 Project 700581). We also thank the team of FP7 Project SALUS (Security and interoperability in next generation PPDR communication infrastructures) and the GEPTD officer Nikolaos Bouzis for the fruitful discussions, feedback and insights on in-field investigation practices.

A SMS Datasets Evaluation Metrics

The appendix contains the analytical metrics for all the datasets tested in Sect. 4 as supplementary resources. Table 3 corresponds to the dataset of the second device (Dev. 2), whereas Table 4 refers to the dataset of the third device (Dev. 3).

Table 3. Evaluation metrics per membership function for the SMS Dev. 2 dataset

M.F.	Algorithm	AUC	Accuracy	Precision	Recall	FPR
Triangular	kNN	0.888	0.864	0.885	0.864	0.045
	SVM	0.875	0.822	0.840	0.822	0.052
	Naive Bayes	0.791	0.740	0.691	0.740	0.078
	AdaBoost	**0.897**	**0.850**	**0.870**	**0.850**	**0.043**
	Random Forest	0.890	0.867	0.888	0.867	0.045
Trapezoidal	**kNN**	**0.801**	**0.665**	**0.850**	**0.665**	**0.082**
	SVM	0.587	0.514	0.307	0.514	0.168
	Naive Bayes	0.727	0.684	0.606	0.684	0.107
	AdaBoost	0.742	0.704	0.647	0.704	0.102
	Random Forest	0.741	0.703	0.646	0.703	0.102
Bell	kNN	0.984	0.980	0.977	0.980	0.005
	SVM	0.976	0.968	0.966	0.968	0.008
	Naive Bayes	0.846	0.809	0.743	0.809	0.054
	AdaBoost	**0.998**	**0.997**	**0.997**	**0.997**	**0.001**
	Random Forest	0.991	0.989	0.986	0.989	0.004

(*continued*)

Table 3. (*continued*)

M.F.	Algorithm	AUC	Accuracy	Precision	Recall	FPR
Gauss	kNN	0.987	0.984	0.982	0.984	0.004
	SVM	0.980	0.972	0.9709	0.972	0.007
	Naive Bayes	0.850	0.815	0.746	0.815	0.052
	AdaBoost	**0.995**	**0.994**	**0.991**	**0.994**	**0.001**
	Random Forest	0.991	0.989	0.986	0.989	0.002
Gauss2	kNN	0.986	0.983	0.981	0.983	0.004
	SVM	0.988	0.984	0.982	0.984	0.003
	Naive Bayes	0.880	0.848	0.781	0.848	0.040
	AdaBoost	**0.989**	**0.986**	**0.983**	**0.986**	**0.003**
	Random Forest	0.988	0.984	0.982	0.984	0.003

Table 4. Evaluation metrics per membership function for the SMS Dev. 3 dataset

M.F.	Algorithm	AUC	Accuracy	Precision	Recall	FPR
Triangular	kNN	0.619	0.310	0.857	0.310	0.158
	SVM	0.611	0.582	0.508	0.582	0.159
	Naive Bayes	0.604	0.573	0.365	0.573	0.160
	AdaBoost	**0.617**	**0.591**	**0.651**	**0.591**	**0.156**
	Random Forest	0.617	0.590	0.610	0.590	0.157
Trapezoidal	kNN	0.608	0.294	0.571	0.294	0.143
	SVM	0.609	0.294	0.571	0.294	0.143
	Naive Bayes	0.600	0.571	0.365	0.571	0.162
	AdaBoost	**0.606**	**0.579**	**0.371**	**0.579**	**0.160**
	Random Forest	0.605	0.578	0.371	0.579	0.161
Bell	kNN	0.971	0.963	0.963	0.962	0.010
	SVM	0.937	0.906	0.922	0.906	0.025
	Naive Bayes	0.722	0.682	0.527	0.682	0.102
	AdaBoost	**0.990**	**0.986**	**0.986**	**0.986**	**0.004**
	Random Forest	0.983	0.978	0.978	0.978	0.033
Gauss	kNN	0.979	0.971	0.972	0.971	0.008
	SVM	0.940	0.909	0.975	0.975	0.025
	Naive Bayes	0.713	0.666	0.519	0.666	0.191
	AdaBoost	**0.990**	**0.986**	**0.986**	**0.986**	**0.006**
	Random Forest	0.981	0.975	0.975	0.975	0.006

(*continued*)

Table 4. (*continued*)

M.F.	Algorithm	AUC	Accuracy	Precision	Recall	FPR
Gauss2	**kNN**	**0.975**	**0.967**	**0.968**	**0.967**	**0.009**
	SVM	0.944	0.915	0.931	0.915	0.023
	Naive Bayes	0.716	0.671	0.521	0.671	0.108
	AdaBoost	0.949	0.920	0.935	0.920	0.022
	Random Forest	0.946	0.917	0.932	0.917	0.022

References

1. Barmpatsalou, K., Damopoulos, D., Kambourakis, G., Katos, V.: A critical review of 7 years of mobile device forensics. Digit. Invest. **10**(4), 323–349 (2013)
2. Gegov, A.: Fuzzy Networks for Complex Systems: A Modular Rule Base Approach, vol. 259. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-15600-7
3. Siddique, N., Adeli, H.: Computational Intelligence: Synergies of Fuzzy Logic, Neural Networks and Evolutionary Computing. Wiley, Hoboken (2013)
4. Stoffel, K., Cotofrei, P., Han, D: Fuzzy methods for forensic data analysis. In: 2010 International Conference of Soft Computing and Pattern Recognition, pp. 23–28 (2010)
5. Kumar, P.A.R., Selvakumar, S.: Detection of distributed denial of service attacks using an ensemble of adaptive and hybrid neuro-fuzzy systems. Comput. Commun. **36**(3), 303–319 (2013)
6. Shalaginov, A., Franke, K.: Automatic rule-mining for malware detection employing neuro-fuzzy approach. In: Norsk informasjons sikkerhets konferanse (NISK) (2013)
7. Nyanyo, A., Marques, H., Wickson, P., Brouwer, F., Blaha, M., Jelenc, D., Brouet, J., Junittila, K., Kolundzija, B.: Deliverable 2.3: SALUS use cases final. Technical report, SALUS Consortium (2014)
8. Wagner, D.T., Rice, A., Beresford, A.R.: Device analyzer: understanding smartphone usage. In: 10th International Conference on Mobile and Ubiquitous Systems: Computing, Networking, and Services, MOBIQUITOUS 2013, Tokyo, Japan (2013)
9. Fuller, R.: Neural Fuzzy Systems. Abo, Turku (1995)
10. de Lima, H.P., de Arruda Camargo, H.: A methodology for building fuzzy rule-based systems integrating expert and data knowledge. In: 2014 Brazilian Conference on Intelligent Systems, pp. 300–305 (2014)
11. Guillaume, S., Charnomordic, B.: Fuzzy inference systems: an integrated modeling environment for collaboration between expert knowledge and data using FISPRO. Expert Syst. Appl. **39**(10), 8744–8755 (2012)