International Journal of Mathematics and
Computer Science, **15**(2020), no. 4, 1221–1234

Classification of phases based on a Principal Component Analysis for Intrusion Detection Methods

El Mostafa Rajaallah

Laboratory: Mathematics, Computer Science and Engineering Sciences
Institut des Sciences du Sport
Hassan First University of Settat
Settat, Morocco

email: rajaallahelmostafa@gmail.com

(Received August 2, 2020, Accepted September 10, 2020)

Abstract

The impacts of an intrusion can be dangerous for the information system of an organizational being. According to the Quebec office of the French language, an intrusion is an operation which consists in accessing, without authorization, the data of a computer system or a network, bypassing or defusing the security devices put in place. The detection of an intrusion is not the end in itself, but also the optimization of the reaction time, that is to say minimize the time between detection and reaction, for this reason we use the of experts to assess the effectiveness of a method and its phases. In this work we will propose an approach based on principal component analysis allowing the development of the typology of an intrusion detection method based on the opinion of experts in the field. Certainly the matrix of the example used is very small, but the aim is to propose a classification approach based on a Principal Component Analysis (PCA). The profiles are drawn up on the basis of a detailed study of the variables and individuals in relation to the axes generated by the PCA.

Key words and phrases: Intrusion detection, PCA, Classification.
AMS (MOS) Subject Classifications: 68M25, 94A05.
ISSN 1814-0432, 2020, http://ijmcs.future-in-tech.net

1 Introduction

In recent years, many approaches have been developed for the management and detection of an intrusion, where one can distinguish two main categories of intrusion system: Abuse (based on the signatures of known models of attacks) and Anomalies (based on a learning phase of the normal functioning of a system).

For the first approach, it finds its limits, especially in the unknown attack models, so it requires a continuous update of its signature database.

For the category based on anomalies, the weak point is false alarms, any activity that derives from the functioning considered normal in the learning period will be considered abnormal.

In our article entitled: "Intrusion Detection System: To an Optimal Hybrid Intrusion Detection System" (Rajaallah et al. 2019) [1], and according to the results obtained from intrusion detection techniques and the comparison of open source IDS, we found that signature-based systems like Snort are powerful at detecting known attacks, but not enough to stop masked attacks. Especially, if a default configuration is used, because the attacker can test in advance with false attacks, to avoid detection. The use of anomaly detection (PHAD, NETAD and ALAD) can detect unusual events, but they require a large amount of training data.

There are three types of learning: 1) Supervised learning of attacks (Bognar, 2016) [2] proposed an approach based on neural networks with a very high rate of accuracy (99.9985%) and a rate of false alerts fairly low (3.006), but with several anomalies such as the need for a lot of memory. 2) Unsupervised learning: (Ajboye et al. 2015) [3] insisted on the usefulness of pre-processing all the data in the training sample before modeling, in order to differentiate the data which seems normal or abnormal and he proposed a model improving the precision of intrusion detection was obtained using the DBSCAN algorithm. This involves identifying points within a class (cluster) using the radius parameter to assign an instance to a class; despite its performance, this algorithm requires large resources of memory and computation, without taking into account the subjectivity at the level of the radius estimation as well as the number of points. 3) Hybrid learning, (D. Pierrot et al. 2018) [4] proposed a method combining supervised and unsupervised learning with a risk rating. According to the results presented, the method allowed optimal detection of security policy violations and behavioral derivations, the table below shows the feedback of five experts following the approach proposed by Ghoneim et al. (2014) [5] :

Tab. 1 Overview of expert feedback

Question about the usefulness of :	**E1**	**E2**	**E3**	**E4**	**E5**
Visualization (Phase 1)	5	4	3	4	4
Policy derivation (Phase 2)	5	4	4	4	5
Behavior derivation (Phase 2)	5	5	5	4	5
Risk management (Phase 3)	3	4	4	2	3
Action plan (Phase 4)	3	4	3	4	3

Despite the positive opinions of the experts, they criticized the way of presenting the information (source IP at risk, modification of behavior on the destination IP and risk rating) which should be easier to understand. With five experts you can easily assess the criteria or the phases, but if you have a large number of experts with a high scale, in this case it will be difficult or impossible to synthesize the data.

Generally, the proposed intrusion detection approaches requires a capture of network traffic and its analysis, which can impact the performance of an information system. which makes them very dependent on the resources used in the assessment. For this reason, we suggest evaluations of intrusion detection methods based on expert opinions.

In this work, we will propose an approach which allows synthesizing the data, even with a large number of experts. The proposed approach is based on the technique of PCA data analysis (Principal Component Analysis).

2 Proposed Approach

Principal Components Analysis PCA is a data analysis tool that allows the reduction of the dimensionality of a set of quantitative variables (Morineau, Aluja-Banet, 2000) [6]. It explores the variables (according to their correlations) and the similarities between individuals (according to their distances).

1. **Step 1: Identify individuals and variables**

   In our approach, we will consider the criteria as individuals and the experts as variables. Concerning the example treated is that of (D. Pierrot et al. 2018) (Table No. 1).

2. **Step 2: Determination of the number of axes to be interpreted**

   The main axes (or principal components, or factorial axes) are constructed so as to minimize the distances between individuals according to the criterion of least squares.

   What is the number of axes to retain?

   - There is no indisputable rule to limit the number of axes to be retained (Saporta, 2006) [7].
   - The choice of the number of axes depends on the inertia restored by each of the axes and the objectives assigned to the principal component analysis.

   Note: In practice, the first proper values often show an irregular decrease.

   For our example, the following table presents the eigenvalues and the eigenvectors:

   Tab. 2 eigenvalues and eigenvectors

	F1	F2	F3	F4
Eigenvalue	3,040	1,379	0,467	0,114
Variability (%)	60,806	27,570	9,344	2,280
% cumulé	60,806	88,376	97,720	100,000

3. **Step 3: Interpretation of the variables**

   - This stage is divided into two phases:
     (a) Analysis of correlations between variables.
     (b) Analysis of the correlations between variables and axes.
   - Retain for the analysis the correlations which seem the most significant (The significant correlations can be presented and classified in descending order).

- For the correlation between variables, it is preferable to distinguish positive correlations from negative ones.

Step 3.1. Correlations between variables.

- If the variables are close, they are strongly correlated, and the correlation is positive.
- If the variables are opposite, they are also strongly correlated, but the correlation is negative.
- If the variables are orthogonal, their correlation is zero.

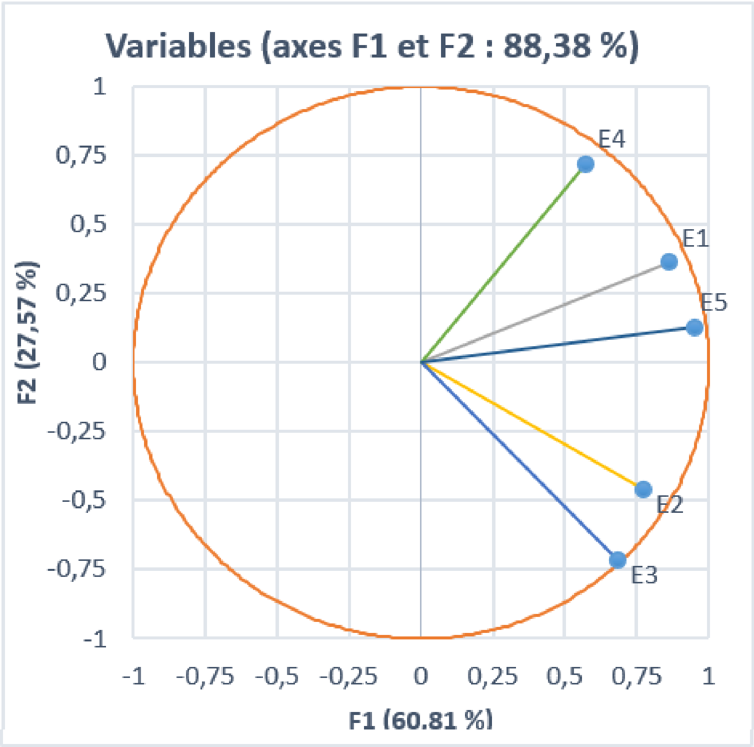

Figure 1: Circle of correlation of variables

According to this correlation circle, the correlation between the two variables E1 and E5 is very strong, the same between E2 and E3. On the other hand, the correlation between E2 and E4 is very weak, while the correlation between E3 and E4 is almost zero.

The correlation between E2 and E5 is medium, and the correlation between E1 and E2 is moderately weak.

Step 3.2.: Correlations between the axes and the variables.

- If the coordinates of the variables are close to the origin of the axes, the correlation with the axis concerned is not significant.
- If the coordinates of the variable are close to 1 on an axis, the correlation with this axis is strong.
- If the coordinates of the variable are close to 1 on both axes, the correlation is strong on the plane.
- The level of correlation between the variables and the Pi-j plane is determined as follows:

$$\mathrm{r}_{P(Axei,Axej)=}\mathrm{r2}_{(Axei)+}\mathrm{r2}_{(Axej).}$$

Note: It is possible to identify the significant correlations to retain between variables and axes (Anderson, 1984) [8]. All values greater than the absolute value of R are used for the analysis.

Abs(R) = 1,96/square(n+2) = 0,741, where n is the number of variables

we will retain the values marked in bold in the table below:

Tab. 3 Correlation of variables with the two axes 1 and 2

	Axe 1	Axe 2	Plan 1-2
E1	**0,862**	0,36	0,87
E2	**0,772**	-0,458	0,81
E3	0,684	**-0,712**	0,97
E4	0,57	**0,718**	0,84
E5	**0,954**	0,127	0,93

So we will retain the variable E1, E2 and E5 for axis 1, and the two variables E3 and E4 for axis 2.

Tab. 4 Axe 1: Variable-axis correlations and mirror effects

Negative informations	**Positive informations**
	+E1+ E2+E5
Mirror effect	
-E1-E2-E5	

Negative information	**Positive information**
+E3	+ E4
Mirror effect	
-E4	- E3

When a variable or a group of variables is correlated to an axis, in positive or negative coordinates, they evolve in the same direction.

Tab. 5 Axe 2: Variable-axis correlations and mirror effects

4. **Step 4: Interpretation of individuals**

Step 4.1.: Reading the plan of individuals

The following table presents the coordinates of the Individuals on the axes:

Tab. 6 Coordinates of individuals on the axes

	Axe 1	**Axe 2**
Visualization (Phase1)	-0,074	1,399
Policy derivation (Phase2)	1,062	0,710
Behavior derivation (Phase2)	2,693	-1,076
Risk management (Phase3)	-1,987	-1,687
Action plan (Phase4)	-1,694	0,653

Step 4.2: Analysis of the qualities of representation

The quality of representation of the Individuals on the factorial axes is measured by the cosine of the angle formed by the segment joining the center of gravity O to the point representing the Individual and the factorial axis.

Cos2(?)= d2(O, i')/d2(O, i), considering all the factor axes

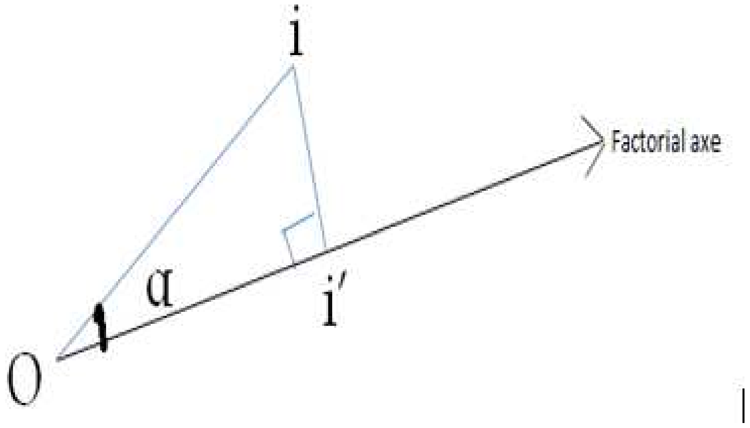

Figure 2: The square cosine

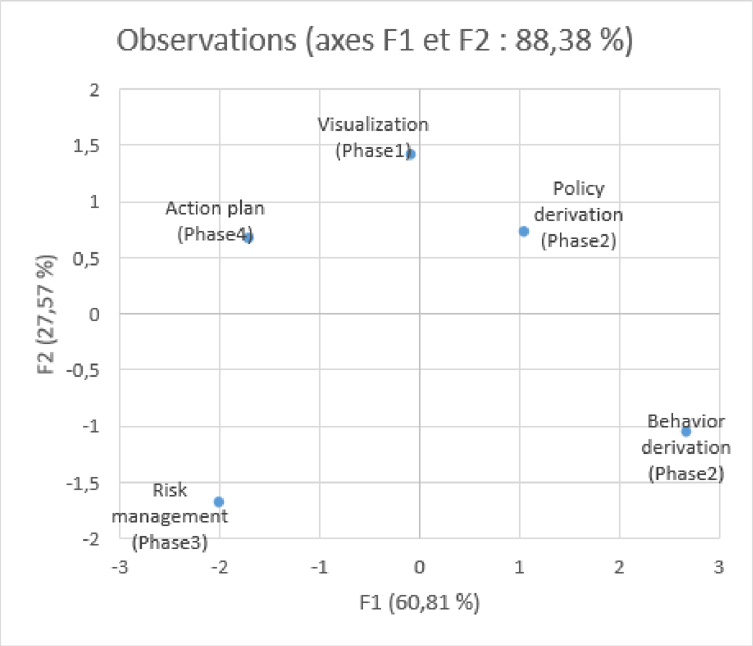

Figure 3: Observations of the variables

The calculation of the square cosines on a factorial plane requires the calculation of the sum of the cosines of the two axes which compose it.

In practice, a cosine rarely reaches the value 1. The mean or median (central value) of the square cosines of the Individuals by axes and on the factorial planes is an indicator of the quality of their representation.

Tab. 7 Representation qualities of individuals on the axes

	Axe 1	Axe 2	Plan 1-2
Visualization (Phase1)	0,002	**0,848**	0,85
Policy derivation (Phase2)	0,453	**0,203**	0,656
Behavior derivation (Phase2)	**0,830**	0,132	0,962
Risk management (Phase3)	**0,558**	**0,402**	**0,96**
Action plan (Phase4)	**0,653**	0,097	0,75
Median	**0,558**	**0,203**	–

The values in bold correspond for each observation to the factor for which the square cosine is greater than or equal to median.

The quality of representation is significant for a plan if it is significant for the two axes that make up the plan.

Tab.8 Summary table: Quality of representation

Axes, plan	**Axe 1**	**Axe 2**	**Plan**
Sig. rep.	Behavior der., Risk man., Action plan	Vis., Policy der., Risk man.	Risk man.

Rule: The Individuals selected therefore necessarily have a square cosine greater than or equal to the central value.

Can we limit ourselves to the quality of representation to assign an Individual to an axis or to the plane?

The factor axes are the synthesis of data relating to several variables and related to a set of individuals. It is therefore essential to identify individuals and measure their contribution to the construction of the axes (Didier Busca et Stéphanie Toutain 2009) [9] **.**

Step 4.3 .: Analysis of the contribution of individuals to the construction of axes.

- The relative contribution of an Individual i to the formation of a main component is the relative inertia of this Individual on the factorial axis F linked to this component. It is defined by:

CTR = (Score of i on the factorial axis F)2/(n*? $_F$**)**

	F1	F2	Plan 1-2
Visualization (Phase1)	0,036	**28,412**	28,448
Policy derivation (Phase2)	7,415	7,317	14,733
Behavior derivation (Phase2)	**47,702**	**16,787**	**64,488**
Risk management (Phase3)	**25,970**	**41,294**	**67,264**
Action plan (Phase4)	**18,877**	6,190	25,067
Median	**18,877**	**16,787**	–

Tab. 9 Quality of representation and contribution of individuals

The values in bold correspond for each observation to the axis for which CTR is greater than or equal to median.

The quality of contribution is significant for a plan if it is significant for the two axes that make up the plan.

Tab. 10 Summary table: Contribution to the construction of the axis

Axes, plan	**Axe 1**	**Axe 2**	**Plan**
Sig. rep.	Beh. der., Risk man., Act. plan	Vis., Beh. der., Risk man.	Beh. der., Risk man.

Tab. 11 Summary table: Quality of representation and contribution to the construction of axes

Individual	**C2A1**	**C2A2**	**C2P1-2**	**CTRA1**	**CTRA2**	**CTR**	**Liaison**
Vis.(Ph.1)	0,002	**0,848**	0,719	0,036	**28,412**	28,448	**Axe 2**
Policy der.Ph.2)	0,453	**0,203**	0,246	7,415	7,317	14,733	-
Beh. der.(Ph.2)	**0,830**	0,132	0,706	**47,702**	**16,787**	64,488	**Axe 1**
Risk man.(Ph.3)	**0,558**	**0,402**	**0,474**	**25,970**	**41,294**	67,264	**Plan 1-2**
Action plan(Ph.4)	**0,653**	0,097	0,436	**18,877**	6,190	25,067	**Axe 1**
Median	**0,558**	**0,203**	–	**18,877**	**16,787**	–	

From the previous summary table, we can develop the summary table axes 1 and 2, and following plan:

Axes, factorial plan	**Axe 1**	**Axe 2**	**Plan**
Sig. rep. and contr.	Beh. der., Action plan	Vis.	Risk man.

Tab. 12 Summary table: Quality of representation and contribution to the construction of the axes

The individual or the question "Policy derivation" has not been classified due to its weak contribution in the construction of the two axes of plan 1-2, we will need to deal with it in another plan with more loss of information.

5. Step 5 : Synthesis of the analysis (Development of profiles of individuals).

Taking into account the summary table Tab. 12 with the two tables Tab. 4 and Tab. 5 of variable-axis correlations and mirror effects of the two axes, as well as figure number 3 observations of the variables; we can deduce the profiles below:

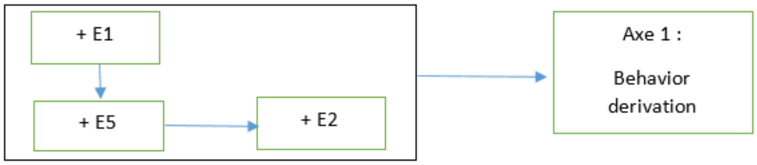

Figure 4: First profile

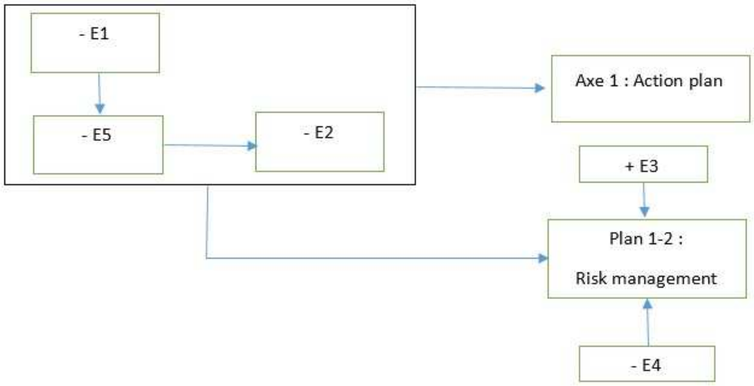

Figure 5: Second profile

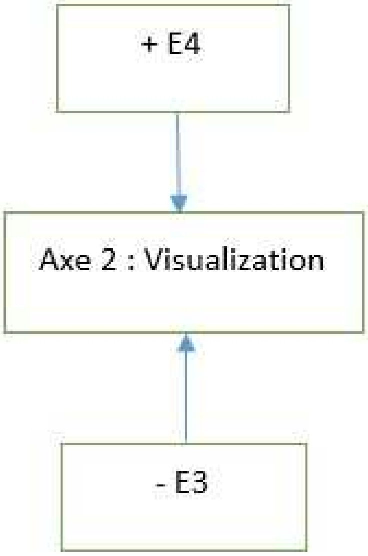

Figure 6: Third profile

3 Conclusion

Three profiles have been identified. The first profile concerns the question " Behavior contribution" which was positively influenced by the opinions of the three experts E1, E2 and E5. The second profile concerns the two questions "Action plan" and "Risk management" which were impacted by the opinion of experts E1, E2 and E5, in addition to these effects, the question "Risk management" is positively influenced by opinion of the expert E3 and negatively by the opinion of the expert E4. In profile number 3, the question ”Visualization” is impacted by the opinion of the E3 expert but positively influenced by the opinion of the E4 expert.

The profiles are produced on the basis of interpretations of the results which take into account the axes generated by the PCA. The proposed approach allowing a classification of the phases of an intrusion detection method based on the opinion of experts in the field. This will help define a targeted improvement policy.

The dimension of the matrix to be considered is not a hindrance. We used this simple example for two reasons, the first because it considers concrete data and the second reason to facilitate the understanding of the approach proposed to the readers of the work.

Acknowledgements. My thanks go to all those who supported me making this work by offering me their time, their energy and their commitments. I present to them my deep gratitude.

I would like to thank professors El Hassan Essoufi, Taoufiq Gadi and Abdellah Lamnii who are members of Laboratory: Mathematics, Computer Science and Engineering Sciences.

References

[1] E. M. Rajaallah, S. A. Chamkar, S. Ain El Hayat, (2019) Intrusion Detection Systems: To an Optimal Hybrid Intrusion Detection System. In: F. Khoukhi, M. Bahaj, M. Ezziyyani (eds), Smart Data and Computational Intelligence. AIT2S 2018. Lecture Notes in Networks and Systems, **66,** Springer, Cham. https://doi.org/10.1007/978-3-030-11914-0_30

[2] E. Bognar, Data mining in cyber threat analysis neural networks for intrusion detection, **15,** (2016), 187–197.

[3] A. Ajboye et al., Anomaly Detection in Dataset for Improved Model Accuracy Using DBSCAN Clustering Algorithm, (2015), 39–46.

[4] David Pierrot, Nouria Harbi, Jèrôme Darmont. Dètection des intrusions et aide a la dècision. 12e Confrence sur les Avancées des Systemes Dècisionnels (ASD 2018), May 2018, Marrakech, Maroc. hal-01761914.

[5] M. Ghoniem et al., VAFLE: Visual Analytics of Firewall Log Events. In Visualization and Data Analysis, (2014).

[6] A. Morineau, T. Aluja-Banet, Analyse en Composantes proncipales, Montreuil, CISIA, 2000, 142.

[7] G. Saporta, Probabilités, analyse des donnèes et statistiques, Technip, 2006, 493.

[8] T. W. Anderson, An introduction to multivariate statical analysis, New York, Wiley, 1984.

[9] D. Busca, S. Toutain, Analyse Factorielle Simple en Sociologie, de boeck, 2009.